

UC Davis

UC Davis Electronic Theses and Dissertations

Title

Investigating skin pigmentation genetic architecture and fine-scale population structure in southern African Khoe-San descendants

Permalink

<https://escholarship.org/uc/item/8bs805dq>

ISBN

9798297647398

Author

Al-Hindi, Dana Raed

Publication Date

2025-09-21

Peer reviewed|Thesis/dissertation

Investigating Skin Pigmentation Genetic Architecture and Fine-scale Population Structure in
Southern African Khoe-San Descendants

By

DANA RAED AL-HINDI
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Anthropology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Brenna M. Henn, Chair

Megan Y. Dennis

Heather L. Norton

Committee in Charge

2025

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

To those who have, and continue to, persevere through life in the diaspora.

To the families who carry the keys to their home from afar.

To my parents, siblings, husband, and extended family.

For the strength and grit taught to us by example.

For all the ways you've help me persevere.

Acknowledgments

I'd like to express my heartfelt gratitude to the participants and communities whose generous contributions made this research possible. It was a privilege to meet some of you, listen to your stories, and discuss the projects and their early findings. I'm grateful for your trust, collaboration, and many memories we shared together. These experiences truly made this research project much more intimate and meaningful.

I am incredibly lucky to have had guidance of Dr. Megan Y. Dennis, Dr. Heather L. Norton, and Dr. Brenna M. Henn, over these last few years. Megan has a gift for guiding students toward questions they never thought to ask and pushing them out of their comfort zone, and for that I am grateful. Heather, your encouragement over the years has been invaluable. Thank you for trusting me with your iris camera – and for not flinching when I returned it coated in Karoo dust. You have always shown such enthusiasm and support for my research, even when the results were confusing and negative.

Brenna, I'm deeply grateful that you took me on as your graduate student. When we first met, I had almost no research or computational experience, yet you saw my potential and welcomed me into your lab. Your mentorship over these past few years has enriched my development and experience while in graduate school. Thank you for training me and entrusting me with fieldwork. The experiences I have gained while in your lab has left me with unforgettable memories, like exploring rock art and fleeing bats in Elands Bay, floating down the American River, and braving the Gannaga Pass (twice).

To the Henn Lab: Chelsea Cataldo-Ramirez, Oshiomah Oyageshio, Justin W. Myrick, Neus Font Porterias, Gillian Meeks, Alaina Brenner, Kristin Hardy, and everyone else that has

passed through, thank you for your unwavering support. You have always been presented to bounce ideas off one another, share knowledge, and push each other to succeed. You've also always been ready to unwind and share a laugh or offer your home for my overnight stays. I am unsure how successful I would have been without your presence. Please always keep our sibling energy alive and remember to be kind to one another.

Austin W. Reynolds, I adopted you as family after our first field season together. Your guidance and support have profoundly shaped me personally and professionally. Thank you for always welcoming my chair at your desk and patiently answering my endless questions. You are truly one of my most honest friends, even if it means admitting you'd leave me during a zombie apocalypse or that you have no space for me in your Cederberg Cult.

Roshanne Bakhtiary, Neetha Iyer, and Mayowa Adegboyega from the Anthropology Department. Thank you for welcoming me during my first year and for teaching me lessons that went far beyond a traditional education. The three of you showed me what it means to be fearless – leading by example as you never hesitated to challenge contrasting ideologies with strangers at Sophia's.

Baba and Mama, my success is your success. If there was one essential reason I was capable of completing this degree, it is the both of you. You persevered and sacrificed so much of your own dreams to guarantee my own. Even when it took years to find the right words in Arabic to describe my research, you supported my decision to study human evolution. Alhamdulillah. I could not have asked for better parents.

Talal, Leena, and Salma – long before I began this journey, you were already teaching me some of life's most valuable lessons. From our wild Halloween candy chases, our clever one

penny-for-twenty trades, to the chaos of babysitting you, you've gifted me with resilience, creativity, and countless moments of joy. I love you three, immensely.

Hussein Naji Sawas, my caring husband, you have been my anchor these past few years. From buoying my spirits with your Baba-jokes during writing marathons to quietly making tea when I couldn't bear another word, your love and kindness carried me through. Thank you for believing in me, for the pep talks, and for always nudging me to reach higher. To the rest of the Sawas family, I am forever grateful to have you three in my corner: Ahmad Sawas, Maysoun Bochi, and Zane Sawas. Each of you have all inspired me in a different sector, and I will spend a lifetime trying to return the favor.

Lastly, to my family who are no longer with us: Seedo Adel Mohammad Al-Hindi, Seedo Mohammed Rashid Al-Ghoul, Amo Ehab Al-Hindi, Taitah Noof Ali, Amto Nadia Al-Ghoul. Your passings have been one of the greatest difficulties faced on this journey. Seedo Adel used to squeeze my hand and (in Arabic) say, "there will be a doctor in the family because of you." I like to think you all know I made it. Allah yerhamkum. I miss you every day.

Abstract

Africa harbours the most deeply divergent human lineages among living populations and is exceptionally rich in phenotypic and genetic diversity. The first lineages to have split are uniquely found among southern African Khoe and San. Traditionally, Khoe and San peoples spoke click-consonant Khoisan languages and practiced pastoralism and hunter-gatherer subsistence strategies, respectively. Though culturally distinct, geneticists often refer to them collectively as “Khoe-San” because of their high genetic affinity. The ancestral signature we refer to as Khoe-San is unique in its genetic diversity, retaining high levels of heterozygosity and ancestral alleles relative to other populations.

Today, there are few communities in southern Africa that still identify ethnically as Khoe or San. Inter- and intra-continental migrations over the past 3,000 years have profoundly shaped the region’s present ethnic mosaic, including signatures from a small East African pastoralist migration and the Bantu expansion. More recently over the last 400 years, European colonial impact has also given rise to highly heterogeneous Coloured communities, who constitute a sizable proportion of present-day South Africans (~42%). As a result of these multi-faceted migrations, it is not uncommon to find Khoe-San, European, South and Southeast Asian, and equatorial African ancestry among South African Coloured (SAC) communities (i.e. SAC are 5-way admixed). In this dissertation, I explore questions surrounding the contemporary phenotypic and genetic diversity of South Africa, with a focus on the SAC, one pastoral Khoekhoe-speaking group (Nama) and three San populations from South Africa (≠Khomani San) and Botswana (!Xoo and Ju|’hoansi).

Chapter 1, *Genetic Divergence within Southern Africa During the Later Stone Age*, sets our understanding of Khoe-San genetics by providing a comprehensive review focused on

uniparental lineage and genome-wide autosomal data divergence patterns within and between indigenous southern African Khoe and San groups. This chapter aims to contextualize the genetic diversity present in southern Africa during the Later Stone Age, a period when the 36,000-year-old Hofmeyr skull dates back to. This skull is impressively intact and might be one of the only anatomically modern humans remains in Sub-Saharan Africa that date to this time period, when modern humans were likely moving across the landscape. I conceptualize the effects that climatic changes may have on population migration. Additionally, I discuss the relationship between linguistics and genetics and also clarify differences in dating between the archaeological record and genetic evidence (i.e. population divergence vs. genetic divergence).

Khoe and San groups from South Africa, Namibia, and Botswana have been well described in previously published studies, but we know little regarding the Coloured populations in southern Africa. Few SAC individuals have been extensively sampled in South Africa's Western and Northern Cape provinces, with most published studies either focusing primarily on the Eastern Cape and Cape Town or relying on small Northern Cape sample sizes. While Cape Town lies in the Western Cape, it has been a focal point of recent migrations, and individuals sampled there may not be representative of SAC populations elsewhere. Chapter 2, *Recovering the Pre-Colonial Population Structure of Khoe-San Descendant Populations*, begins by characterizing the ancestral components present in the largest SAC cohort collected outside of Cape Town to date ($n = 620$). I put effort to distinguish which ancestral populations contributed these components into the population, while considering which populations are implicated by the historical record and previously published genetic research. For instance, the historical record surrounding the Dutch East Indian slave trade implicates Indonesia as a source of the Southeast Asian component found among the SAC, although we lack genetic evidence to support this. In

this chapter, I estimate SAC global ancestry proportions using ADMIXTURE, characterize their genetic composition with principal component analysis (PCA), infer local ancestry across the genome, and plot ancestry-specific multidimensional scaling to assess putative reference populations for each of the 5 ancestries. Khoe-San ancestry predominates across all SAC participants (>48%), indicating that these communities are primarily Khoe-San descendants.

After grasping the genetic diversity that exists among the Khoe, San, and SAC participants in this study, I move on to characterizing the phenotypic diversity among the populations. Most pigmentation studies have focused on Eurasian populations, with only a handful having turned to describe skin and eye color genetic variation among continental Africans. Chapter 3, *Assessing the effect of rs2470102 (SLC24A5) on iris pigmentation in southern Africa*, reports a candidate gene study of the intronic variant rs2470102 in *SLC24A5*, a strong-effect and well-characterized pigmentation gene whose lightening allele rs2470102*A has been associated with lightening brown eye color among Cape Verdeans. I assess the locus's putative influence on iris pigmentation among 262 individuals with high Khoe San ancestry and find the analysis largely underpowered due to fieldwork challenges in image collection and quality control, but the lessons learned inform recommendations for future iris image collection in remote regions.

Lastly, I build on previously published skin pigmentation studies which focused on continental Africans, including the Khoe and San. These studies report the genetic architecture of skin pigmentation among continental Africans to be far more complex than in admixed populations or populations adapted at higher latitudes (such as Europeans or East Asians). In Chapter 4, *Partitioned heritability estimates highlight genomic regions omitted by GWAS*, I merge previously published Khoe and San data with newly collected SAC samples to run the

largest Khoe-San focused skin pigmentation study ($n = 1119$). I begin by characterizing the heritability distribution across the genome by partitioning by chromosome. Partitioned heritability reveals enrichments on chromosomes 15, 5, 6 and 9, yet standard genome-wide association study (GWAS) reproduces only a previously known hit on chromosome 15, *SLC24A5*. Functional knockouts in zebrafish fail to validate any additional loci implicated in skin pigmentation. Moreover, I demonstrate how strong phenotype-ancestry correlation reduces power to detect pigmentation variants and breaks assumptions used in polygenic scores (PGS) development and transferability. Finally, I argue for the development of novel approaches capable of identifying other contributors to the genetic architecture of skin pigmentation.

Together, these studies illustrate how deep and recent demographic processes have molded genetic and phenotypic diversity in southern Africa and underscore the importance of population history and evolutionary forces in shaping the genetic architecture of skin pigmentation among Khoe-San and Khoe-San descendant communities.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	III
ABSTRACT	VI
CHAPTER 1. GENETIC DIVERGENCE WITHIN SOUTHERN AFRICA DURING THE LATER STONE AGE	1
1.1 ABSTRACT	1
1.2 INTRODUCTION: THE CRADLE OF MODERN HUMANS	2
1.3 CLICK-SPEAKING HUNTER-GATHERERS WITHIN AFRICA	3
1.4 SUBSTRUCTURE AMONG THE KHOE-SAN	5
1.5 LATE HOLOCENE GENE FLOW	14
1.6 FUTURE DIRECTIONS	15
1.7 ACKNOWLEDGMENTS	16
CHAPTER 2: RECOVERING THE PRE-COLONIAL POPULATION STRUCTURE OF KHOE-SAN DESCENDANT POPULATIONS	17
2.1 ABSTRACT	18
2.2 INTRODUCTION	19
2.3 RESULTS.....	23
2.3.1 <i>Genetic composition of Khoe-San descendant communities.....</i>	<i>24</i>
2.3.2 <i>Identifying potential sources for African population structure.....</i>	<i>28</i>
2.3.3 <i>Identifying potential ancestry sources for non-African population structure.....</i>	<i>31</i>
2.3.4 <i>Geographic predictors of population structure</i>	<i>33</i>
2.4 DISCUSSION	36
2.4.1 <i>Western African ancestry reflects both ancient Bantu expansion and colonial slave trade</i>	<i>37</i>
2.4.2 <i>Indonesian-ancestry and other Inter-Continental Migration.....</i>	<i>38</i>
2.4.3 <i>South African Coloured populations retain high Khoe-San ancestry.....</i>	<i>40</i>
2.4.4 <i>Colonial Impact on Ecogeographic Population Structure.....</i>	<i>42</i>
2.5 CONCLUSION.....	44
2.6 METHODS	45
2.6.1 <i>Sample Collection</i>	<i>45</i>
2.6.2 <i>Sample Consolidation, Genotyping, and PCA</i>	<i>45</i>
2.6.3 <i>Phasing.....</i>	<i>47</i>
2.6.4 <i>Identifying Relatives</i>	<i>47</i>
2.6.5 <i>ADMIXTURE</i>	<i>49</i>
2.6.6 <i>Local Ancestry Inference.....</i>	<i>50</i>
2.6.7 <i>Ancestry-specific MDS.....</i>	<i>52</i>
2.6.8 <i>Euclidean Distance Computation.....</i>	<i>53</i>
2.6.9 <i>SPRUCED.....</i>	<i>53</i>
2.6.10 <i>Within and Between Population IBD Sharing.....</i>	<i>55</i>

CHAPTER 3. ASSESSING THE EFFECT OF RS2470102 (SLC24A5) ON IRIS PIGMENTATION IN SOUTHERN AFRICA.....	57
3.1 ABSTRACT	57
3.2 INTRODUCTION	58
3.2.1 <i>Variation and Evolution of Iris Pigmentation</i>	58
3.2.2 <i>SLC24A5 lightening allele's introduction into southern Africa</i>	59
3.3 RESULTS.....	60
3.3.1 <i>Candidate gene study</i>	60
3.4 DISCUSSION	62
3.4.1 <i>Fieldwork Complications</i>	62
3.4.2 <i>Fieldwork Recommendations</i>	64
3.5 METHODS	65
3.5.1 <i>Genetic Data</i>	65
3.5.2 <i>Phenotype Data</i>	66
CHAPTER 4: PARTITIONED HERITABILITY ESTIMATES HIGHLIGHT GENOMIC REGIONS OMITTED BY GWAS.....	68
4.1 ABSTRACT	69
4.2 INTRODUCTION	70
4.3 RESULTS.....	74
4.3.1 <i>Significant differences in average melanin across Khoe-San and Khoe-San descendant populations</i>	75
4.3.2 <i>Correlation between ancestry and melanin index</i>	77
4.3.3 <i>Models including rs1426654*A dosage increase ancestry against melanin correlation</i>	79
4.3.4 <i>Heritability is unevenly distributed across the genome</i>	81
4.3.5 <i>Genome-wide association study</i>	82
4.3.6 <i>Constraints with including PCs in GWAS</i>	85
4.3.7 <i>Poor Transferability of European-Derived Polygenic Scores</i>	87
4.3.8 <i>Prioritizing candidates for CRISPR-driven gene knockout in zebrafish</i>	88
4.3.9 <i>Interrogating causality of candidate genes</i>	90
4.4 DISCUSSION	91
4.4.1 <i>Partitioned heritability estimates highlight genomic regions omitted by GWAS</i>	92
4.4.2 <i>Functional Annotation and Validation</i>	94
4.4.3 <i>Poor Transferability of European-Derived Polygenic Scores</i>	95
4.5 CONCLUSION.....	96
4.6 METHODS	96
4.6.1 <i>Sample collections and quality control</i>	96
4.6.2 <i>Merging within genotyping platforms</i>	97
4.6.3 <i>Phasing</i>	98
4.6.4 <i>Imputation</i>	98
4.6.5 <i>Post-imputation merging across platforms</i>	98

4.6.6 <i>Modeling Pigmentation Variation</i>	99
4.6.7 <i>Genome-wide association studies</i>	100
4.6.8 <i>Partitioned heritability</i>	101
4.6.9 <i>Transferability of Polygenic Scores</i>	102
4.6.10 <i>Functional Annotation of GWAS</i>	103
4.6.11 <i>Experimental gene-knockouts</i>	104
SUPPLEMENTARY INFORMATION	105
SECTION 1: SUPPLEMENTAL INFORMATION FOR CHAPTER 2	106
SECTION 2: SUPPLEMENTAL INFORMATION FOR CHAPTER 3	150
SECTION 3: SUPPLEMENTAL INFORMATION FOR CHAPTER 4	158
REFERENCES	194

Chapter 1. Genetic Divergence within Southern Africa During the Later Stone Age

Authors: Dana R. Al-Hindi, Austin W. Reynolds, Brenna M. Henn

Modified from DOI: 10.1007/978-3-031-07426-4_3

1.1 Abstract

Southern Africa is currently home to a wide array of diverse ethnic groups and ancestries resulting from agriculturalist and pastoralist expansions, colonialism, slave trade and modern global trade networks. The vast majority of this change occurred in the last 2,000 years, however. Prior to these migrations, southern Africa was populated by hundreds of hunter-gatherer populations which can be loosely grouped as “Khoes-San”. Many publications refer to these groups as Khoisan, San, Basarwa, Khoekhoe, Bushmen, Coloured, etc., terms which we unpack in greater detail. Here, our goal is to situate the Hofmeyr skull in the context of the genetic diversity present in southern Africa during the Later Stone Age. The primary source of information for this review is DNA samples from contemporary populations distributed from Angola to South Africa. Over the past decade, we have seen a remarkable surge in DNA data generated from genome-wide autosomal loci, i.e. “genomic” data. These new genomic datasets have been used to estimate the earliest population divergence between Khoes-San groups and other populations. Linguistic data is often used to inform genetic sampling projects; thus, we provide a brief introduction to linguistic diversity across southern Africa. We describe how geneticists have identified population structure among Khoes-San groups, how this ancestry is structured across the landscape, the time depth of population structure, and changes in population size through time. We aim to elucidate common misconceptions of genetic data (such as the idea

that population diverge at the same time as genetic lineages) and how/if genetic dates can be related to archaeological dates.

1.2 Introduction: The Cradle of Modern Humans

South Africa is famously home to the “Cradle of Humankind” World Heritage site, referring to the Sterkfontein cave system which has produced numerous Australopithecine fossils and, more recently, *Homo naledi* skeletons from the Rising Star Cave. The prevailing view among many geneticists is that southern Africa (broadly speaking) is also the cradle of modern humans. By this we mean that the majority of the genetic ancestry in modern human populations has its source in an ancestral population which lived in southern Africa. This ancestral population is the last time that *all* present-day humans can trace back their common ancestors which contributed genetic material to future generations, and the deepest human lineages are among indigenous southern African populations.¹⁻⁵ Consider an example for contrast: the 1.5% of Neandertal DNA identified in Eurasian/Oceanian populations is neither the majority of their genomic ancestry nor is it shared by all human populations. It is important to consider the implications of this definition because it does not *a priori* exclude other regions of Africa or even other hominin species from contributing to subsequent modern human evolution.

Based on the primacy of fossil evidence, eastern Africa has typically been considered the locus of origin for humans.⁶ However, in just the past decade, genomic evidence is nearly unequivocal in demonstrating that the deepest population divergence is between the Khoe-San and all other human groups. Eastern Africans simply do not carry the highest levels of genetic diversity nor the deepest divergent lineages among sampled populations. The absolute date of this Khoe-San divergence varies widely depending on the types of loci included, the mutation rate, and the method of analysis. Almost all dates range from 200 ka to 100 ka with a mode

around 150 ka-120 ka, which corresponds roughly the marine isotope stage (MIS) 5 to MIS 6 transition.⁷ The assumption in most tested genetic models is that human populations approximate an “isolation-migration” model of bifurcating population divergence which led to the modern genetic landscape. However, we note that more complex models of population reticulation or isolation by distance have not been widely tested).⁷⁻¹² With nearly a dozen genetic publications having examined the date of divergence with broadly congruent results, the focus has turned to how assumptions about how our methods influence the results (i.e., model misspecification, reference bias, ghost populations, isolation by distance) and how calibrated rates accord with paleoanthropological discoveries.

1.3 Click-Speaking Hunter-Gatherers within Africa

While Africa encompasses a considerable amount of genetic diversity, it also harbors great deal of linguistic diversity, exhibiting a similar pattern of (phoneme) diversity decreasing with distance away from central and southern Africa under a founder effect model.¹³ This is particularly useful when unraveling population demographic histories. Both biological and social interaction between populations is necessary for linguistic and genetic change to arise, leading language and genetic trees to have a strong correlation.¹⁴ This correlation has become a useful tool to genetic anthropologists. The linguistic classification of an individual (or their parents/grandparents) is often used as a proxy to help identify populations to sample.

There are four major linguistic phyla within Africa – Nilo-Saharan, Niger-Kordofanian, Afroasiatic, and Khoisan.¹⁵ It is worth mentioning that there have been exchanges of views regarding the nomenclature of the Khoisan language family and the populations, where “Khoisan” is often used interchangeably. Schultze (1928) combined Khoi (now Khoe), meaning ‘people’ among Nama pastoralists, and “San” meaning ‘forager’, to collectively reference

hunter-gatherer and pastoralist groups that reside within southern Africa. Since then, it has been borrowed by linguists to refer to their languages. The combined ethnonym has left some contemporary groups uncomfortable and culturally indistinguishable when referred to jointly.¹⁶ Mainly, members of the South African San community have asked to remain distinct from their Khoe-speaking neighbors and to be referred to by their specific ethnic identities or generally as the San.¹⁶ To preserve their request, this paper will reference the *populations* cumulatively as Khoe-San, separate than that of the Khoisan languages, and to acknowledge their difference in identity through hyphenation.

The Khoisan linguistic phylum is unique for its many and various click phonemes.^{15,17,18} These click-incorporating languages are only found within eastern and southern Africa, with the exception of Damin, a language that has independently established click-consonants in northern Australia.¹⁹ Five independent language families have been historically classified under Khoisan: Tuu, Kx'a, and Khoe-Kwadi from southern Africa, and Hadza and Sandawe from Tanzania.^{17,19} The Khoisan language map is illustrated in Figure 1.1. The Dahalo people in Kenya use click consonants but speak a Cushitic language that likely adopted click phonemes through past exposure with click-speaking populations (although an independent innovation of click consonants has not been ruled out).¹⁹ While the Khoe-San are closely related groups that reside relatively near one another, language families vary across populations and are not necessarily consistent with subsistence strategies. For example, the Taa and G||ana are primarily hunter-gatherers but speak Tuu and Khoe-Kwadi languages, respectively. The Nama prove to be another example, as they speak a Khoe-Kwadi language despite being genetically more similar to their hunter-gatherer Tuu-speaking ≠Khomani San and the Karretjie people.^{5,20,21}

1.4 Substructure Among the Khoe-San

The dispersal of the various subsistence and language practices have led researchers to question what factors have driven the distribution exhibited today. Investigating the genetic relationships between the Khoe-San groups has provided an interesting explanation for group distribution. Previous studies using genotyped single nucleotide polymorphisms (SNPs) have found that the Khoe-San populations separate into two clusters that correlate with geography, stretching in a northwest and southeast fashion through the Kalahari.^{5,20} Non-genetic variables such as language and subsistence strategies do not correlate as well with the ancestry distribution,²⁰ albeit genetic component predictions have suggested language plays a complicated role in the divide.⁵ More recent studies found the Khoe-San populations can be further divided into three ancestries,^{21,22} complicating the geographical correlation previously observed^{5,20} beyond a northeast and southwest pattern. The authors proposed that the population substructure was a result of the ecological boundaries.²¹ While an isolation-by-distance model can explain a 2-dimensional cline,^{5,20,22,23} it does not explain the circum-Kalahari signature (Fig. 1.2).²¹ There, the three ancestry signals are referred to as central Kalahari, northern Kalahari, and circum-Kalahari -- corresponding to the rim of the Kalahari.²¹

Taa (North, East, and West), =Hoan, Naro, Shua and Kua, G|ui, and G||ana (mixture of Tuu and Kx'a speakers). The remaining populations are highly composed of the northern Kalahari ancestry – Ju/'hoansi (North and South), Khwe, and !Xun. An uneven distribution of autochthonous mtDNA haplogroups L0d and L0k reinforce the ancestry clusters. While both haplogroups are commonly found among the northern Kalahari ancestry, L0k is generally absent from the circum-Kalahari ancestry (such as the Nama, Karretjie, ≠Khomani) and is present at low frequencies in the central Kalahari ancestry.^{21,24}

This is within the same timeframe many of the mtDNA and Y-chromosome Khoe-San haplogroups coalesce. Divergence within the autochthonous mtDNA L0d and L0k haplogroups and Y-chromosome A2, A3b1, and B2b lineages have been estimated to take place during the MIS 3-4.^{11,20,24–}
²⁷ The L0d1 branch is the most frequent Khoe-San mtDNA haplogroup.^{27,28} L0d1 is found in all Khoe-San populations and splits into two branches (L0d1a and L0d1c) that have been reported to coalesce by Barbieri et. al (2013) and Chan et al. (2019) at ~55 kya and ~47 kya, respectively. Similarly, L0d3a and L0d3b have varying coalescent dates to L0d3, previously reported at ~45 kya^{27,28} and more recently ~60 kya.¹¹ Within L0d, L0d2 is more common than L0d3 and is found in frequencies higher than 10% in the Naro (25%), G||ana (13%), Hai||om (27%), and Nama (34%). In Tuu-speakers from eastern Botswana, L0d2 is found as high as ~53% in the population. Furthermore, the coalescence dates of Y chromosomal haplogroup B2b (~45-60 kya) and those of the most common Khoe-San haplogroup A lineages (A2 at ~30kya and A3b1 at ~55kya) are among the oldest divergences in the human Y chromosome tree.^{26,29,30}

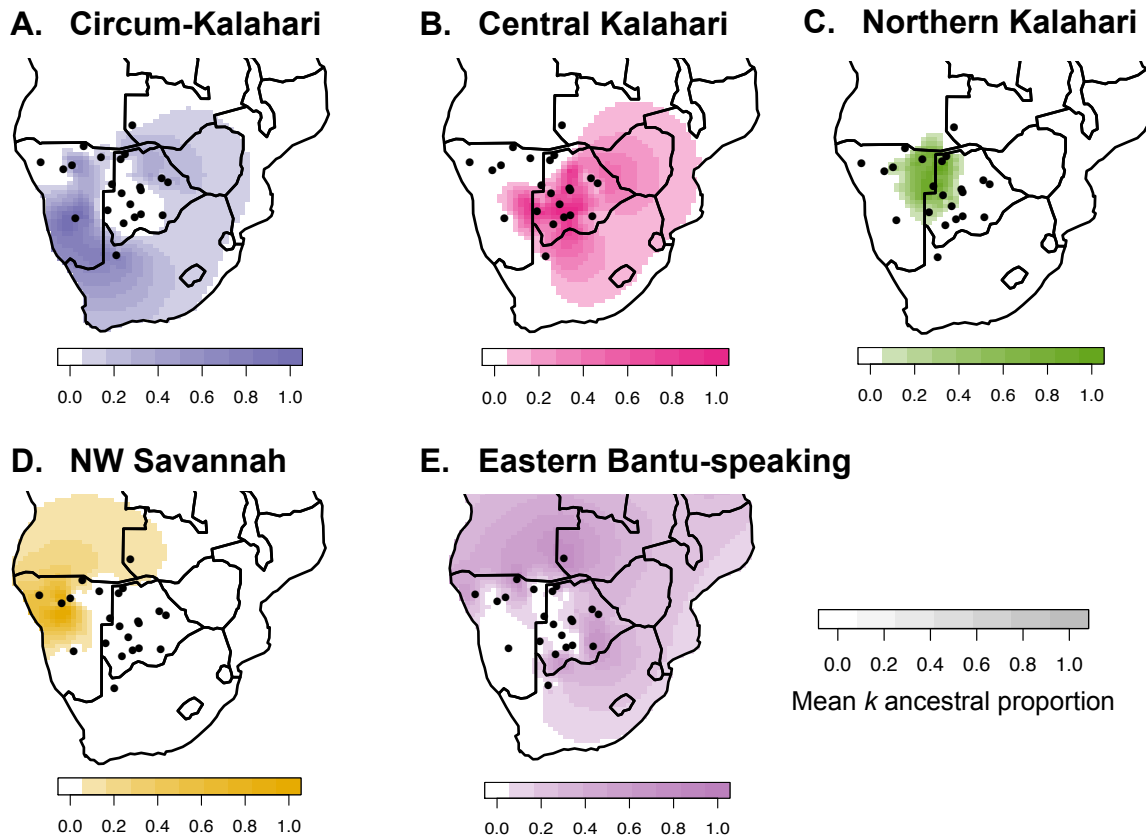


Figure 1.2) Five spatially distinct Khoe-San ancestries. The five most common ancestries in contemporary southern Africa populations are shown separately in A–E. Black dots represent the sampling location of populations in southern Africa. The third dimension in each map (depth of color) represents the mean ancestry proportion for each group for a given k ancestry, calculated from ADMIXTURE using unrelated individuals, and indicated in the color key as 0– 100% for each k ancestries. Reprinted with permission from Uren et al. (2016).

Low frequencies of the L0d branch are also found in eastern Africa. Previous studies hypothesized that Khoe-San populations inhabited eastern Africa prior to southern Africa, and the present-day distribution of L0 lineages could be explained by loss in drift and recent migrations into the region³¹. More recent work using 2,330-year-old aDNA samples predating the introduction of

pastoralism in southern Africa explains the various L0 haplogroups in southern and eastern Africa as a result of an ancient population's migration event out of modern-day Botswana starting ~130 kya¹¹. In this scenario, harsh climatic conditions during MIS 6 restricted an ancestral population to a refuge around palaeo-lake Makgadikgadi, until green corridors opened toward the northeast and southwest. Population migration in both directions resulted in the mtDNA distribution of L0 haplogroups observed today. Additional ancient genome-wide analyses including southern African samples ranging between 2,000 and 1,200-years-old have suggested an ancestral forager population wide-spread between eastern and southern Africa, based on signals ascribed as Khoe-San ancestry in ancient samples from eastern Africa^{8,32}. However, the linguistic affinity between eastern and southern Khoisan speakers is still debated. (Sands 1998; Güldemann and Vossen 2000).

It should be stressed that genetic divergence predates population divergence. Gene genealogies diverge when a new mutation occurs in a lineage, such as a new mutation in the mitochondrial genome. Both the ancestral and the derived mutation may continue to segregate in a population for some time. Whereas population divergence occurs when gene flow between two subpopulations ends, ceasing to homogenize allele frequencies. In the presence of minimal gene flow, subpopulations are subject to genetic drift separately from one another, allowing for allele frequencies to fluctuate such that, given enough time, distinguish ancestry clusters. Notice in 1.3, although genetic divergence (D_G) is observed in the ancestral population, population divergence (D_P) does not take place until later. This concept is important when synthesizing dates across disciplines, specifically between geneticists and paleoanthropologists. Of particular importance is building context around genetic divergence estimates within the Khoe-San alongside the available fossil record from this time period (i.e., the Hofmeyr skull).

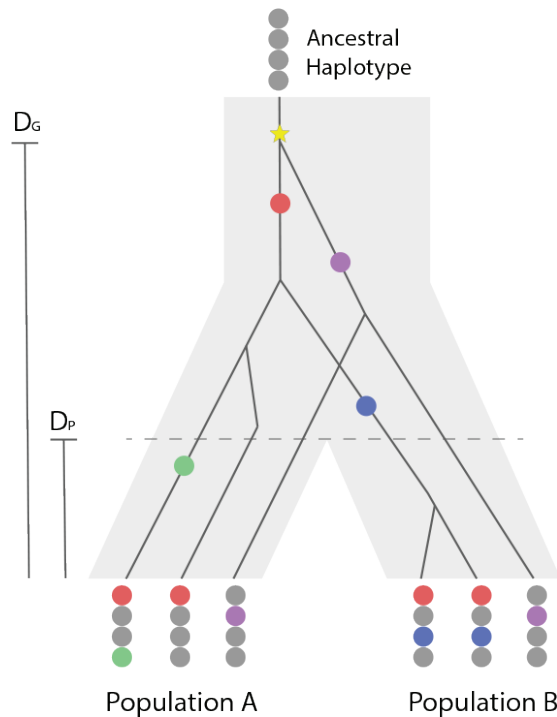


Figure 1.3) Genetic divergence versus population divergence adapted from Kimura (2015). Colored circles (red, purple, blue, and green) represent mutations on each haplotype. Each split along the lineage marks a point of genetic divergence (D_G). Population divergence (D_P) after genetic divergence, highlighted by the dashed line. While the populations share mutations (red and purple), private mutations (green and blue) along haplotypes differentiate them from one another. The four haplotypes coalesce and share a most recent common ancestor (MRCA) marked by the star. To clearly visualize divergence and the MRCA, the star is positioned slightly over the initial split in the lineage.

The Khoe-San ancestry clusters and estimated divergence time have often been based on Wright's F_{ST} , a commonly used statistic that describes genetic differentiation, or structure, between populations under the assumption of neutrality (Box 1, Equation 1) (Holsinger and Weir 2009). The timing of the divergence within Khoe-San using autosomal F_{ST} has been estimated to date back to ~25kya – 43kya, during MIS 3.^{5,20} Genetic differentiation between the Khoe-San

ancestry clusters are moderate (Wright's $F_{ST} \sim 0.012 - 0.034$),⁵ though these values can vary widely ($\sim 0.011 - 0.055$).²¹ F_{ST} values less than 0.05 and between 0.05 to 0.15 are usually considered low and moderate, respectively. These values can be seen between populations such as the Tuscans from Italy and South Asians (0.03) or the Han Chinese and South Asians (0.07).³³ Although, given that Khoe-San who reside within the same region and have similar subsistence, these values are substantial.

The northern Kalahari Ju/'hoansi have a relatively high F_{ST} score between the central Kalahari ancestry groups Shua (0.054), Kua (0.044), and G||ana (0.036).²¹ Interestingly, the Khwe cluster with the Ju/'hoansi within the northern Kalahari ancestry,^{5,21} but share an F_{ST} score (0.055) comparable to that of the Ju/'hoansi and the Shua. Their distinction within the northern Kalahari cluster has been suggested to be a result of gene flow with Bantu-speaking populations within the region.⁵ By contrast, the Nama have the greatest genetic differentiation with the Ju/'hoansi (0.038) and east Taa (0.025) who have high northern Kalahari and central-Kalahari ancestries, respectively. The Nama, who previously clustered with the southern San,⁵ are central Khoisan language speakers and pastoralists who are genetically more similar to the ≠Khomani (0.003) despite the difference in subsistence strategies. While the Damara are a Khoisan-speaking population, they are genetically distinct from other Khoe-San groups, and it is assumed that their language was recently adopted from the Nama.^{5,20,21}

The divergence patterns and distribution of ancestries within the Khoe-San has previously been suggested to be a result of climate change within southern Africa between the Middle to Later Stone Age (MSA and LSA).^{12,34} The diversification of archaeological artifacts during the MSA and LSA also suggests population fragmentation.^{35,36} Mackay and colleagues (2014) hypothesized that wetter and warmer climates (i.e., MIS 5 and MIS 3) promoted

population dispersal into more environments, causing an increase in diversification within the archaeological record (Fig. 1.4). Conversely, cold and arid climatic conditions (i.e., MIS 4 and MIS 2) are envisioned as having caused populations to coalesce leading to shared attributes between artifacts. Though initially proposed to describe the fragmentation and coalescence patterns observed in the archaeological data of southern Africa, this hypothesis could be applied to the Khoe-San genetic data.

Given that the divergence among the Khoe-San ancestries date back to generally ~30 kya, it is plausible the warm climate during MIS 3 encouraged population dispersal into environments that may not have been previously as favorable; although contradicting hypotheses regarding the influence interglacial periods have on the social landscape have been proposed (see Marean et al. 2020). During the harsher climate conditions of MIS 2, populations may have been geographically constricted by difficult to traverse ecological boundaries, thus limiting gene flow and leading to the varying ancestries found within the Khoe-San today (see Fig. 1.3). While the Khoe-San ancestry distributions across the landscape have commonalities that may support the hypothesis put forward by Mackay et al. (2014), further collaborative research and modeling will be necessary to examine how uniformly applicable this model is. The Hofmeyr Skull falls directly in the time frame of increasing genetic diversity. This fossil dates back to ~36kya³⁷ and has different morphological characteristics than those found in regional contemporary populations.³⁸ As it is currently the most complete human cranium from the sub-Saharan Africa during this timeframe,^{39,40} understanding its behavioral, cultural, and environmental context will paint a more thorough image of southern Africa's prehistory.

Box 1

Mean pairwise difference, denoted by π , is calculated based on single nucleotide variation between haplotypes randomly selected from a single population. This equation uses the frequencies of haplotypes i and j in the population, the nucleotide differences between i and j , and the summation of these three values multiplied together for all pairs of haplotypes to describe genetic diversity.

$$\pi = \sum_{ij} x_i x_j \pi_{ij}$$

(Eq. 1)

Where π_t is the average pairwise difference between two subpopulations, and π_s is the average pairwise difference within a single population. Various factors can increase or decrease the nucleotide diversity of a population, including mutation, genetic drift and gene flow. Populations that are genetically more similar to one another will have an F_{ST} value closer to 0, while more variation will result in a value closer to 1. In practice, values of 1 are not observed, as no human population carries the exact opposite set of alleles compared to another. The highest F_{ST} values observed between human populations would *at most* be around 0.30. The !Kung San and Karitiana from Brazil, for instance, are about as geographically separate as two populations can be and have an F_{ST} of about 0.25 (Ramachandran et al. 2005). Measurements of F_{ST} have a strong relationship with geography, where F_{ST} increases with increased distance between the populations (Ramachandran et al. 2005). Much of this stems from continental or insular barriers that cause population divergence and limit gene flow.

While F_{ST} is useful to describe genetic structure between or within populations, it can also be used to calculate genetic divergence times (Eq. 2). This equation is related to coalescent theory, a method that geneticists use to calculate the probability two alleles of the same gene coalesce one or more generations back in time. When doing so, it is necessary to take into account the effective population size, N_e , because genetic drift has a stronger effect on smaller populations than it does with larger populations. The effective population size is the number of individuals in an idealized (Wright-Fisher) population that has the same magnitude of genetic drift as the actual population. This number is usually smaller than that of the census population size because not every individual will bare offspring. In diploid organisms, both sets of chromosomes need to be accounted for when estimating time of coalescence, thus the effective population is doubled, $2N_e$.

$$1 - F_{ST} = \left(1 - \frac{1}{2N_e}\right)^t$$

(Eq. 2).

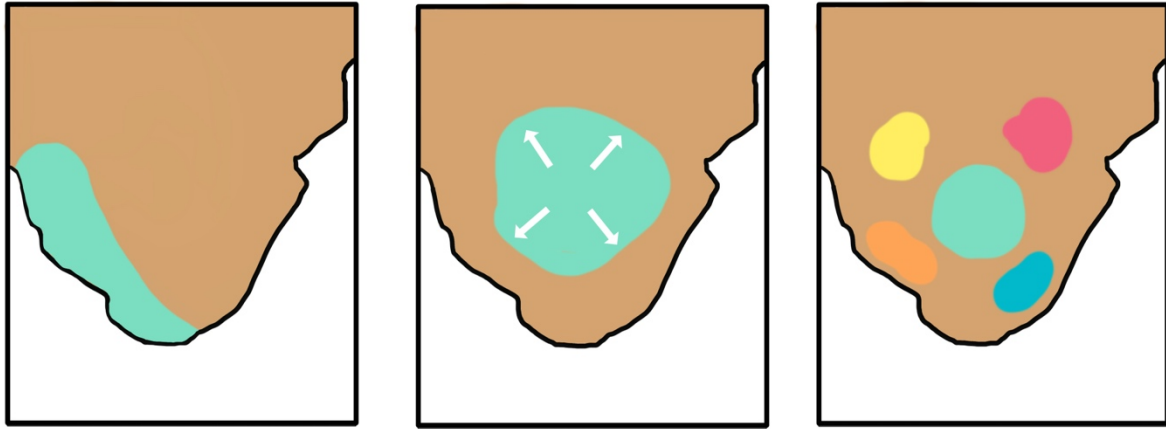


Figure 1.4) Southern Africa’s population coalescence and fragmentation conceptually depicted from MIS 4 to MIS 2. This figure illustrates a possible scenario in which population fragmentation leads to strong population differentiation among southern African Khoe-San. During MIS 4, populations (in teal) may have been sequestered along the southwestern coastline of southern Africa. Climate change and relative warming of MIS 3 promoted populations to expand and disperse inland. As the climate gradually cooled into the next glacial period, MIS 2, populations become isolated, restricted by ecological boundaries such as an arid Kalahari. Depiction here does not include the Paleo-Agulhas Plain off the Cape which would have extended out an additional 90 km and was a productive zone for coastal foraging.⁴¹ Artistic contribution: Anna Goldfield.

1.5 Late Holocene Gene Flow

Other lineages commonly found in contemporary Khoe-San groups are likely due to recent demographic processes such as the Bantu expansion and European colonization. Haplogroup B2a (M150) and many of the lineages in haplogroup E, for instance, are associated with the Bantu migration.⁴² However, one branch of haplogroup E (E3b1f-M293) makes up a significant proportion of the E lineages found Khoe-San groups in southern and eastern Africa and is consistent with the spread of pastoralism into southern Africa ~2,000 years ago.⁴³ These recent demographic processes have resulted in large variation in the Y chromosome haplogroup

frequencies in Khoe-San populations. The Ju/'hoansi have the highest frequency (~90%) of the characteristic Y-chromosomal haplogroups A and B, followed by the Nama (~70%), ≠Khomani San (~60%) and !Xuun (~60%).^{25,43,44} The lowest levels of KhoeSan Y-chromosomal haplogroups are found in the Damara, Khwe, and the Karretjie people (10-15% each).^{45,46}

1.6 Future Directions

Future interdisciplinary research is necessary to further unravel southern Africa's prehistory, though this is not without challenges. Different mediums, methods, and varying definitions across disciplines have given rise to opposing conclusions, further complicating interdisciplinary discussions.^{47,48} While in the field of population genetics, populations are defined by differentiation in allele frequencies, in archaeology they are often defined by site and location. This can entail a difference in the null hypotheses related to population movement versus continuity during the Late Pleistocene. Archaeologists may perceive hunter-gathers as mobile, non-sedentary groups. The extent of that mobility can vary, but the null hypothesis still presumes that cultural continuity does not imply population continuity. Rarely are archaeologists confident that the assemblage at hand is from a single episode, but rather a time-averaged palimpsest. That is, an observed archaeological site typically depicts an average behavioral trend over generations rather than single groups. On the other hand, geneticists generally begin with a null hypothesis that, while people may migrate (at varying rates), meta-populations as a whole are not moving across large landscapes during prehistoric conditions unless they are expanding into unpopulated environments. That being said, a common critique of geneticists is the scarcity of hypothesis testing.

The trans-disciplinary contrast in definitions seems natural given the different mediums and scales used between the fields. For example, mtDNA demographic modeling by Powell et al.

(2009) showed patterns indicative of population growth within southern Africa during the late Pleistocene, though the archaeological record came to contrasting conclusions.⁴⁹ The paleoanthropological field is also subject to specific constraints. The acidity of southern Africa's soil has left the paleontological record slim, handing paleoanthropologists and archaeologists little preserved organic material. The Hofmeyr skull serves as a glimpse of past modern human morphology but is the only nearly complete sub-Saharan cranium from its time period and region that can do so.⁴⁰

Reconciling differences across disciplines is crucial if we wish to elevate our anthropological understanding of modern humans in Africa during the Later Stone Age, and how the Hofmeyr skull fits within the archaeological and genetic data. More broadly, it is imperative to address these differences to build a more in-depth comprehension of human evolution. With gaps in the tangible record and the various resources and methods used across disciplines, complimentary and collaborations approaches tackling similar questions is a necessity.

1.7 Acknowledgments

We extend our appreciation to all the African participants that have contributed to the genetic studies reported in this paper. We thank Fred Grine for the invitation to be a part of this volume. Furthermore, we are grateful to Dr. Fransesco Montinaro, Dr. Teresa Steele, Sara Watson, and an anonymous reviewer for their comments. Lastly, we would like to acknowledge and thank Dr. Monika Feinen for her efforts in updating the Khoisan language map, and Dr. Anna Goldfield for creating our population coalescence and fragmentation figure. We thank the Genetics Society of America for granting permission to re-print figure 1.3.

Chapter 2: Recovering the pre-colonial population structure of Khoe-San

descendant populations

Authors: Stacy L. Edington*, Dana R. Al-Hindi*, Neus Font-Porterías, Alexandra Surowiec, William J. Palmer, Justin W. Myrick, Paul J. Norman, Marlo Möller, Brenna M. Henn[^], Austin W. Reynolds[^]

Authors contributions:

This manuscript was co-authored by Stacy L. Edington and Dana R. Al-Hindi. Dana R. Al-Hindi conducted ancestry analyses for the study populations, including ADMIXTURE, PCA, local ancestry inference, ancestry-specific multidimensional scaling, and centroid plots, as well as identifying Khoe-San populations from previously published data and merging samples across datasets. Stacy L. Edington estimated migration rates across South Africa using MAPS, conducted SPRUCE random forest analyses, and ran ancestry-by-distance linear models. Neus Font-Porterías assisted in assessing local ancestry inference models. Dana R. Al-Hindi, Alexandra Surowiec, Justin W. Myrick, Dr. Brenna M. Henn, and Dr. Austin W. Reynolds contributed to fieldwork and sample collection in South Africa. William J. Palmer aligned and quality-controlled the #Khomani San whole-genome samples used in the study. Dr. Brenna M. Henn and Dr. Austin W. Reynolds served as the lead principal investigators.

2.1 Abstract

San populations from Botswana and Namibia retain exceptional linguistic, cultural and genetic diversity, but few Khoisan-speaking groups remain south of the Kalahari Desert. However, historically, far southern Africa was home to many San and Khoekhoe groups. Popular opinion often implies that such populations do not contribute to the ancestry of contemporary South Africans. Here, we characterize the genetic ancestry of South African groups and reconstruct pre- and colonial population structure from 620 newly sampled individuals. These groups retain majority Khoe-San genetic ancestry (>48%), suggesting the persistence of Khoe-San ancestry to the present day. By isolating the Khoe-San ancestry component, we show that it is intermediate between the $\neq$ Khomani San and Nama, and distinct from Kalahari Khoe-San populations. We also find that signatures of the Indian Ocean slave trade can be traced to Indonesian islands such as Sulawesi, Java, and Flores, while the South Asian ancestry is regionally non-specific.

2.2 Introduction

Present-day populations in South Africa reflect diverse origins, from the Indigenous San and Khoe peoples to migrants from varying regions throughout Africa, Europe, and Asia. The region's ethnic mosaic results from a series of complex migrations over the last two thousand years (ky). The earliest migrations of pastoralist groups from eastern Africa into the region and later influx of agro-pastoralist groups of western African ancestry have been well documented in both the archaeological and genetics literature.^{50,51} However, the impacts of more recent population movements during the colonial period have been less well studied from a genetic perspective.

The arrival of the Dutch East India Company (in Dutch: *Vereenigde Nederlandsche Geoctroyeerde Oostindische Compagnie* or “VOC”) in 1652 initiated South Africa's colonial period in Table Bay, situated near modern day Cape Town.⁵² At the time, the Company intended for the Cape to act as a refreshment post for ships traveling between Europe and Asia, as its primary colonies were situated in Sri Lanka and Jakarta – historically known as Ceylon and Batavia.⁵³ But by 1690, the resupply station had grown into a permanent Dutch colony. The VOC maintained its occupation of the Cape for nearly 200 years, its rule ending when the British seized control of the region in 1806.⁵³ During this period, the VOC is estimated to have imported upward of 63,000 to 80,000 enslaved individuals from eastern coastal Africa, Madagascar, India, Sri Lanka, and the Malay Archipelago.^{52,53} While historians have reconstructed the port of origin for many enslaved individuals, significant gaps remain in our understanding of how they contributed to the genetic diversity of contemporary South Africans.

Over the next several decades, European colonists continued their expansion past Cape Town reaching the Cederberg mountains to the northwest by the 1720s, extending further north

into the Richtersveld and southern Kalahari in the early 19th century, and advancing into the arid inland Karoo by the 1840s (Figure 2.1).⁵⁴ Their incursion resulted in population decline and the systematic displacement of the San and Khoe Indigenous groups.^{55,56} In some cases, San and Khoe individuals (hereafter collectively referred to as Khoe-San, reflecting their shared genetic ancestry) were incorporated into colonial society, forming new genetically heterogeneous communities with VOC-forced migrants and colonists, later subsumed under the term “Coloured”. This was a pre-apartheid term historically imposed to categorize people of mixed ancestry; that is, individuals who did not fit within the constructed ‘White’ or ‘Bantu’ racial categories.⁵² The term is still in-use today and a formally recognized ethnic classification in South Africa, although we acknowledge that it may carry derogatory connotations in some contexts.

Despite this history, it is commonly thought that there is no Khoe-San ancestry in South African Coloured (SAC) communities today. SAC individuals make up ~41% of the population of the Northern and Western Cape Provinces⁵⁷ and are frequently described as five-way admixed,^{58,59} retaining Khoe-San, European, non-Khoe-San African, South and other Asian ancestries. While the global genetic ancestry of these communities along the southern coast have been characterized,^{58,60–62} the temporal dynamics of admixture and the relative genetic contributions of different ancestral groups of SAC communities further from Cape Town remain largely unexplored.

We address these gaps in knowledge by analyzing 620 newly genotyped samples from self-identified SAC and Khoe-San individuals from the Western and Northern Cape Provinces, along with 22 reference populations that serve as putative sources of genome-wide ancestry from regions proximal to VOC operation sites and trade routes. These include populations from the

Coromandel and Bengal coasts and across the Indonesian archipelago. Using this new robust reference panel, we investigate the genomic impacts of colonial dynamics in these regions and assess putative source populations for each of the five major distinct ancestries previously found among the SAC populations: Khoe-San, Bantu/non-Khoe-San African (NKA), South Asian, European, and East and Southeast Asian.

Specifically, we focus on the Khoe-San ancestry found in these communities. Compared to other global populations, the Khoe-San exhibit the highest levels of genetic heterogeneity, retain more ancestral alleles, and harbor autochthonous Y-chromosome and mitochondrial DNA (mtDNA) haplotypes.^{12,26,30,63,64} Previous studies focused on Khoe-San substructure for groups in and around the Kalahari Desert, estimated to have diverged ~20-30kya.^{5,20} The Botswanan San–Tuu speaking !Xoo and Kx’a speaking Ju|’hoansi–harbor mostly northern-Kalahari Khoe-San ancestry, while the South African Tuu-speaking ≠Khomani San and Khoekhoe pastoralists Nama retain majority southern-Kalahari ancestry. Models including geographic location, language and subsistence strategies improved the predictive power of these genetic components than geography alone, suggesting that all of these factors contribute to genetic substructure within the Khoe-San.⁵ However, the genetic affinity of Khoe-San peoples south of the Kalahari Desert, including our study region, remains largely unknown, despite this region historically being home to thousands of individuals.⁶⁵ We characterize pre-colonial population structure and haplotype sharing of these Khoe-San descendant individuals by isolating the Khoe-San ancestry haplotypes and assessing their genetic similarities relative to four Khoe-San reference populations: the ≠Khomani San, Nama, !Xoo, and Ju|’hoansi.

Finally, we estimate the proportion of Khoe-San ancestry that persists across more than a dozen contemporary communities, based on sampled individuals from the Cederberg, Karoo,

Middle Orange River, and Upington regions and investigate how geographic and environmental factors influence gene flow between these communities. Previous efforts to characterize the Khoe-San ancestry among SAC populations were more heavily focused on the Western Cape^{23,60} or were limited by small sample sizes from the Northern Cape Province (~50 total),^{5,58,66} potentially missing representation for the broader population.

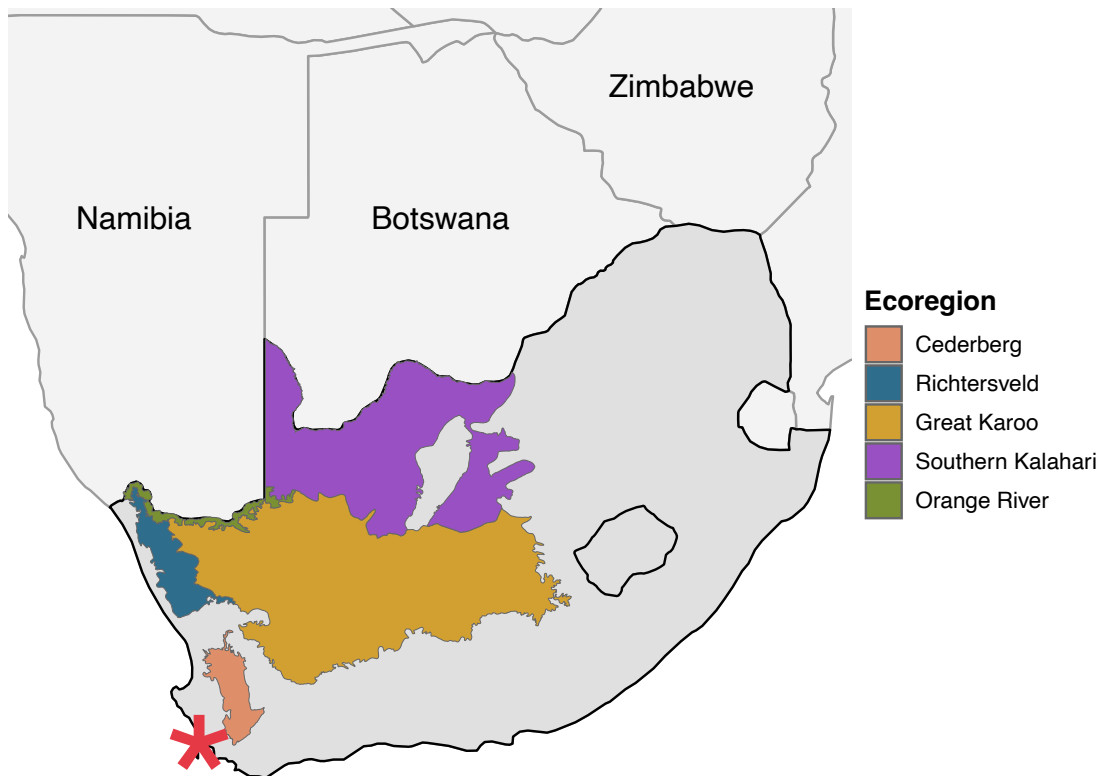


Figure 2.1) Map of South Africa highlighting the sampling regions included in this study. Colored markers indicate the geographical locations where participants were recruited: Cederberg (n=204), Richtersveld (n=206), Karoo (n=176), Southern Kalahari (n=310), and the Orange River (n=101). Total sample size across all regions was 997 individuals. The red star indicates Cape Town.

2.3 Results

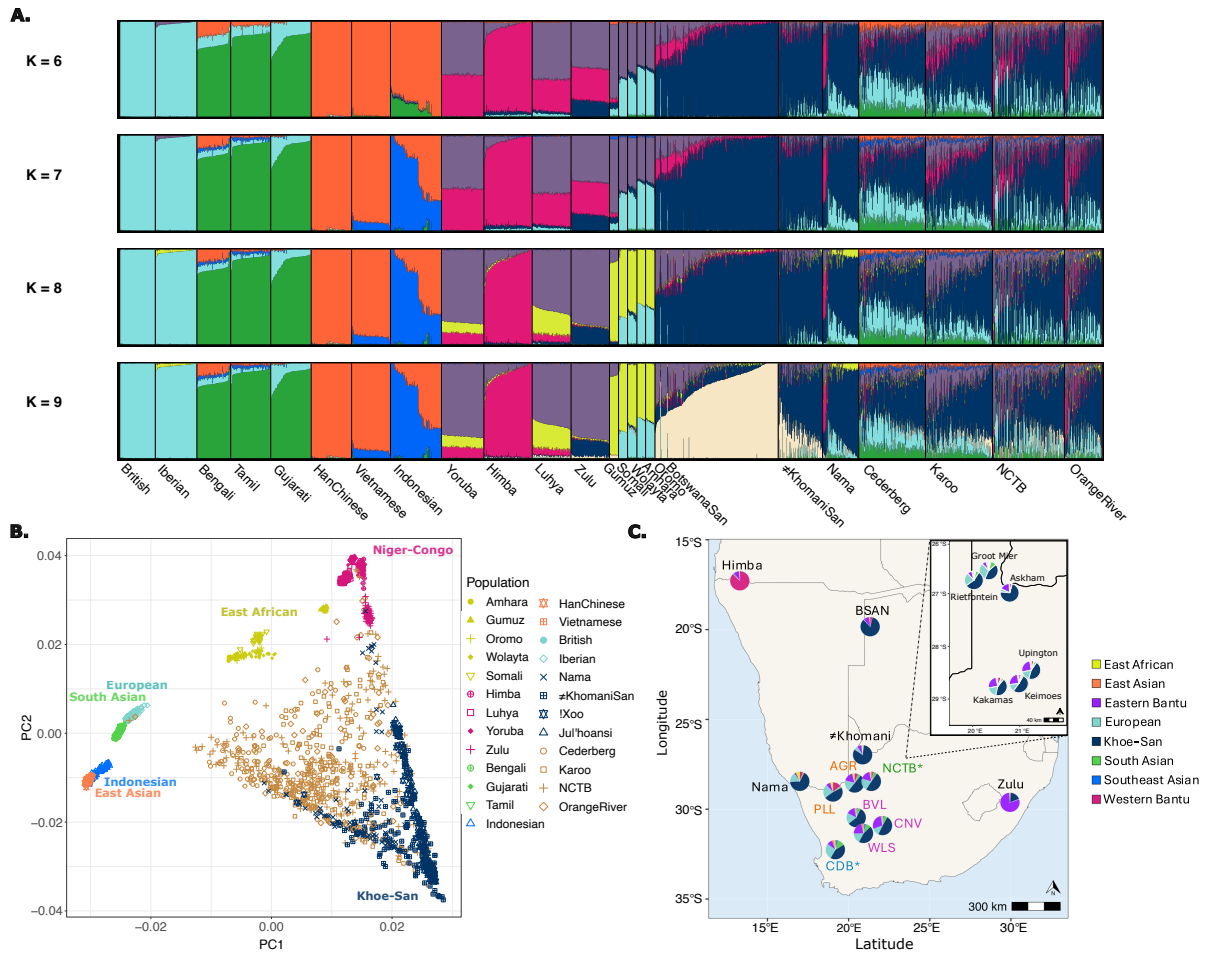


Figure 2.2) Characterizing global population structure within newly collected samples. Colors throughout the figures corresponding to broadly inferred ancestry from ADMIXTURE $k=8$. Population labels and abbreviations correspond the respective genome project (1000GP, African Genome Variation Project, and HGDP), with the exception of Indonesians (IND). In some instances, the !Xoo and Jul'hoansi, who were inferred based on a series of analyses (*see Supplemental Text*) are merged into one population: Botswanan San (BSAN). **A)** Unsupervised ADMIXTURE run of $k=6$ through $k=9$. **B)** PCA on 806,460 congruent H3Africa array markers. Visualized are 1,165 individuals along PC1 and PC2, with shapes distinguishing populations within broader inferred ancestries ($k=8$). Study populations are colored beige, while shapes indicate sampling locality within South Africa. **C)** Approximate locations and ADMIXTURE $k=8$ results of southern African populations used in the study. Newly collected samples are segregated and colored within sampling regions,

with the exception of the Cederberg (CDB) in blue and the Northern Cape Tuberculosis (NCTB) project in green. Towns sampled across the CDB are too close in proximity to segregate, while the inset map depicts sampling locations for the NCTB project. Orange River (orange) samples were collected from Pella (PLL) and Augrabies (AGR), while Karoo (purple) samples were collected across three towns: Williston (WLS), Brandvlei (BVL), and Carnarvon (CNV).

2.3.1 Genetic composition of Khoe-San descendant communities

We characterize the genetic composition of the largest set of SAC samples (n=997) derived outside of Cape Town, including newly collected samples spanning four southern African regions (Cederberg, Karoo, Upington area, and Middle Orange River (n=620); Figure 2.1, Figure 2.2C). Newly collected southern African samples were assayed for 2.5 million markers on Illumina's H3Africa array, then merged with a reference panel comprised of 1,165 individuals from inter- and intra-continental populations (see *Sample consolidation and PCA Methods*). Building upon previous studies, we assembled a comprehensive reference panel that aims to capture greater depth and fine-scale genetic diversity of southern African populations. The reference panel was also assembled with the goal of historically contextualizing the various migrations into southern Africa^{58,60,62,66}. Here, we include populations that represent six broad genetic ancestries: Niger-Congo, Khoe-San, European, South, Southeast (Indonesian), and East Asian ancestries, with >120 samples representing each ancestry (Figure 2.2A and 2.2B). Niger-Congo languages are the most commonly spoken language families spoken across sub-Saharan Africa, as a result of the Bantu-expansion. Populations included in this study and are those whose broader ancestry is labeled Niger-Congo reside western, eastern and southern Africa, but can largely trace back their genealogies to western Africa. We include populations to capture western and eastern African Bantu-expansion streams,^{67,68} 5 East African populations, 130

Indonesian genomes, and over 500 individuals representing 4 Khoe and San populations to capture both northern- and southern-Kalahari genetic signals (*Supplemental Table 1*). The final merged dataset was comprised of 806,460 autosomal markers.

Following quality control filtering (see Methods), newly collected samples were assessed using principal component analysis (PCA).⁶⁹ Principal Component (PC) 1 segregates populations out-of-Africa from Khoe-San populations, while PC2 separates within-Africa populations, mainly the southern African Khoe-San groups from Niger-Congo speaking populations originating from western/central Africa (Figure 2.2B). The majority of Karoo (KRO), Cederberg (CDB), Upington area (NCTB), and Middle Orange River (ORG) communities are distributed across the center of PC1 and PC2. Similar distribution of reference population and sampled SAC communities has previously been observed.^{58,62}

We further ran unsupervised ADMIXTURE⁷⁰ runs across multiple levels to infer ancestry clusters, setting $k=4:10$ (Figure 2.2A; *Supplemental Figure 2.6*), with $k=9$ and $k=10$ demonstrating the lowest cross-validation error rates (cv error=0.519, see *Supplemental Figure 2.7*). At $k=5$, we observe ancestries that correspond to European, South Asian, East and Southeast Asian, Khoe-San and non-Khoe-San African (NKA). Using ADMIXTURE $k=5$ results, we assessed these ancestry proportions within the South African sampled populations and their sampling region's geographical distance from Cape Town. We find an increasing gradient of Khoe-San ancestry with increased distance from Cape Town, ranging from an average of 40% near Cape Town to 80% in the southern Kalahari ($R^2=0.74$, $p=0.013$; Figure 2.2C and Figure 2.3). Conversely, South Asian and East and Southeast Asian ancestries show a negative correlation with increasing distance from Cape Town ($R^2=0.92$, $p=6.37 \times 10^{-4}$ and $R^2=0.86$, $p=2.85 \times 10^{-3}$, respectively; Figure 2.2C and *Supplemental Figure 2.9*), while distance

from Cape Town explains a lower amount of the variability in European ($R^2=0.63$, $p=0.033$) and NKA ancestry ($R^2=0.042$, $p=0.66$; Figure 2.2C and *Supplemental Figure 2.9*).

The inferred NKA ancestry begins to segregate into two components at $k=6$, one of which is highest in the Himba from Namibia, likely a result of their endogamy and recent bottleneck.⁷¹ At $k=7$, the southeastern Asian component distinguishes inferred ancestry in Indonesians from the Han Chinese and Vietnamese. A proportion of Indonesian ancestry is present in the sampled SAC populations, though the extent varies across groups. The CDB exhibits the highest proportion at $3.1\% \pm 2.1\%$, followed by the KRO at $2.1\% \pm 1.4\%$, while the NCTB and ORG communities show slightly lower proportions, $1.3\% \pm 1.4\%$ and $1.1\% \pm 1.3\%$, respectively. Unsurprisingly, this likely has a correlation with distance from Cape Town, as previously published SAC samples from District 6, Cape Town were reported to have a higher Indonesian (~20%) and South Asian ancestry (21%).⁶²

Eastern African ancestry begins to differentiate from the other populations at $k=8$, shown in yellow (Figure 2.2A and 2C). The Gumuz show highest frequency of this component compared to other Ethiopian and Kenyan groups, with smaller proportions also present in the Yoruba, and Nama (~6.5%). The relatively elevated East African ancestry in the Nama has been previously identified by genetic and archeological research, due to the sheep/goat/cattle pastoralist diffusion around 2kya.^{58,72,73} However, this ancestry is a minority among the ≠Khomani San ($\sim 1.5\% \pm 2.1\%$) and Botswanan San populations ($\sim 1.5\% \pm 1.4\%$), with most of NKA ancestry similar to the Yoruba, Luhya and Zulu (and variably affiliated with the Himba across k 's). The ORG sampling sites (Pella and Augrabies) collectively show higher East African ancestry ($3.4\% \pm 2.3\%$) compared to the other populations (CDB $2.4\% \pm 1.5\%$, NCTB $2.0\% \pm 1.9\%$, and KRO $1.0\% \pm 1.1\%$; see Figure 2.2A and 2C).

The northern- and southern-Kalahari Khoe-San ancestries segregate at $k=9$, primarily differentiating the ≠Khomani San and Nama from the two Kx'a speaking Botswanan San populations, with the latter representing a northern-Kalahari ancestry. Together, the Botswanan San populations harbor the most northern-Kalahari ancestry ($76.4\% \pm 18.5\%$), relative to other Khoe-San reference populations, and minimal southern-Kalahari ancestry ($10\% \pm 6.6\%$). The majority of the Khoe-San ancestry both the ≠Khomani San and the Khoekhoe-speaking Nama is southern Kalahari ($60.0\% \pm 12.7\%$ and $69.5\% \pm 19.2\%$, respectively), while the ≠Khomani contain a notably higher proportion of northern Kalahari ancestry ($25.5\% \pm 15.2\%$) compared to the Nama, who show only $2.2\% \pm 2.9\%$. Populations sampled from the CDB, KRO, ORG, and in Upington area (NCTB), share relatively consistent southern-Kalahari Khoe-San ancestry, ranging between 42.0% and 46.8% , while northern-Kalahari frequencies are uniformly lower, spanning from 2.6% in the CDB to 9.0% among NCTB participants, similarly to other sampled SAC populations from South Africa.⁵⁸ Interestingly, at this k value we find the Zulu encompass $16.7\% \pm 1.7\%$ southern-Kalahari Khoe-San ancestry, compared to their low northern Khoe-San component of $2.3\% \pm 0.9\%$ (Figure 2.2A; *Supplemental Figure 2.6*).

Finally, at $k=10$, a new cluster (lime green) appears in a subset of Indonesian individuals. At this resolution, the previously inferred East Asian component (orange, highest among the Han Chinese) found across the CDB, KRO, ORG, and NCTB participants $k=7:9$ now clusters with new Indonesian ancestry (*Supplemental Figure 2.6*). The CDB displays the highest proportions of this component $3.0\% \pm 1.8\%$, followed by $2.3\% \pm 1.7\%$ in the KRO, $1.4\% \pm 1.8\%$ in the NCTB, and $0.8\% \pm 1.3\%$ in the ORG. The ≠Khomani San results replicate previously published findings, showing little to no Indonesian ancestry from either cluster (0.005%), while the Nama and Botswanan San show no detectable Indonesian clusters across all k 's (<0.001).⁶²

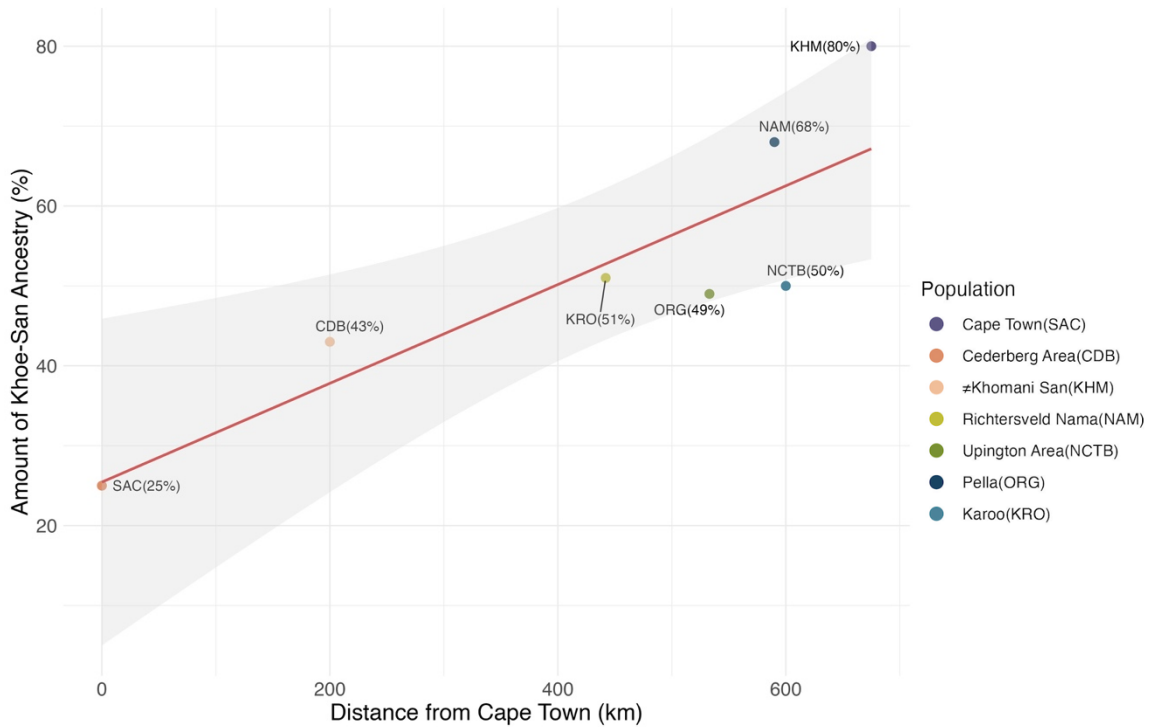


Figure 2.3) Percent of Khoe-San ancestry across southern African populations by distance from Cape Town, South Africa. $R^2 = 0.74$

2.3.2 Identifying potential sources for African population structure

To investigate potential sources for African population structure, we performed multidimensional scaling (MDS) analysis on our sampled populations from the CDB, KRO, Upington area (NCTB) and the ORG. Ancestry specific MDS (asMDS) allows for a “zoomed in” view on specific ancestry components in admixed populations. Local ancestry inference (LAI) was preformed using GNOMIX with five broad reference groups (Khoe-San, European, South Asian, non-Khoe-San African and East and Southeast Asian).^{74,75} Ancestry-specific haplotypes were then projected via ancestry-specific multidimensional scaling (asMDS) (*see Methods*;

Figure 2.4A and 2.4B).⁷⁴ We aim to best characterize specific source populations for the five primary ancestries within the admixed study populations.

Our Khoe-San asMDS analysis incorporates South African and Botswanan Khoe and San groups with diverse Khoisan languages. Specifically, the reference populations included are the !Xoo, ≠Khomani San, Ju|'hoansi, and the Nama (Figure 2.4A).^{76,77} Additionally, included with our sampled populations was a control group, the Zulu, who have previously been characterized to have ~20% Khoe-San ancestry⁷⁸ (*Supplemental Figure 2.17*). We replicate previously observed structure within the four Khoe-San reference populations, with the first axes largely distinguishing between the northern and southern Khoe-San ancestries (i.e. structure between the Botswanan San and the ≠Khomani San and Nama).^{20,22} The refined !Xoo and Ju|'hoansi groups (originally subsumed under “BSAN”) centroids are distant from ≠Khomani San than the Nama (*see Supplemental Table 5*). The second axis separates the !Xoo and Ju|'hoansi from one another. Khoe-San inferred haplotypes from the Cederberg, Karoo, and NCTB, and Middle Orange River cluster between the overlapping tail ends of Nama and ≠Khomani San haplotype distributions along asMDS1 (Figure 2.4A), though all study populations' centroids lie closer to the ≠Khomani San than the Nama (*Supplemental Table 5*).

The first dimension of our NKA asMDS segregates Yoruba and western Bantu-speaking populations (Zulu, Himba, and Luhya) from the eastern African reference populations (Gumuz) (Figure 2.4B). Along the second axis, we find the Himba pull away from the other broadly western African populations, likely due to a bottleneck which the population has experienced. Unexpectedly, the haplotypes from our sampled populations largely fall on top/near the Yoruba, Zulu, and Luhya, with some individual haplotypes' exhibiting a cline towards the Himba (Figure 2.4B, *Supplemental Figure 2.19*). Notably, the Middle Orange River individuals exhibit a cline

toward the Himba, though still more similar to other broadly southern African and western Bantu-speaking NKA groups (*Supplemental Figure 2.19*).

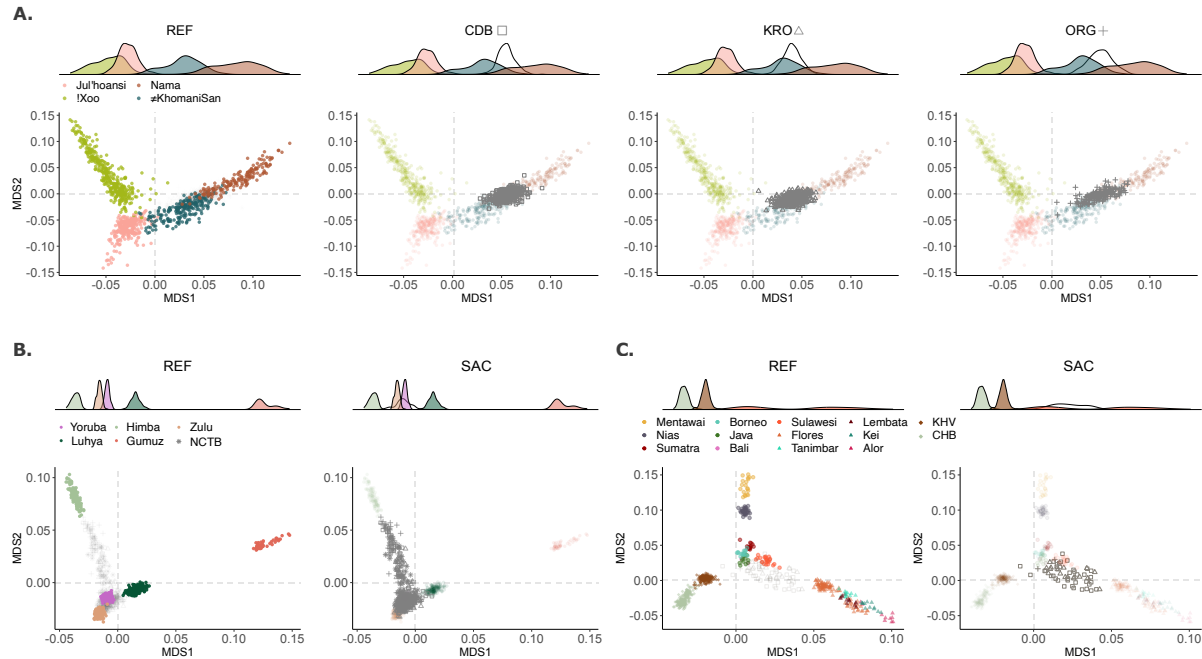


Figure 2.4) Ancestry-specific MDS analysis across three ancestries. In all panels, the left ancestry-specific MDS (asMDS) plot highlights the reference population, while the right focuses on the queried population's projected haplotypes, mainly the Cederberg (CDB), Orange River (ORG), and Karoo (KRO). Haplotypes were filtered for a minimum 75% probability for the respective ancestry being assessed. Shapes for our sampled populations are found beside their respective asMDS. Each asMDS figure is accompanied by a density plot corresponding to asMDS1. **A)** Khoe-San specific MDS analysis with each of our sampled populations (labelled SAC) in the right three figures. **B)** Non-Khoe-San African (NKA) specific MDS analysis. The panel on the right jointly visualizes the CDB, ORG, and KRO projected haplotypes, with the addition of NKA haplotypes found among #Khomani San (KHM) and Nama individuals with less than 85% Khoe-San ancestry. **C)** Southeast and East Asian specific MDS analysis. Here, Indonesian samples are labeled and colored by island. Shapes geographically differentiate Asian populations into three regions: mainland (diamond), western Indonesians (circles), and eastern Indonesians (triangles).

2.3.3 Identifying potential ancestry sources for non-African population structure

Here we assess potential ancestry sources for non-African population structure with asMDS analyses on European, South Asian and East and Southeast Asian haplotypes. The European asMDS includes 5 reference populations: British from England and Scotland (GBR), Iberian populations in Spain (IBS), Toscani in Italia (TSI), Finnish in Finland (FIN), and the French. The inclusion of Dutch genomes would have been an ideal European ancestral reference; however, we were unable to obtain access of the available data for our analyses. The local ancestry prediction accuracy was the lowest for this ancestry (89.7%), with most of the model's confusion lying on misassignment of South Asian segments as European 6.9% of the time; though the $k=5$ ADMIXTURE correlation is still high ($R^2=0.963$; see *Supplemental Text and Figures 10-13*). We assess European haplotypes found among our sampled individuals and within ≠Khomani San and Nama individuals with less than 85% Khoe-San ancestry and filtered our study samples for those with at least 20% European ancestry. We find European-assigned haplotypes in our sampled populations cluster within and between the British and French populations (*Supplemental Figures 2.16A and 20*), though population centroids are closer to the French (*Supplementary Figure 2.20 and Table 7*). This pattern is concordant with historical migration from these regions into South Africa.⁷⁹

The South Asian asMDS is less specific regarding the putative source populations (*Supplemental Figure 2.16B*). We used 5 reference populations: the Bengali in Bangladesh (BEB), Indian Telugu in the UK (ITU), Punjabi in Lahore, Pakistan (PJL), Gujarati Indians in Houston, Texas, USA (GIH), and Sri Lankan Tamil in the UK (STU). As with the European ancestry analysis, we included haplotypes from ≠Khomani San and Nama individuals with less than 85% Khoe-San ancestry and filtered our samples for those with at least 20% South Asian

ancestry. We observe a regional cline from North-West to South-East along MDS1, with the Gujarati (western India) and Punjabi (Pakistan) anchoring one end and the Bengali, Telugu, and Tamil anchoring the other. Along MDS2 the Bengali separate from the Telugu and Tamil. Haplotypes found within the Cederberg, Karoo, and Orange River participants generally pulled between the Bengali, Gujarati, and Punjabi but are generally scattered, with some clustering towards the Telugu and Tamil (*Supplemental Figure 2.16B*). Though we are unable to definitively state which source populations among these references best captures the South African haplotypes, there is a noticeable increase of South Asian haplotypes derived from Cederberg samples compared to our other sampled populations likely reflecting their closer proximity to Cape Town (*Supplemental Figure 2.16B*).

Included in the East and Southeast Asian asMDS analysis are the Han Chinese and Vietnamese from 1000Genomes alongside 130 Indonesians (Figure 2.4C). The East and Southeast Asian ancestry comprises a small fraction of the study population's overall genetic composition, reaching 4.7% in the CDB and as low as 1.4% in the ORG. Given the region's slave trade and historical record, we hypothesize Indonesians would best represent this component found among SAC populations. Our LAI model's accuracy at assigning this ancestry is 96.3% and correlates well with reported $k=5$ ADMIXTURE proportions ($R^2=0.943$) (*Supplemental Figures 2.10-13*).

We refine previous research, which demonstrated SAC individuals from the Western Cape shift towards Indonesians,⁶⁶ by implementing an haplotype-based analysis using Indonesian genomes sampled across 12 islands and a larger, more diverse SAC cohort.^{62,66} The first dimension separates eastern Indonesian islanders from mainland East Asians, while the second dimension captures previously observed regional variation within Indonesia,

distinguishing western from eastern islanders.⁸⁰ We find Indonesians are more accurate proxies for the East and Southeast Asian components by observing clear haplotype clustering from the study populations near Indonesian samples as opposed to the Han Chinese or Vietnamese (Figure 2.4C). Here, we observe more specific clustering near western islands Sulawesi, Sumatra, and Java, with some trending towards Flores near regions that correspond with major ports during the Dutch East Indian Ocean trade routes.^{81,82}

2.3.4 Geographic predictors of population structure

To estimate how geographic and environmental factors impact human population structure, we used the machine learning tool SPRUCE (Spatial Prediction using Random Forest to Uncover Connectivity among Environments). SPRUCE uses the migration rates and demes outputs from MAPS (Migration and Population Surface estimation) as the dependent variable in the random forest regression (*see SPRUCE methods*).⁸³ The inferred symmetric migration rate (m) is converted to dispersal distance (s) by MAPS “which can be interpreted roughly as the expected distance an individual disperses in one generation”⁸³. We estimated two time periods, approximately 56-13 generations ago (individual IBD segments 2-6cM) and less than 13 generations ago (individual IBD segments greater than 6cM).⁸⁴ The migration surfaces were similar for the two time periods. In Figure 2.5C, corresponding to ~13 generations ago, we see there are slightly more barriers to migration in the Cederberg and relatively more migration within the Greater Karoo compared to 56-13 generations ago (Figure 2.5A). Random forest regression indicates that environmental variables remain similar across time periods, though some variation exists. The most important variables explaining variation in migration rates are mean temperature, slope, and minimum temperature during the coldest month (Figure 2.5B, 2.5D, *Supplemental Figure 2.27*).

To better understand the demographic connections in relation to geography, we further investigated IBD sharing within and between communities using GERMLINE2 (*see Methods*). We see higher mean shared IBD segments within communities than between, with the Nama having the highest mean shared IBD (Figure 2.6; *Supplemental Figure 2.29*). The IBD network shows patterns of sharing between communities that are consistent with the expectation that geographically closer communities are more likely to share IBD than more distant groups. However, this relationship is relatively weak, suggesting that other social and geographic factors may play a role.

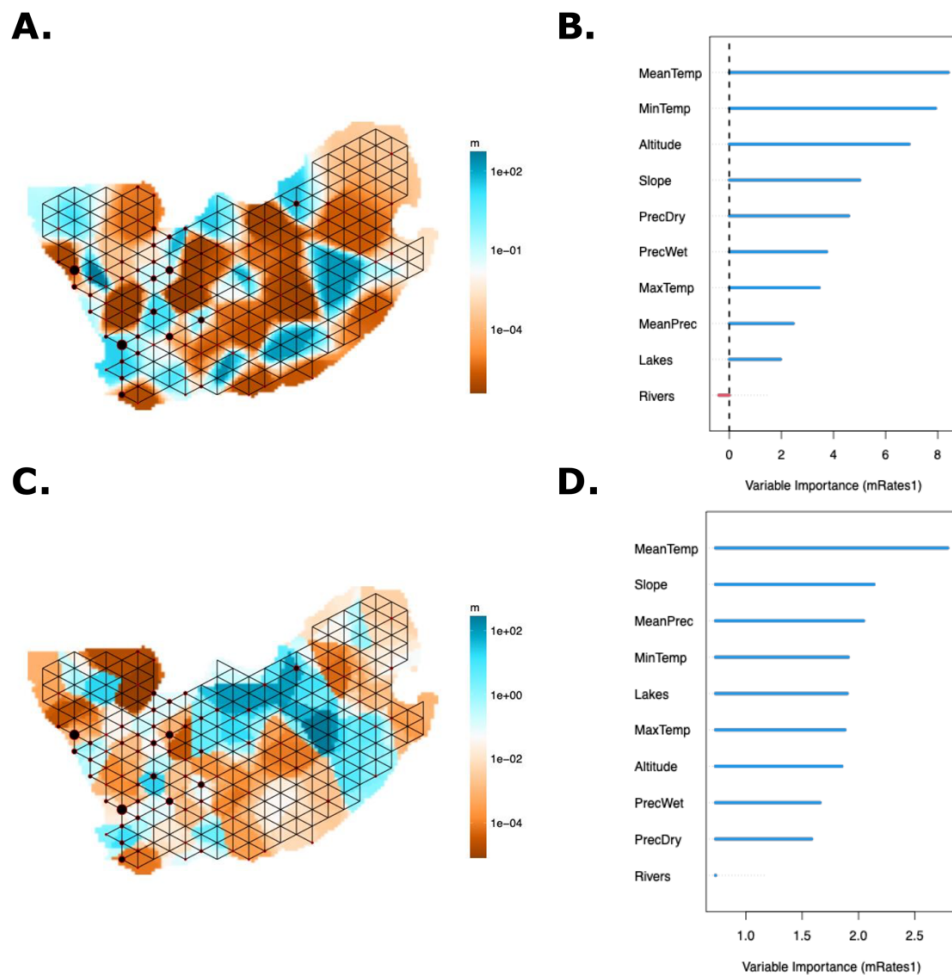


Figure 2.5) Environmental and geographical factors affect population structure. In both panels A and C, the blue color represents higher migration rates and fewer obstacles to migration; the orange color represents barriers to migration and a lower migration rate. **A)** MAPS results showing estimated migration rates for IBD segments 2-6cM representing 56-13 generations ago. **B)** SPRUCE Random Forest importance table with mean annual temperature and minimum temperature during the coldest month being the most important (Pearson Correlation $R = 0.259$). **C)** MAPS results showing estimated migration rates for IBD segments greater than 6cM representing less than 13 generations ago. **D)** SPRUCE Random Forest importance table with mean annual temperature and slope being the most important variables explaining the variation in estimated migration rate (Pearson Correlation $R = 0.387$).

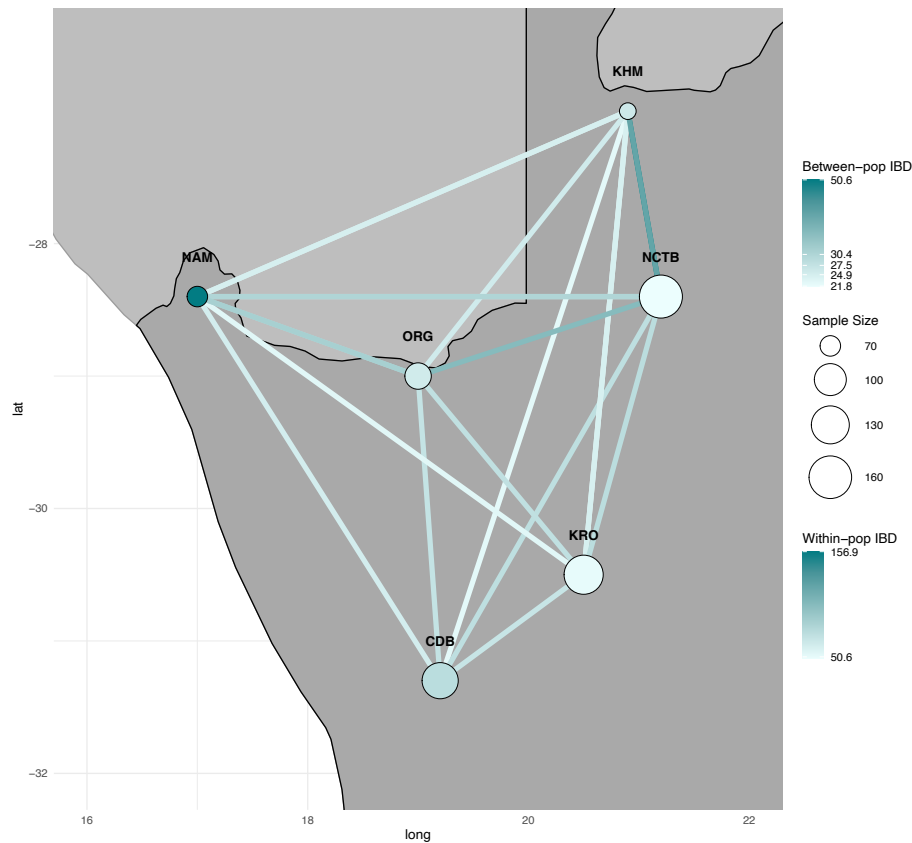


Figure 2.6) IBD sharing indicates low levels of migration across communities. The nodes are located on geographical sampling coordinates from each community and are scaled by the number of individuals sampled

in that region. The nodes are colored based on within-community mean IBD sharing. The edges between groups are colored based on mean IBD sharing between communities.

2.4 Discussion

The genetic diversity of present-day South African populations has been significantly shaped by a series of complex migrations over the past two thousand years. Previous studies have predominantly focused on disentangling Khoe-San population structure and understanding deep human evolutionary history with relatively less attention given to colonial impact and population dynamics.^{5,60,85–87} There is limited knowledge of how colonization impacted the genetic ancestry of Indigenous populations in South Africa, largely due to the systematic displacement of these populations during and after the colonial period. Today, few individuals self-identify as members of South Africa's Indigenous Khoe and San groups.

Previous studies have examined the complex ancestry components of SAC individuals in Cape Town and communities along the southwestern coast.^{58,73} Based on a large-scale study of over 600 participants, we provide a comprehensive genetic analysis of SAC populations outside of Cape Town (n=620) and characterize genetic composition across multiple regions of South Africa (Cederberg, Karoo, Upington, Middle Orange River; Figure 2.1, Figure 2.2C). Our data demonstrates that the SAC population in the northwest of the country is shaped by multiple sources of gene flow, consistent with prior findings in Cape Town and the surrounding areas.^{5,23,60,88,89} European colonization introduced not only European ancestry but also the forced migration of enslaved peoples through the Dutch East India (VOC) slave trade. This brought additional ancestry components into Indigenous Khoe-San populations, including gene flow from western and eastern Africa, South Asia, Malayasia, Madagascar, and Indonesia (*Supplemental Figure 2.6*).^{5,23,58,60,73,88,89} Using ancestry-specific multidimensional scaling

(asMDS) we were able to differentiate between putative ancestral source populations for the five ancestries commonly found within SAC communities^{88,89} and identify better proxies for these sources.⁵

2.4.1 Western African ancestry reflects both ancient Bantu expansion and colonial slave trade

Our analysis reveals distinct patterns of admixture timing and intensity across the Cederberg, Karoo, and Middle Orange River/Upington areas. NKA ancestry varies geographically, from approximately 11% in the Cederberg to about 22% in the Karoo and along the Middle Orange River (*Supplemental Table 2*, Admixture $k=5$). This distribution likely captures signatures of the Bantu expansion, with this component accounting for roughly 80% of genetic composition among southern Bantu-speaking Zulu, or it may represent NKA lineages introduced by the VOC's slave trade.

Previous studies have found the NKA admixture into the Khoe-speaking hunter-gathers and Kxa-speaking hunter-gatherers to vary in timing, ranging from as early as 1.2k to 1.5kya.²⁰ The Richtersveld Nama show a pulse of NKA gene flow at 700 years ago. These dates suggest much of the gene flow observed in the northern Khoe-San populations dates to the initial wave of Bantu expansion across the Kalahari.^{67,68} However, another wave of NKA ancestry was introduced to the Cape Colony as the VOC forcibly relocated enslaved individuals from Madagascar, Mozambique, and along the Swahili Coast.⁵² These multiple waves of admixture are supported by tract length patterns in the Cederberg Mountains. The decay pattern of NKA tracts resembles that observed for South Asian ancestry, both suggesting colonial-era admixture

events. This contrasts with tract length profiles in Karoo and Middle Orange River participants, which indicate older gene flow (*Supplemental Figure 2.14*).

After performing ancestry-specific MDS, the NKA component often closely resembles southeastern Bantu-speaking groups rather than southwestern Bantu populations (Figure 2.2). In the Karoo town of Carnarvon, this ancestry is consistent with self-reported Xhosa parents or grandparents among study participants. Conversely, the Nama and Middle Orange River participants have higher proportions of southwestern Bantu ancestry (~5.6% and 9.5%, respectively) relative to other sampled populations ($\leq 3.4\%$). This pattern positions their population centroids closer to the Himba rather than western African reference populations in NKA asMDS analysis (Figure 2.2C and *Supplemental Figure 2.19*). Ethnographic data from participants sampled along the Middle Orange River frequently report being “Nama-Dama” or having a Damara ancestor. The Damara are a historically mobile population who today reside in the semi-arid regions of Namibia and Northern Cape, South Africa.⁶⁵ Although the Damara are not included in our study, the closely related Himba serve as the best available proxy to reference the southwestern Bantu-expansion ancestral component.²¹

2.4.2 Indonesian-ancestry and other Inter-Continental Migration

Initially established as a refueling station in 1652, the Cape Colony (now known as Cape Town) became the gateway for intercontinental migration into southern Africa. While many of the early colonists were Dutch, migrants from other western European regions also became part of the Cape Colony. Protestant refugees (known as the French Huguenots) arrived in the late 17th century fleeing religious persecution. They lived in South Africa under Dutch rule in areas like Franschhoek and Drakenstein Valley in the mountainous region to the northeast of Cape

Town.⁹⁰ The British gained control of the colony from the Dutch in 1806.^{91,92} Reflecting the European colonial history at the Cape, our study populations retain European haplotypes intermediate to French and British populations in asMDS analysis, with population centroids positioned closer to the French (*Supplemental Figure 2.20*). Although French and British populations serve as reasonable proxies, a limitation of our study is the lack of access to Dutch reference genomes, which would have been most appropriate given Dutch colonial history.

The VOC's extensive Indian Ocean trade network also facilitated the forced migration of enslaved populations from Asia to the Cape Colony.⁹³ Distinguishing among Asian sources has been challenging in previous studies, potentially due to methodological constraints, limited sampling, genetic drift following initial admixture in South Africa, or a combination of these factors.^{62,66} We find a noticeably higher signal of South Asian haplotypes in the Cederberg samples compared to the other sampled locations, likely due to their proximity to Cape Town (*Supplemental Figure 2.22*).⁵⁸ Although our analysis yields ambiguous results for distinguishing among specific South Asian regional ancestries, further methodological developments could potentially refine these haplotype assignments.

In contrast to the ambiguous South Asian patterns, our analysis is able to provide further resolution to East/Southeast Asian ancestry components in the SAC. The asMDS for this component clearly point to Indonesia as the best proxy source population, rather than the mainland Asian populations (e.g. Han Chinese) typically used to capture this signal (Figure 2.4C).^{58,60,89} Inferred proportions of East/Southeast Asian ancestry in our study populations is relatively low, ranging from trace levels (mean <1%) in the ≠Khomani San to about 6% in the Cederberg, following a South-to-North gradient as seen in some other ancestry components (*Supplemental Figure 2.9B and Supplemental Table 2*). We included only those whose combined

maternal and paternal haplotype proportions reflected at least 10% Southeast Asian ancestry, which is a limited proportion of the study cohort. Additionally, input segments were required to have a minimum probability of 0.75 to be classified as East/Southeast Asian. Although the low proportion of East/Southeast Asian ancestry and the applied thresholds introduced some limitations, they also increased confidence in the results.

Detection of this Indonesian component in the Cederberg and other sampled regions aligns with historical records.^{94,95} An estimated quarter of all enslaved individuals at the Cape Colony were of Malay or Indonesian descent.⁹⁶ The clustering patterns observed in our southeastern Asian asMDS correspond with major VOC trade routes connecting South Africa to predominantly western Indonesian islands.^{81,82} Islands such as Sulawesi, Sumatra, and Timor had formalized ‘slave-clauses’ with the Dutch in the 1600s.⁷⁷ Interestingly, we see higher proportions of the South Asian component than Southeast Asian, approximately double the Southeast Asian inferred global ancestry proportions across populations. This pattern may reflect enslaved individuals from Indonesia remained concentrated in the central Cape Colony rather than dispersing northward, or that those who did migrate north, whether through forced relocation or escape, experienced demographic decline resulting in their minimal genetic contribution to the SAC populations outside of present-day Cape Town.

2.4.3 South African Coloured populations retain high Khoe-San ancestry

The VOC initially relied on livestock trade with Indigenous populations during the first 30-40 years of their occupation at the Cape Colony, but made little effort to expand or incorporate Indigenous groups into the colony.⁹⁷ However, as European colonists began expanding outward from the Cape Colony in the 17th century, they successively forced Indigenous populations into the interior, and more marginal areas such as the Kalahari and

Cederberg mountains.⁵⁵ However, colonial ambitions intensified significantly in the 1730s, leading to heightened Indigenous resistance. A frontier war erupted between the colonists and the Khoe-San living in the Cederberg in 1739,⁹⁸ followed by significant population decline due to smallpox epidemics. Historical records indicate that many surviving Khoe and San individuals were forced into servitude and employed as laborers on settler farms.⁹⁹ In some cases, these Indigenous individuals admixed with those enslaved by the VOC, forming new communities that lead to people who largely self-identify as SAC today.¹⁰⁰

The systematic displacement and marginalization of Khoe-San populations during and after the initial European colonization has led to a widespread belief in contemporary South Africa that Khoe-San ancestry is largely absent from modern populations, yet our analysis demonstrates substantial Khoe-San genetic contributions across study populations. Our results indicate that SAC populations from the Western and Northern Cape Provinces retain a substantial proportion of their genetic ancestry from Khoe-San populations, averaging 41–53%. Haplotype-based analysis of Khoe-San segments (with >75% posterior probability) finds this component to be more similar to the southern Kalahari ancestry -- represented by the *≠Khomani* San and Nama -- rather than the northern and central Kalahari, represented in this study by *!Xoo* and *Ju|’hoansi* (Figure 2.4). This result corresponds with the sampling sites’ geographic location south of the Kalahari (Figure 2.1 and 2C).²¹ Previous studies have shown a clear north-south structure in Khoe-San genetic diversity among modern populations, with more recent work suggesting an even more geographically complex pattern, reflective of the Kalahari Desert as a barrier to gene flow.^{5,20,27,28,101} Interestingly, the population centroids of the Karoo and Orange River in our asMDS results are closer to the *≠Khomani* San than the Nama (*Supplemental Figure 2.18* and *Supplemental Table 5*), with the Cederberg situated intermediate.

We hypothesize the observed patterns could result from one or more of the following: 1) European disruption led to widespread displacement and increased gene flow among Khoe and San populations south of the Kalahari, resulting in a Khoe-San ancestry signature generally intermediate between the $\neq$ Khomani San and Nama. 2) A gradient of ancestry between the Nama and $\neq$ Khomani could reflect a demic diffusion of Khoekhoe pastoralists both along the western coast and along the Orange River into the interior. If the diffusion incorporated local hunter-gatherer populations, rather than with replacement, (¹⁰² such a process could fit the asMDS. 3) Finally, the Richtersveld Nama have experienced a severe post-colonial bottleneck; it is possible their extreme MDS values represent localized drift separating them from similar groups (such as Cederberg Khoe-San). Future research collecting additional samples from individuals in Namibia, and formal demographic modeling is necessary to address these hypotheses.

2.4.4 Colonial Impact on Ecogeographic Population Structure

Colonial-era processes, centered at Cape Town and extending outward through systematic displacement and the VOC trade network, have created spatial gradients in the distribution of the five main ancestry components. Using ADMIXTURE $k=5$ results, we considered how these ancestry proportions varied across sampling regions. We observe a gradient of increasing Khoe-San ancestry with increased distance from Cape Town, following colonial patterns of systematic displacement of the Khoe and San groups, (⁵⁵ but observe the opposite in European, South Asian and southeastern Asian ancestries (Figure 2.3 and *Supplemental Figure 2.9*). We find higher proportions of European and Asian ancestries in individuals closer to Cape Town, likely because Cape Town was the initial heart of the settlement.^{52,58} Unlike the other ancestries, the proportion of non-Khoe-San African ancestry has

no relation with the sampling site's distance from Cape Town (*Supplemental Figure 2.9*), reflective of the demographic history mentioned above.

In addition to these geographic gradients of ancestry, our results reveal how European colonization fundamentally altered population movement and structure across the landscape. Using effective migration rate analysis of shared identical-by-descent tracts among our study populations, we compared two distinct time periods: pre-colonial (approximately 450 CE to 1650 CE; 56-13 generations ago) and colonial (1650 CE to present; <13 generations ago).^{96,103} During the pre-colonial period, we observe higher migration rates along the Cederberg and within the arid Greater Karoo interior, with natural barriers to movement between these regions (Figures 1 and 5A). The Cederberg's rugged landscape may have constrained population movement, while the comparatively flatter semi-arid Karoo allowed for greater mobility within the region.

Colonial contact dramatically altered these patterns, resulting in an overall decrease in migration rates (Figure 2.5C). European colonization systematically restricted Indigenous movement through natural resource monopolization, policies limiting individual mobility, and military outposts protecting colonial boundaries.^{55,104} Colonial control of water resources became a particularly powerful tool for controlling Indigenous populations, limiting Khoe-San movement and undermining pastoralist autonomy.⁵⁵ This shift is reflected in our environmental modeling: water resources show low importance in predicting pre-colonial migration rates, but the presence of lakes becomes significantly more important post-contact, potentially reflecting increased sedentism during the colonial period (Figure 2.5B, Figure 2.5D).⁹⁹

As the British took control of the Cape Colony from the Dutch in the early 19th century, land became property of the Crown and Indigenous peoples were viewed as obstacles to colonial

success.¹⁰⁵ Legal restrictions further constrained mobility— the colonial government's 1809 law requiring Khoe-San individuals to adopt sedentism institutionalized the restriction of movement and facilitated labor control.¹⁰⁶ These policies systematically dismantled the traditionally mobile lifestyles of Khoe and San populations, forcing cultural assimilation as European colonizers imposed colonial norms while devaluing Indigenous lifeways as "primitive." The persistence of this disruption is evident today, as most Khoe-San descendant individuals speak Afrikaans and English rather than traditional click-phoneme languages.²¹

2.5 Conclusion

This study provides the first large-scale characterization of genetic composition across multiple regions in South Africa revealing that Khoe-San ancestry constitutes a significant proportion of ancestry among SAC populations (41–53%) from Cederberg, Great Karoo, and Middle Orange River. Despite historical colonial disruption, our results show that the Khoe-San descendant communities retained a significant portion of Khoe-San genetic ancestry. Utilizing local ancestry assignment, we are able to further pinpoint possible sources for colonial-era migration into South Africa. For example, we identify Indonesia as the likely source population for the previous “East Asian” genetic component, further supporting the historical record that a fourth of all enslaved individuals at the Cape Colony were of Malay and Indonesian descent. This ancestry remains a small minority in our sampled individuals, however. The substantial Khoe-San ancestry persisting in modern populations, the geographic gradients reflecting colonial settlement patterns, and the disrupted mobility patterns evident in contemporary genetic structure all demonstrate how colonial processes fundamentally altered the demographic landscape of South Africa. We emphasize that genetic signatures alone do not fully capture the complex power dynamics that shaped these demographic changes. Understanding the genetic legacy of

colonization requires recognizing both what these patterns reveal about historical processes and the limitations of genetic data in capturing the full scope of colonial impact.

2.6 Methods

2.6.1 Sample Collection

Genetic and demographic data from several towns in the Cederberg, Karoo, and Orange River were newly collected between 2017 and 2023; locations are indicated in Figure 2.1. DNA from saliva samples were collected with Oragene OGR-600 kits (DNA Genotek) and ethnographic information regarding ancestry, language, and parental place of birth were collected for all participants. Institutional review board (IRB University of California - Davis (1220180-7) and the Health Research Ethics Committee of Stellenbosch University (N11/07/210 and N11/07/210A). All individuals gave signed written and verbal consent with a community witness present before participating. Participants were asked to self-identify without categorical prompts. Prior to sampling, municipal or non-profit leaders in each community were consulted on the aims and approach of the project. Community level results were returned to the majority of towns between 2023-2025, including community meetings, pamphlet and poster distribution paired with discussions involving local municipalities, high schools, and community members.

2.6.2 Sample Consolidation, Genotyping, and PCA

Newly collected samples were genotyped for >2.2million SNPs using the Illumina H3Africa array. Raw genotype data was processed using Illumina's GenomeStudio to call common variants (MAF>0.05) on the autosomes, X- and Y-chromosomes, followed by zCall to call rare variants. Details of this pipeline are available at <https://github.com/hennlab/snake->

SNP_QC. All datasets were aligned to the 1000 Genomes Phase 3 reference panel and all ambiguous variants were removed.

We consolidated a reference panel incorporating data from four published genome sequencing projects, the African Genome Variation Project¹⁰⁷ (AGVP), Human Genome Diversity Project¹⁰⁸ (HGDP), Indonesian Diversity Genome Project¹⁰⁹ (IDGP), and 1000 Genomes Project¹¹⁰ (1000GP), an unpublished cohort of 94 ≠Khomani San (KHM) whole genome sequences [Soto et al., in preparation], and previously published African samples¹¹¹ genotyped using the Omni5 array.

We extracted populations of interest from each dataset listed above. The Zulu, Omoro, Amhara, Wolayta, Somali, Nama (NAM), and Gumuz were extracted from the AGVP¹⁰⁷. Indonesian samples were sourced from the IDGP¹⁰⁹. The British (GBR), Iberians from Spain (IBS), Han Chinese (CHB), Vietnamese (KHV), Yoruba (YRI), Luhya (LWH), Bengali (BEB), Tamil (STU), and Gujarati (GIH) were obtained from 1000GP¹¹⁰. Lastly, the French were obtained from HGDP and the Botswanan San from previously genotyped African samples.¹¹¹ The Botswanan San include two San populations, the Ju|'hoansi and !Xoo. However, the published dataset¹¹¹ did not include population IDs, leaving us to infer these IDs using a combination of methods further detailed the supplement (*see Supplemental Text and Figures 1–4*). The Botswanan San samples were processed before the lifting or merging with the larger dataset in order to maximize the variants used in the analysis. This left the Botswanan San on hg19 with 2,993,754 variants, and the NAM, KHM, ORG, CDB, NCTB, and KRO aligned to hg38 with 1,015,964 markers. Sample sizes for each population can be found in supplemental table 1.

Datasets were filtered for biallelic variants congruent with markers assayed using Illumina's H3Africa array. With the exception of the included 1000GP and HGDP populations, each dataset was lifted from hg19 to hg38 using Picard Tools LiftoverVCF.¹¹² Samples were then merged with the newly collected cohort of South African samples genotyped on H3Africa.¹¹³ The newly collected samples include SAC individuals sampled along the Cederberg Mountains (CDB), Karoo (KRO), Orange River (ORG), as well as Northern Cape Tuberculosis Project (NCTB). The merged data was filtered for Hardy Weinberg Equilibrium (<0.05), minor allele frequency (<0.00047 to remove singletons and doubletons) and individual and SNP missingness ($>5\%$) using plink (v1.9). The compiled dataset contained 806,460 autosomal single nucleotide polymorphisms (SNP) and 1,165 individuals. Merging was assessed using PCA's 1-10, with PC loadings generated through smartPCA⁶⁹. VCF files were converted to `.ped` and `.map` files for eigensoft smartPCA input.

2.6.3 Phasing

The dataset was phased using the program Shapeit4 (v4.2.2)¹¹⁴ and the Consortium on Asthma among African Ancestry Populations (CAAPA2) reference panel (unpublished data). This reference panel includes 78 $\neq$ Khomani San and 49 Nama individuals, in addition to other continental African populations, making it the most representative reference panel available for the individuals sampled in this study. All Shapeit4 parameters were set to default, with the exception of `--mcmc-iterations` being set to `10b,1p,1b,1p,1b,1p,1b,1p,10m`.

2.6.4 Identifying Relatives

We used Parent Offspring Pedigree Inference Robust to Endogamy (PONDEROSA), a software designed to identify relatives within more endogamous populations and/or populations

that have undergone recent bottlenecks, resulting in elevated levels of shared identical-by-decent (IBD) regions.¹¹⁵ The program provides an option to simulate related individuals from the population for various relationships classes (i.e. maternal/paternal half-siblings, maternal/paternal grandparents, avuncular relationships). It accomplishes this using integrating the program ped-sim.^{116–118} The IBD distributions for the simulated relationships are then used to train the model that will be applied on the real data.

PONDEROSA was independently preformed on each of the southern African populations, including the newly collected SAC individuals. We segregated the ≠Khomani San, Nama, CDB, KRO, ORG, and NCTB samples by region/population and proceeded to run simulations on each of the 6 southern African groups included in this study.

With the exception of the Botswanan San, samples were lifted from hg38 to hg19, without variant loss, in order to use the PONDEROSA this required a sex-recombination map aligned to hg19.^{116–118} IBD calling was then carried out using PONDEROSA's built-in script initiating phasedIBD¹¹⁹, all settings set to default. We simulate 200 pairs of individuals for each population and then use the training classifiers to identify relatives in the populations. Relatives having a greater than 80% probability were assigned a specific class, while those below the threshold were assigned either 2nd, 3rd, or 4th degree relatedness. Finally, we used PONDEROSA 'remove_relatives.py' script to filter individuals who were related closer than 3rd degree. This script does not remove all relatives but rather retains one of the related pair of individuals, in order to maximize the number of unrelated individuals. For example, if a sibling pair existed in one of the populations, the function would retain one of the siblings in the unrelated subset of samples, while placing the other in the relative's subset.

2.6.5 ADMIXTURE

Prior to running ADMIXTURE, we first identified all first- and second-degree relatives using KING¹²⁰ and partitioned relatives into multiple non-overlapping “running groups”. We then performed separate ADMIXTURE analyses for each running group. ADMIXTURE assumes that all samples are unrelated because shared IBD tracts between close relatives can bias allele-frequency estimates and consequently inflate or deflate inferred ancestry proportions. Excluding close kin from each run ensures unbiased allele-frequency estimates, while systematically rotating relatives across groups guarantees that every individual contributes to at least one model fit.

The three Khoe-San reference populations were independently assessed for relatedness using PONDEROSA (as above). An internal R script was utilized to generate: 1) an unrelated subset of individuals and 2) randomized groups of first- and second-degree relatives while ensuring minimal overlap of related individuals across groups 3) a list of unrelated individuals in the dataset. A total of 10 groups were created to divide relatives; each then merged with the larger list of unrelated individuals. Corresponding PLINK files were generated for each of the final 10 running groups (RG).

The three southern African reference populations underwent a similar process independently of one another. However, instead of partitioning all first- and second-degree relatives into different running groups, we retained one individual from each related pair in the unrelated subset using PONDEROSA `remove\_relatives.py` script. Individuals classified as relatives were divided into 10 randomized groups and added to their respective RG PLINK file with the. This method allows us to retain a maximal number of Khoe-San individuals in the core,

an unrelated list of individuals shared across the RG, while the related individuals are randomly subset across the 10 running groups with little overlap in relatives. A flowchart of this pipeline is included in Supplemental Figure 2.5.

European and Asian populations were thinned to only include the GBR, IBS, GIH, BEB, SES, CHB, KHV, and Indonesians in admixture. Linkage disequilibrium (LD) was calculated on the merged dataset of 806,460 markers, including the reference populations and newly genotyped southern African samples using PLINK 1.9¹²¹. Data were conservatively pruned using the PLINK --indep-pairwise flag for markers with LD of 0.5 or greater within a window size of 50 SNP, with sliding window size of 20 SNPs. The pruned dataset contained 479,497 variants. Ten iterations of ADMIXTURE $k=4:10$, k being the number of clusters ADMIXTURE infers, were performed for each running group, all with fivefold cross-validation and randomly selected starting seed. With respect to the level of k , results for each running group were averaged across individuals. ADMIXTURE results were visualized using pong¹²² (Figure 2.2A, Supplement Figures 6 and 7).

2.6.6 Local Ancestry Inference

Local ancestry (LAI) was inferred using Gnomix⁷⁵, which has been shown to outperform other programs in accuracy and speed. Similar to other LAI programs, Gnomix treats each reference population as an independent ancestral group, without considering the relationships between populations. In other words, the individuals within each ancestral group must have a high ancestry proportion with respect to the ancestral group they are representing.

This was relatively simple requirement to meet for most of the references, but particularly difficult when using reference populations who are admixed, such as the southern African

reference individuals. Based on the ADMIXTURE results for $k=6$, we filtered for $\neq$ Khomani and Nama individuals having at least 85% Khoe-San ancestry, and the Botswanan San for minimum 90% Khoe-San ancestry. We chose to use varying cut-offs to retain the maximum number of individuals while accounting for within Khoe-San population structure and differences in sample sizes. Southern African individuals below the cut-off were then added, in addition to the Zulu as a control.

Given that the SAC populations are often described as five-way admixed, we assembled a LAI reference panel that accounts for each putative ancestry. Putative source populations for the KHS components in our study samples include the $\neq$ Khomani San, Nama, and the Ju|'hoansi and !Xoo from Botswana. Putative source populations for European ancestry include GBR and IBS (EUR), for South Asian ancestry include BEB, GIH, and STU (SAS), and for East/Southeast Asian ancestry included CHB, KHV, and Indonesians (EAS). We separated the KHS populations, that is the Nama, $\neq$ Khomani San, and Botswanan San, from other Africans and aggregated all NKA populations into an independent category. We opted to use this method because 1) the Khoe-San are highly diverged from other NKA populations, providing confidence that we can distinguish between Khoe-San and NKA haplotypes, and 2) the study samples likely have NKA ancestry from various source populations, making it difficult to discriminate between the putative populations. Populations included in NKA are the Yoruba, Luhya, and Gumuz. The Omoro, Amhara, Wolayta, and Somali individuals were removed from the NKA reference panel due to the back-to-Africa¹²³ ancestry which weakened the model's assignment of EUR and NKA segments. In summary, our five-binned local ancestry reference panel had populations representing EUR, SAS, EAS, KHS, and NKA ancestry.

Gnomix⁷⁵ was executed using the recommended array config, a logistic regression base and xgboost smoother modules. All settings were set to default, with the exception of adjusting for larger window (0.4cM) and smoother size (82). We assessed multiple models and found this model to have the most accuracy and to be more stable against phase switching due to the increased window and smoother size. These changes reduced sensitivity with decreased window size but increased model accuracy. Local ancestry results were evaluated using five methods: 1) a confusion matrix, 2) ADMIXTURE and Gnomix global ancestry correlations, 3) track lengths, 4) probability distribution for each ancestry, and 5) visualizing karyograms. See supplement for more detail on local ancestry assessment.

2.6.7 Ancestry-specific MDS

We used multi-array ancestry specific multidimensional scaling (maaMDS)⁷⁴ to evaluate whether the projected ancestry-specific haplotypes produced by GNOMIX in our southern African study samples would gravitate towards one or more putative source populations relative to others. This method calculates the dissimilarities among the assigned reference populations, then incorporates ancestry-specific haplotypes from the study population by their distances relative to the existing reference clusters, effectively projecting them onto the pre-defined coordinate axes. The analysis was carried out using average pairwise genetic distances, including only haplotypes with at least 75% probability of the target ancestry. Each ancestry had an independent maaMDS run where only the respective putative reference populations were specified. For example, when running the non-Khoe-San African (NKA) maaMDS, the KHS, EUR, SAS, and EAS populations were removed from the analysis. The flag 'IS_WEIGHTED' was set to true. Reference populations were given a weight of '1', while study samples were

assigned ‘0.01’ in order to ensure they are projected within the MDS space of the reference haplotypes. Visualization of the output data was done using the ggplot2¹²⁴ library in R.

2.6.8 Euclidean Distance Computation

Using the coordinates from the asMDS1 and asMDS2, we calculated the Euclidean distances between population averages distribution. Individuals were assigned to populations based on predefined labels, and for each population, the MDS1 and MDS2 centroid was calculated by computing the arithmetic mean of the individual MDS1 values. Pairwise Euclidean distances between population centroids were computed in one dimension (MDS1) using the absolute difference between the centroids. The final pairwise Euclidean distance matrix was visualized as a table. Using the gridExtra package in R, the formatted distance matrix was converted into a table with a custom theme.

2.6.9 SPRUCE

To estimate and identify potential environmental factors contributing to gene flow, we used SPRUCE (Spatial Prediction using Random Forest to Uncover Connectivity among Environments)¹²⁵, a machine learning approach to integrate genetic and environmental data, to determine which factors may be influencing population structure in individuals from Nama, Khomani, and Khoe-San descendant communities. SPRUCE utilizes the MAPS output to extract environmental values from the coordinates corresponding to the produced demes.

MAPS (Migration and Population Surface estimation) (⁸³) estimates time-resolved dispersal rates and population densities based on the number of shared long Pairwise Coalescent Segments (IPSC). Prior to calling identical by descent (IBD) segments, we performed phasing with SHAPEIT4¹¹⁴ removed first- and second-degree relatives using the output from

PONDEROSA (¹¹⁵ (see *Identifying Relatives* and *Phasing* methods). IBD segments were then called using hap-ibd v1.0¹²⁶ were repaired with the program merge-ibd-segments. For running MAPS, an IBD matrix was constructed using code provided by the SPRUCE authors and the IBD calls were split into two groups: 2-6 and greater than 6 cM. The categories were chosen based off of historical dates corresponding to pre- and post-European contact in South Africa. The mean ages of the categories are 56.25 and 12.50 generations. (¹⁰³ Following the recommendation of the SPRUCE authors, we chose to use 300 demes for higher resolution migration patterns. Migration surfaces were visualized using the R package plotmaps⁸³.

Climate data such as mean annual temperature, maximum temperature of the hottest month, minimum temperature of the coldest month, mean annual precipitation, precipitation in the wettest month, and precipitation in the driest month were sourced from the free, online repository CHELSA (Climatologies at High Resolution for Earth's Land Surface Areas)¹²⁷. Slope data was obtained from the Geomorpho90m dataset,¹²⁸ and elevation was obtained from MERIT DEM (Multi-Error-Removed Improved Terrain Digital Elevation Models)¹²⁹. Processing steps were performed using the Geospatial Data Abstraction Library (GDAL/OGR Contributors 2021) in a Bash environment. The geographic and environmental data were clipped within the geographic range of interest and the demes.txt output file from the MAPS software.

For the random forest regression, the predictor variables were the values of 10 spatial variables related to climate, ecology, and geography at each deme. All variables were numeric and continuous with the exception of lakes and rivers which were coded as “1” for a body of water being present and “0” for a body of water being not present. The response variable was calculated by using the migration rate at each deme from the “mRates.txt” file outputted by MAPS and converting them to 10^m .¹²⁵ The R package randomForestSRC was used to perform

the random forest regressions (<https://CRAN.R-project.org/package=randomForestSRC>). We performed random forest regressions for both 56-13 generations ago (IBD segments greater than 6cM) and less than 13 generations ago (IBD segments between 6 and 2 cM), which were our time intervals of interest. All other model specifications and considerations were applied as described in Pless et al. (2023).

2.6.10 Within and Between Population IBD Sharing

GERMLINE2¹³⁰ was used to call identity by descent (IBD) segments to investigate IBD sharing both within and between communities across South Africa. Phasing was performed using SHAPEIT (see *Phasing* methods). First and second-degree relatives were removed using PONDEROSA (see *Identifying Relatives* methods) and KING IBDseg. IBD segments were repaired using *merge-ibd-segments*. GERMLINE2 reports all pairwise IBD segments for each chromosome. We filtered for IBD segments greater than 6cM and calculated the sum of the IBD segment lengths for each pair of individuals to get the total length of pairwise shared IBD across the genome. We then filtered the total length of pairwise shared IBD for segments greater than 13cM. A heatmap analyzing within and between community IBD sharing was generated in R using pheatmap (version 1.0.12). The mean of the total shared IBD segments was calculated for the heatmap to look at the average shared IBD within and between communities. IBD networks were generated in R using iGraph.

Acknowledgments:

We would like to thank all the participant communities in the Western and Northern Capes for their continued trust and support. We would especially like to thank our many community research assistants and translators who assisted in data collection for the project. This

research was supported by a NIH grant R35GM133531 and R35GM133531-04S1 (to BMH), R01 AI151549 (to PJM and BMH), and NSF grant #2419417 (to SLE and AWR). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We acknowledge the UC Davis Genome Center High-Performance Computing Cluster for data analysis support. We also thank Alexander Ioannidis, Daniel Mas Montserrat, and Santiago Gerardo Medina-Muñoz for their help in assessing the local ancestry models, and Carmina Barberena-Jonas for guidance on running the ancestry-specific MDS analysis.

Chapter 3. Assessing the effect of rs2470102 (SLC24A5) on iris pigmentation in southern Africa

3.1 Abstract

We sought out to test whether the derived intronic single nucleotide variant rs2470102 (*SLC24A5*) influences iris pigmentation among southern Africans with high Khoe-San ancestry. *SLC24A5* has a direct role on pigmentation and has been shown to have pleiotropic effects on skin, hair, and eye pigmentation. The variant rs2470102 has previously been associated with variation in skin color^{131,132} and lightening brown eyes among admixed Cape Verdeans.¹³³ We aimed to replicate the variant's association with eye color using a candidate gene study among. We sampled paired genotype and phenotype data from 262 individuals in South Africa and genotyped the data on the H3Africa Illumina array. Difficulties during fieldwork required conservative iris photos quality control cut offs. After genetic data and iris photos were quality controlled, we found most of the iris photos to be unreliable and the analysis to be largely underpowered, leaving us unable to detect a strong signal at the rs2470102 locus. We assess whether raw images can detect a signal, and discuss the field work complications that arose, our attempts at assessing a signal at the locus of interest, and our recommendations for future iris pigmentations studies when collecting phenotypic data in the field.

3.2 Introduction

3.2.1 Variation and Evolution of Iris Pigmentation

Similar to skin pigmentation, iris pigmentation is a polygenic trait that is largely based on an individual's genetic inheritance (98%)¹³⁴. The concentration and ratio of eumelanin to pheomelanin present in irises are responsible for the phenotypic color observed. For example, dark hair, skin, and eyes have been observed to contain higher quantities of eumelanin than pheomelanin, while the opposite is true with light hair, skin, and eyes (citation). Brown eyes make up the most common eye color across human populations (79%), while blue, hazel, and green eye color account for 8-10%, 5%, and 2%, respectively¹³⁵.

There are several genes that contribute to iris pigmentation variation, such as *TYR*, *TYRP1*, *SLC45A2*, *IRF4*¹³⁶⁻¹³⁸, and our gene of interest *SLC24A5*¹³³, but arguably the most notable region is *OCA2-HERC2*. *OCA2*, or ocular albinism 2, is located in 15q12–q13 and is believed to encode for a transmembrane protein responsible for the transportation of tyrosine, an essential amino acid used in melanin synthesis. Variants in *OCA2* can result in green/hazel eye coloring^{139,140}. *HERC2* is located about 10kb upstream of *OCA2* and codes for a protein that regulates *OCA2* expression. The rs12913832 variant within *HERC2*'s 86th intron is accredited to be the discriminating factor between brown and blue eyes, explaining about 74% of the phenotypic variation^{136,139,141}.

There have been several proposed hypotheses that attempt to explain eye color variation, in particular blue eye color, which interestingly increases in frequency with latitude¹⁴². These hypotheses include sexual selection^{143,144}, adaptation to low light^{145,146}, and hitchhiking^{142,143}. Previous research has found an excess of homozygous *HERC2* (rs12913832) among contemporary

and ancient population ^{143,147}, which could provide supporting evidence for the sexual selection and adaptation to low light conditions hypotheses.

Alternatively, there may not be a fitness advantage to lighter colored eyes. It seems that the protective layer of melanin in the iris derives from the iris pigment epithelium, the posterior layer of the iris. This layer is heavily pigmented in all individuals, regardless of the shade of their iris ¹⁴⁸. Rather, under the hitchhiking hypothesis, the pleiotropic nature of some pigmentation genes that contribute to eye and skin color hitchhiked along with the selection for skin color, resulting in more variation in eye color among populations at higher latitudes ¹⁴³. For example, *HERC2* (rs12913832) has been shown to have pleiotropic effects on eye, hair, and skin pigmentation ^{149–151}. This hypothesis is further supported by the many genes that contribute to skin color also exhibiting a pleiotropic effect on eye pigmentation, such as *SLC24A5* ¹³³, *OCA2-HERC2* region^{149–151}, *SLC45A2*, and *TYRP1*.

3.2.2 *SLC24A5* lightening allele's introduction into southern Africa

SLC24A5 is a well characterized pigmentation gene that codes for a potassium-dependent sodium/calcium exchanging protein, *NCKX5* ¹⁵². The derived allele, *SLC24A5**A, contains a non-synonymous variant in the third exon, rs1426654, which disrupts *NCKX5*'s normal function in melanosomes, resulting in lighter skin in humans and model organisms ¹⁵². In complete linkage disequilibrium (LD) with rs1426654 is a variant in the eighth intron of *SLC24A5*, rs2470102. This single nucleotide variant has previously been found to have an independent effect on skin color among populations from the Indo-Gangetic Plain ¹³² and Brazilians ¹³¹, as well as an association with lightening brown eyes in highly admixed Cape Verdean populations ¹³³. The *SLC24A5* lightening allele is fixed in Near Eastern and European populations ^{152–155} and explains about 30% of the skin pigmentation variation observed between West Africans and Europeans ¹⁵².

SLC24A5 lightening allele is present at about 41% frequency in our study population. Considering this allele coalesces with its ancestral counterpart sometime around 25kya¹⁵⁶, after the initial out-of-Africa migration, this is an unusually high frequency to have resulted solely from European admixture within the last 400 years. Given the complex demographic history of southern Africa, Lin et al. (2017) sought to test whether selection is playing a part in the allele's high frequency by modeling a series of demographic events that can be broken down into 3 parts: (1) back-to-Africa migration around 5-10kya, (2) East African pastoralists migration to southern Africa about 2-3kya, and (3) Europeans colonialist migrating into southern Africa around 400ya. Under the null model, where Europeans are the sole source of *SLC24A5**A's introduction into southern Africa, the selection coefficient is unreasonably high ($s=0.32$). Lin et al. (2017) concluded the most likely scenario explaining *SLC24A5**A's high frequency is by introduction from East African pastoralists ($s=0.04$). Under this hypothesis, East African pastoralists acquired the allele during the back-to-Africa migration (5-10kya) and then introduced it to the Khoe-San population around 2-3kya, after their initial migration to southern Africa. A second pulse of this allele was then contributed by Europeans during their migration into southern Africa around 400ya. Since then, this allele has been under strong, directional selection, exhibiting little heterozygosity across the haplotype.

3.3 Results

3.3.1 Candidate gene study

We focused our efforts on the entirety of chromosome 15 and the phenotypic measurements for L, lightness of the iris. We remove individuals with major arcus rings and use the remaining 255 participants and their raw L* measurements (unedited) to assess our model. A genetic related matrix was generated based on the array data (1,052,778 markers) using SNPRelate

and GENESIS. Principal component analyses (PCA) were calculated using PC-Air and were later used as covariates in our model to further correct for population structure. The best fit model includes the genetic relatedness matrix and the first 5 PC-Air vectors as covariates, although it is slightly inflated ($\lambda = 1.025$, see supplemental fig.2.4). We found it more difficult to control for structure in the data when running the same model for a^* ($\lambda = 1.31$) and b^* ($\lambda = 1.63$) measurements but continue to assess the *SLC24A5* locus since our primary interest is whether rs2470102 has an effect on lightening brown eye.

We then extracted regions between chr15:48115990 and chr15:48147672, 5kb up- and down-stream of *SLC24A5*, and use emmax to assess unmodified iris pigmentation values at this locus. There are no significantly associated loci when assessing L^* using a Bonferroni correction, but we do find significant associations when using a false discovery rate (FDR) cut off of 0.05. Although this should be evaluated with caution considering the slight inflation and unaccounted for structure in the phenotypic data used in this model. Similarly, a^* and b^* association results surpass the FDR threshold but are even more inflated than the L^* model. Thus, they are reported in supplemental figure 3.5 but are not considered well calibrated models.

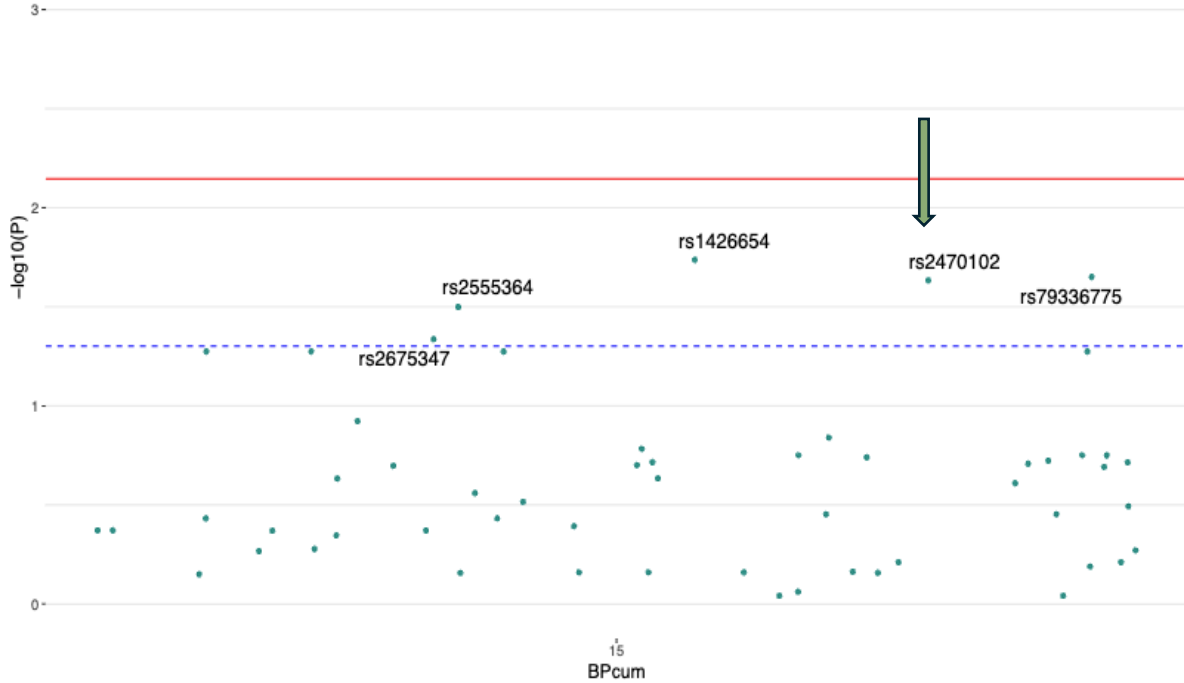


Figure 3.1) Raw L^* association results on all individuals. Red line represents the Bonferroni threshold. Blue line is the FDR (0.05) cut off. The green arrow highlights our SNP of interest, rs2470102.

3.4 Discussion

3.4.1 Fieldwork Complications

Iris pigmentation collection is efficient when restricted to a stationary setting, where a chin rest and controlled environment can provide a base-line level of consistency across image quality. Unfortunately, this can also constrain who part-take in the study, especially in the case of our study population who largely reside in remote parts of the Northern Cape, South Africa. Here, we collected samples from 267 individuals in the Northern Cape, South Africa by visiting participants at their own homes, report the challenges faced while in the field, and make suggestions for future researchers interested in iris analyses in smaller, diverse cohorts.

We then began examining the quality of the images. Despite using the Direct Flash Light laminator, which is meant to control for lighting with the Miles Eye Camera, we found it difficult to consistently control lighting during fieldwork collections. Three things might have contributed to this: 1) the lack of a chin rest and tripod to aid consistency between the iris and camera, 2) the variability introduced due to the constant reconstruction of the camera's parts while traveling to participants' homes, and 3) the variation in lighting across time and location of sampling. Evidently, we observed variation across image lighting and distance between the iris camera, with several iris photos appearing too dark. Though we attempted to correct for lighting using Adobe Light Room, it was impossible to remove subjectivity when the color chart was missing from several participants.

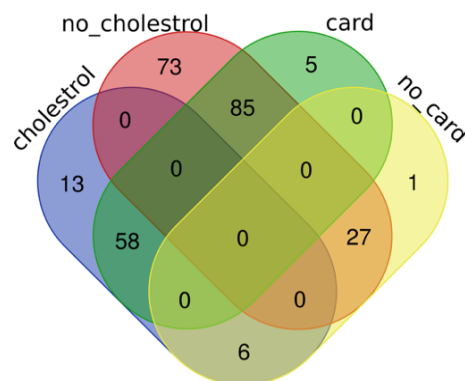


Figure 3.2) Pie chart of quality-control variables across images. In the figure “cholesterol” refers to an arcus ring.

Additionally, some individuals express an arcus ring, that is a buildup of cholesterol around the cornea which can appear in a gray/blue color. Some cases were mild and manageable to work around during processing, while other major cases left very little imaging area to process. For mild cases, we processed the outer ring around the cornea within the iris, to remove any discoloration

the arcus ring contributed. Individuals with prominent arcus rings had to be removed as their arcus ring left little room to process the iris without discoloration.

3.4.2 Fieldwork Recommendations

We found several of our mistakes could have been prevented with better fieldwork preparation. To start, an extra color chart is absolutely necessary. Our color chart went missing during sample collections and our field site in the Northern Cape, South Africa had little resources that could replenish it. Secondly, we highly recommend other researchers who are working in the field to document the conditions each image was taken in. For example, roughly documenting what time of the day an image was taken, whether it was inside or outside, and whether it was taken in a sunny, semi-shaded, or completely shaded area would be useful for image quality control. Even without the use of a color chart, we could have potentially controlled for some inflation in our model using covariates that describe the conditions at the time of imaging.

Arguably, our most important recommendation would be to include a method that controls for lighting outside of the Direct Flashlight laminator. We recommend using a box-like that would enclose the participants head from any external lighting. This would include a circular opening for the camera. This follows Beleza, Johnson and colleagues (2013) methods while collecting iris pigmentation during fieldwork. When build properly the box would also control for the distance between the iris and camera, minimizing variation that could be introduced due to the lack of a tripod.

3.5 Methods

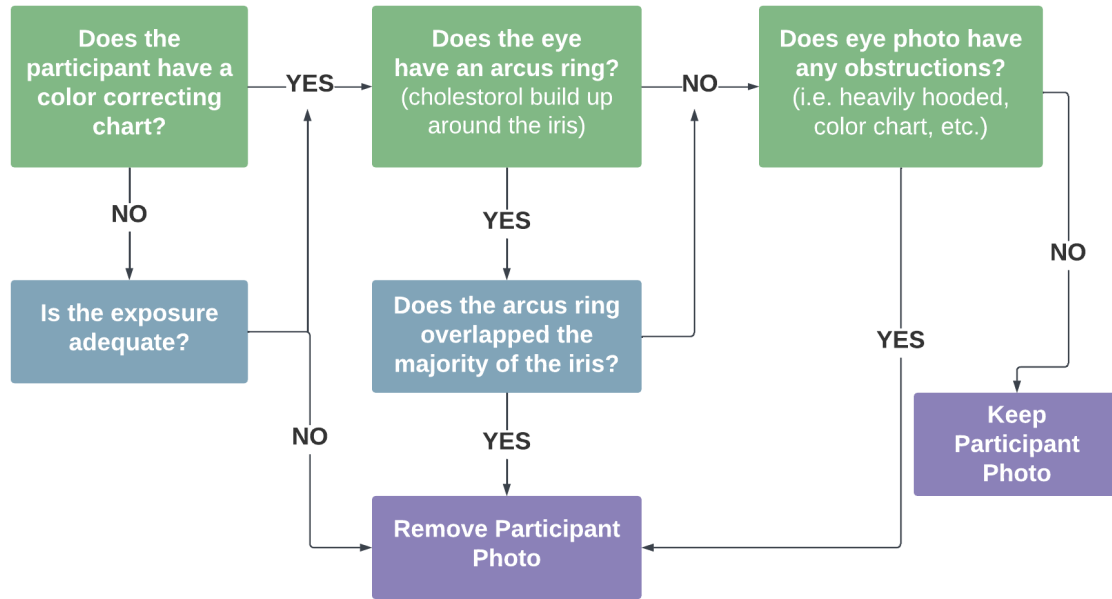


Figure 3.2) Qualitative steps taken to quality control iris photos.

3.5.1 Genetic Data

We collected 262 paired iris photos and genetic data from individuals in the Northern and Western Cape of South Africa. All participants were at least 18 years old. Genetic data was collected from saliva, extracted using Chemagic, and then genotyped on Illumina's H3Africa array. They were then processed using the Henn Lab SNP QC Pipeline on github [https://github.com/hennlab/snake-SNP_QC]. The final genotype data contained 1,804,093 variants prior to imputing. Since our focus was on *SLC24A5*, we extracted and imputed chromosome 15 (34,328 markers) using the African Genome Reference Panel through the Sanger Imputation Server [<https://www.sanger.ac.uk/tool/sanger-imputation-service/>]. The imputed chromosome contained 2,464,308 markers. Variants with less than 80% quality score were removed from the data, leaving 333,900 variants to use in our study.

Iris Color	n
dark brown	202
light brown	41
hazel	17
green	1
blue	1

Table 3.1) Sample size across iris categories.

3.5.2 Phenotype Data

All images were taken on a Nikon D3200 Miles Eye Camera using the same settings. The camera was set to photograph on RAW, using a 1/200 shutter speed, aperture setting f20, and ISO 100. Auto distortion control, noise reduction, active d lighting, VR, and ISO sensitivity control were all set to OFF. Flash control was set to manual at $\frac{1}{4}$ of a second. Picture control was set to neutral. All participants were photographed without a chin rest. When possible, photos were taken indoors and with a color chart, but several photos lack a color chart and were taken outside, though in the shade when possible. Images were processed using a web application developed by David Cha (<https://iris.davidcha.ca/>). The application guides the user through 11 steps to process and quantify iris color and features. For the purposes of this project, we skipped steps used to highlight iris surface features and focused on steps 2-6 which guides the user to

assess the best fit outline of the pupillary ruff, collarette, and iris. These outlines provide boundaries for the application use while calculating pigmentation measurements. The program returns CIE Lab color space for averaged L^* , a^* , and b^* values for each iris. These colors correspond to how light or dark ($L=100$ yields white and $L=0$ is black) an image is and where along the red to green ($-a^*$ are greener, $+a^*$ values are redder) and yellow to blue ($-b^*$ is blue, $+b^*$ is yellow) spectrum the iris sits (supplemental figure 3.1).

Chapter 4: Partitioned heritability estimates highlight genomic regions omitted by GWAS

Authors: Dana R. Al-Hindi, Aidan Baraban, Justin W. Myrick, Alexandra Surowiec, Austin W. Reynolds, Heather L. Norton, Megan Y. Dennis, Brenna M. Henn

Authors Contributions:

Dana R. Al-Hindi conducted the primary analyses for this study, including characterizing the genetic composition of newly collected southern African samples using PCA and ADMIXTURE; modeling ancestry by melanin; performing the genome-wide association study (GWAS) and post-GWAS annotations; and estimating heritability and the transferability of polygenic risk scores to the study population. Aidan Baraban led the validation efforts, including zebrafish knock-out experiments, confirmation of knock-out success, and imaging and processing pigmentation results. Justin W. Myrick, Alexandra Surowiec, Dr. Brenna M. Henn, Dr. Austin W. Reynolds, and Dana R. Al-Hindi contributed to sample collection, with the latter two overseeing quality control of newly genotyped data. Dr. Heather L. Norton, Dr. Megan Y. Dennis, and Dr. Brenna M. Henn advised on pigmentation, CRISPR/zebrafish knock-out, and Khoe-San genetics, respectively.

4.1 Abstract

Recent publications have demonstrated that the genetic architecture of human skin pigmentation variation is far more complex than previously thought, with upwards of a hundred loci contributing to the phenotype. However, it is unclear how architecture varies across populations, beyond a few large effect alleles with distinct ancestry origins. In admixed populations with recent European contribution, these large effect alleles dominate the pigmentation mapping signals. We compile a new dataset of paired skin pigmentation and genomes from southern African populations with admixed Khoe-San ancestry ($n = 1119$). Khoe-San-descent groups have lighter skin pigmentation than equatorial Africans. We find that skin pigmentation follows a south to north geographic gradient across 7 populations (i.e. lighter to darker), but after controlling for historical admixture there is little correlation between Khoe-San ancestry proportions and pigmentation. We partitioned heritability across chromosomes, and show strong enrichment on chr 15, 5, 6, 9 indicating pigmentation loci are distinctly clustered. However, standard GWAS only significantly replicates canonical SLC24A5 ($p = 9.45 \times 10^{-23}$). Functional validation of other candidate loci in zebrafish knockouts failed to verify novel genes. We show how correlation between overall ancestry and phenotype strongly attenuates power to associate loci in admixed individuals when the phenotype is highly differentiated among ancestral populations, as is the case with skin pigmentation. Improved methodological models will be needed to move beyond ‘low-hanging’ large effect alleles and fully characterize other contributors to the genetic architecture of skin pigmentation.

4.2 Introduction

Skin pigmentation is a highly heritable, polygenic trait whose phenotype largely follows an environmental gradient. The global distribution of melanin, the biopolymer responsible for pigmentation across various tissue types, is primarily shaped by variation in ultraviolet radiation (UVR) exposure across latitude,^{157,158} specifically ultraviolet A (UVA) and ultraviolet B (UVB) radiation.^{159–163} Melanin is synthesized in melanocytes and, in humans, exists in two forms: eumelanin (brown/black) and pheomelanin (red/yellow).^{157,158} Latitudinal variation of UVR explains approximately 86% of observed human melanin index differences.¹⁶² Multiple hypotheses have been proposed to explain the selective pressures shaping this variation, but few establish direct connections to reproductive success. However, two complementary and widely supported hypotheses—the folate photoprotection hypothesis and the vitamin D3 synthesis hypothesis—provide plausible fitness-related mechanisms.^{162,164,165}

Khoe-San and Khoe-San descendant populations have much lighter skin compared to other sub-Saharan African populations.^{166–168} While the Khoe-San genetic signature carries varying degrees of European admixture, this alone does not explain their reduced pigmentation.¹⁶⁹ Although few pigmentation alleles are shared between the Khoe-San and Europeans, several key depigmentation alleles common in Europeans are entirely absent in the Khoe-San.¹⁶⁹ This suggests that distinct sets of pigmentation alleles underlie the lighter complexions observed in the Khoe-San. Khoe and San populations sampled in South Africa also reside at an intermediate-latitude region (24–29° S), where UVR exposure is lower than in equatorial zones, aligns with adaptive expectations for lighter pigmentation at such latitudes.¹⁶⁹

Among the shared alleles between the Khoe-San and Europeans is *SLC24A5* rs1426654*A lightening allele. This allele found its way to Africa through the back-to-Africa

migration (5-10kya) and was subsequently introduced into the Khoe and San via gene flow with East African pastoralists migrants approximately 2kya.¹⁶⁸ Since its introduction to the ~~≠~~Khomani San and Nama, and an additional pulse from European admixture, the allele has risen to moderate frequencies (~42%) across the populations.^{168,169} Previous research has found this locus to be under strong, directional selection ($s = 0.04$).¹⁶⁸ This allele alone contributes to a measurable reduction in melanin index by about 7 units per copy.^{133,152,169} Though remarkably strong in its effect, rs1426654*A lightening allele explains 30% of the variation in skin color among Khoe-San and Khoe-San descendants. Previous Khoe-San focused pigmentation studies proposed that skin color is shaped by an architecture of some large-effect loci and many smaller-effect variants, which contrasts the simpler pigmentation models of in Europeans.^{168,169} This complexity is likely due to the Khoe-San retaining high levels of genetic diversity, including pigmentation alleles not commonly found in other populations.^{169,170}

Narrow-sense heritability (h^2) estimates for skin color are consistently high across human populations (0.55 to >0.90).^{158,169,171} When compared to other highly heritable complex traits, such as height ($h^2 = 0.8$),¹⁷² skin pigmentation ranks among the highest for any human complex trait. Among southern Africans, both SNP- and pedigree-based heritability estimates for baseline melanin index are remarkably high, reaching 0.90–0.97 and 0.96, respectively.¹⁶⁹ Although skin pigmentation is highly heritable, the proportion of variance in melanin index explained by ancestry proportions and locus-specific heritability estimates varies substantially across populations, reflecting differences in both ancestry composition and the underlying genetic architecture of pigmentation. For instance, in Cape Verdeans, who are of admixed European- and West African-ancestries, genome-wide West African ancestry proportions explain a large fraction of pigmentation variance (44%), while specific loci such as *SLC24A5* ($h^2 = 0.18$) and *SLC45A2*

($h^2 = 0.07$) contribute disproportionately to phenotypic variance.¹³³ Four loci—*SLC24A5*, *SLC45A2*, *OCA2*, and *KITLG*—explain 35% of skin pigmentation variance in Cape Verdeans but all of these are loci which are highly differentiated between Europeans and West Africans.¹³³ Thus, they do little to explain variation in the phenotype coming from West African ancestry haplotypes.

Heritability patterns among continental African populations point to a more complex genetic architecture. While *SLC24A5* has a large effect on skin color and is present in some African populations due to back-to-Africa migrations,^{111,133,168,169} other African-specific variants at *MFSD12*, *DDB1/TMEM138* and *OCA2* strongly contribute to skin pigmentation variation across eastern and southern Africans ($h^2 = 0.045$, 0.022 and 0.039 , respectively).¹¹¹ A study focused solely on southern Khoe and San pooled the top 50 hits in their GWAS but found these accounted for only 23% of the variance in skin color, highlighting the complexities underlying skin pigmentation genetic architecture. The genetic architecture responsible for pigmentation variability within and between populations remains poorly understood, with most skin pigmentation research largely focused on Eurasian populations^{136,141,155,173,174} despite Africa harbors the greatest range of pigmentation diversity.^{166,167} Only in the past decade have genome-wide approaches been applied to study pigmentation in continental African populations, but much of the trait's heritability remains unexplained.^{111,169,170}

Here, we characterize different elements of skin pigmentation's genetic architecture among individuals with high Khoe-San ancestry by analyzing 359 newly collected and 760 previously published^{111,169} samples ($n=1119$). We describe the distribution of skin pigmentation heritability across chromosomes, rather than specific loci, and perform a genome-wide association study (GWAS) to interrogate chromosomes enriched with high heritability. We

functionally annotate and assess candidate genes of interest and discuss the constraints in standard GWAS practices and their limitations when the phenotype is correlated with ancestry. Lastly, we assess whether polygenic scores (PGS) are useful with highly structured phenotypes which correlate with ancestry by implementing two European-derived PGS and assessing their transferability to the study populations.

4.3 Results

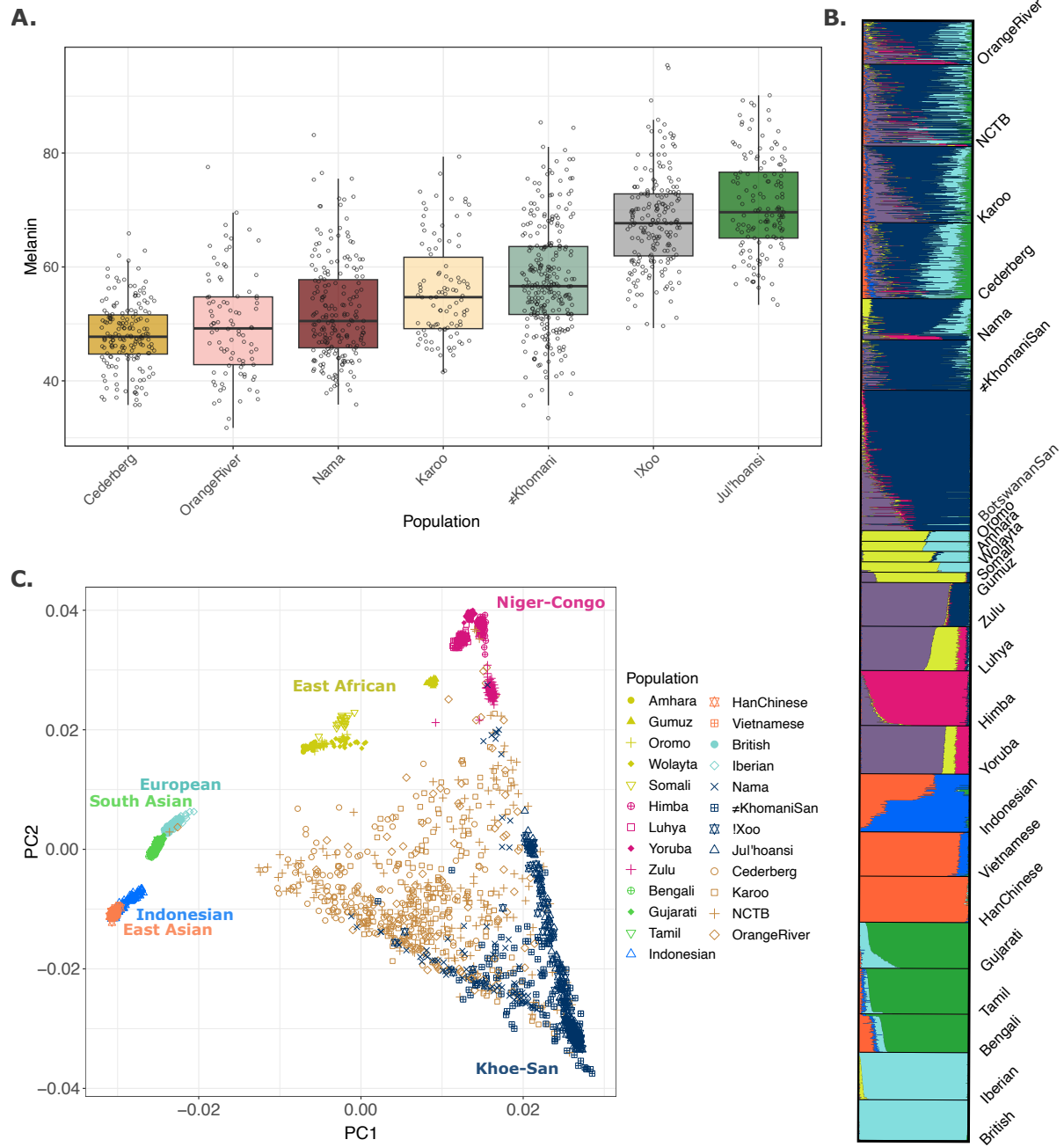


Figure 4.1). Pigmentation distribution and global population structure for study cohorts. A) Distribution of melanin in southern African study cohort. Populations are arranged by lighter to darker complexion based on their average melanin index. In some instances, the !Xoo and Ju|'hoansi, who were inferred based on a series of analyses (see *Supplemental Information for Chapter 2: Inferring Botswanan Populations*) are merged into one population: Botswanan San. C) Unsupervised ADMIXTURE run of $k=8$, retaining the same individuals

and set of variants as the PCA. Colors correspond to broader ancestry labels specified in the Figure 4.1A. C) PCA on 806,460 congruent markers with previously merged genomes (see previous chapter), with shapes distinguishing populations within broader inferred ancestries (ADMIXTURE $k=8$). Newly collected SAC populations are colored beige, while shapes indicate sampling locality within South Africa. Colors throughout the figures corresponding to broadly inferred ancestry from ADMIXTURE $k=8$.

4.3.1 Significant differences in average melanin across Khoe-San and Khoe-San descendant populations

Southern African Khoe and San populations have traditionally practiced hunter-gatherer and pastoral subsistence strategies, respectfully, and spoke click-constant Khoisan languages. They occupy diverse ecological regions across the region and share a common genetic signature frequently referred to as “Khoe-San” in literature (see *sections 1.3 and 2.4.3*), with the hyphen representing their cultural distinction from one another.¹⁶ We include four previously published Khoe-San populations (the Nama and ≠Khomani San from South Africa^{12,169} and the !Xoo and Ju|’hoansi from Botswana¹¹¹) along with newly sampled South African Coloured (SAC) individuals from the Cederberg, Karoo, and Middle Orange River regions of South Africa (see *sections 2.4.3*). The SAC communities frequently self-identify as “Coloured,” an ethnic term with roots in South Africa’s apartheid regime, though we and others^{58–60,62} have demonstrated Khoe-San ancestry comprises a large proportion of their genetic composition, leading us to refer to them as Khoe-San descendants (see *sections 2.4.3*).

Other ancestries found among the SAC include European, western/central (WC) African, and South and Southeast Asian ancestries.^{58,59} Collectively, this provides the largest and most comprehensive paired phenotype-genotype dataset of populations with high Khoe-San ancestry spanning a latitudinal gradient from southern South Africa to northern Botswana (32–19°S). Newly collected and previously published samples exhibit wide variation in skin pigmentation,

forming a south to north gradient in melanin index values across the seven study populations (Figure 4.1A; *Supplemental Table 4.1*). Post-hoc pairwise comparisons of average melanin highlighted significant differences (adjusted $p < 0.05$; *Supplemental Table 4.2*). The melanin index difference between the Cederberg and Ju|'hoansi was larger than all pairwise comparisons ($\Delta = -22.8$, $p < 1 \times 10^{-14}$; *Supplemental Table 4.2*). These populations are also the most geographically distant among the study populations: the Cederberg towns are closest to Cape Town and also exhibit the highest European ancestry (ADMIXTURE $k5 = 0.278$; see *section 2.3.1*), while the Ju|'hoansi are located in northern Botswana and have minimal inter-continental admixture (see *section 2.3.1*).

The Ju|'hoansi exhibited the darkest average melanin index (mean = 70.9), followed by the !Xoo (mean = 67.8). The Ju|'hoansi have higher melanin index values than the !Xoo ($\Delta = 3.10$, $p = 0.026$), though the magnitude of this difference is modest. The Botswanan San populations are distinctly darker than the South African cohorts ($p < 0.0001$) The ≠Khomani San (mean = 57.8) and Karoo individuals (mean = 55.9) have similar average melanin index values ($\Delta = 1.94$, $p = 0.48$), while the Nama were slightly lighter than both, differing significantly from the ≠Khomani San ($\Delta = -5.46$, $p < 1 \times 10^{-9}$) and Karoo ($\Delta = -3.53$, $p = 0.016$). Average melanin values among the Nama were indistinct from the Orange River ($\Delta = 2.62$, $p = 0.20$). Similarly, the Cederberg were not significantly different from individuals sampled along the Orange River ($\Delta = -1.6$, $p = 0.77$). The lightest complexions were observed among participants from the Cederberg (mean = 48.1).

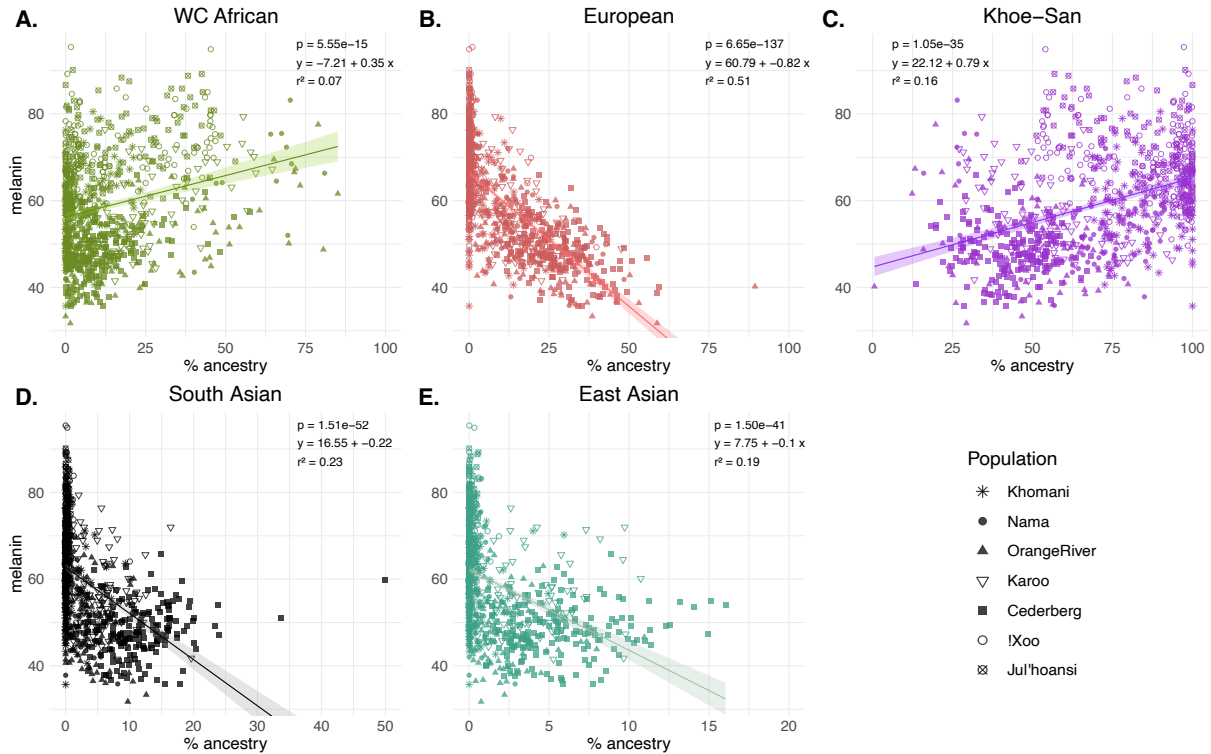


Figure 4.2) Modeling melanin by global ancestry. Ancestry estimates were inferred from unsupervised ADMIXTURE $k=5$ across 806,460 markers (see section 2.6.5). Each panel represents a single ancestry modeled against pigmentation measurements, A) Bantu, B) European, C) Khoe-San, D) South Asian, and E) East Asian inferred ancestry, with p-value, model equation, and R^2 values annotated within the plot.

4.3.2 Correlation between ancestry and melanin index

We characterize the genetic ancestry of the southern African Khoe-San and Coloured populations, whom we refer to as “Khoe-San-descent”, using 806,460 autosomal markers, principal component analysis (PCA)⁶⁹ and ADMIXTURE⁷⁰ (see section 2.6.2 and 2.6.5). Unsupervised ADMIXTURE⁷⁰ runs across multiple levels to infer ancestry clusters, setting $k=4:10$ (see section 2.6.5: Supplemental Figure 2.7), with $k=9$ and $k=10$ demonstrating the lowest cross-validation error rates (cv error=0.519). At $k=5$, we observe ancestries that correspond to European, South Asian, broadly eastern Asian, Khoe-San and WC African. ADMIXTURE $k=8$ highlights 8 broad ancestries across the reference populations: European,

West and East African, South, Southeast, and East Asian, and Khoe-San (Figure 4.1B). Principal Component (PC) 1 segregates populations out-of-Africa from Khoe-San populations, while PC2 separates within-Africa populations, mainly the southern African Khoe-San groups from Niger-Congo speaking populations originating from western/central Africa (Figure 4.1C). The majority of the newly collected samples (Cederberg, Orange River and Karoo) are distributed along the center of PC1 and PC2, a pattern previously observed^{58,62} reflecting their complex admixture.^{58,59}

ADMIXTURE $k=5$ results (see *section 2.6.2 and 2.6.5*) were regressed against melanin index values for each of the five primary ancestries commonly found among the southern African study cohorts (Figure 4.2A – 4.2E). European ancestry alone explained 51% of the variance in skin color, with higher proportions of European ancestry strongly associating with lighter skin pigmentation (Figure 4.2B). These values shift when the Botswanan San are excluded or examined independently from the South Africans ($R^2 = 0.33$ and $R^2 = 0.05$, respectively; *Supplemental Figure 4.2*), likely because the Botswanan San included in the study have higher melanin index values but trace European ancestry (<0.01). Following patterns of European ancestry, South and East Asian ancestries independently explain 23% and 19% of the variance in skin color (Figures 2D and 2E), despite their low proportion among the cohort (<0.12 and <0.06 , respectively). These three ancestries are highly correlated with each other, due to similar colonial-era gene flow patterns introduced at similar points in time: South and East Asian ancestry ($r=0.81$), European and South Asian ($r=0.61$) and European and East Asian ($r=0.55$).

Interestingly, ancestry contributed by darkly pigmented western/central (WC) Africans explains only a small portion of the observed variation in skin pigmentation ($R^2 = 0.07$; Figure 4.2A). This value is slightly higher when the Botswanan San are excluded ($R^2 = 0.15$; *Supplemental Figure 4.2*), who are significantly darker in complexion than the South African

cohorts ($p < 0.0001$; *Supplemental Table 4.2*). Analyzing the Botswanan San independently yields a similar estimate ($R^2 = 0.139$; *Supplemental Figure 4.3*), reinforcing the minimal explanatory power of WC ancestry alone in accounting for pigmentation variation in our study. Khoe-San ancestry is the predominant component across all populations (ranging between 43–86%; see *section 2.6.2 and 2.6.5*) yet it explains less variance in skin pigmentation ($R^2 = 0.16$; *Figure 4.2C*) than the European component. The variance explained by Khoe-San ancestry is reduced even further when we exclude the Botswanan San ($R^2 = 0.03$; *Supplemental Figure 4.2*) or examine the Botswanan San independently from the South African cohort ($R^2 = 0.08$; *Supplemental Figure 4.2*). This likely occurs because the Botswanan San harbor high proportions of Khoe-San ancestry and have higher melanin index than the South African Khoe-San populations included in the study.

4.3.3 Models including *rs1426654*A* dosage increase ancestry against melanin correlation

We investigated how skin color varies across different *rs1426654*A* (*SLC24A5*) genotypic dosages (i.e. 0 = GG, 1 = GA, 2 = AA) and found a clear dosage-dependent effect. Individuals carrying zero copies of the derived allele exhibit an average melanin index of approximately 63.9, while each additional copy is associated with a decrease in melanin by about 9.2 units (*Supplemental Figure 4.4*). An additive model explains approximately 30.2% of the variation in skin color ($p = 2.24 \times 10^{-92}$; *Supplemental Figure 4.4*), reinforces the loci's previous estimates as a strong effect locus among southern African Khoe-San groups.^{168,169}

The ancestral allele is virtually fixed in western and central Africans and has been selected out of Europeans populations entirely (0-2%). The Botswanan San carry the lowest frequencies of the *rs1426654*A* allele (*Supplemental Tables 4.3*). We replicate previous findings reporting the lightening *SLC24A5* allele at higher frequency in the Nama (51.3%) than the

Khomani San (32.9%)¹⁶⁹, and find similarly frequencies in the Cederberg (48.5%) and Orange River communities (41.7%), and in the Karoo (31.6%; (*Supplemental Table 4.3*). Despite the Nama from the Richtersveld exhibiting higher frequency of the rs1426654*A allele, individuals from the Cederberg are significantly lighter in complexion ($\Delta = -4.2$, $p = 5.8 \times 10^{-5}$; *Supplemental Tables 4.1 – 4.3*), potentially due to their overall higher proportions of European ancestry.

Taking this loci's strong effect on pigmentation into consideration, we included rs1426654*A allelic dosage as a fixed effect to partition out the known variance attributable to *SLC24A5*, and assess how much ancestry proportions explain the rest of the variation in pigmentation observed. Accounting for dosage of rs1426654 led to an increase in the proportion of variance explained by ancestry (*Supplemental Figure 4.5*), even when the Botswanan San were excluded (*Supplemental Figure 4.6*). Specifically, the adjusted R^2 rose to 0.39 for the Khoe-San, 0.34 for WC, 0.57 for European, and 0.43 and 0.42 for the South and East Asian models, respectively. Excluding the Botswanan San from the South African cohort decreased these estimates (*Supplemental Figure 4.6*), likely because the Botswanan San are more darkly pigmented, with trace levels of inferred EUR ancestry (mean <1%) and comparable WC African proportions to the South Africans included in the study (~13%).

This increase led us to subset the data with respect to an individual's copy number of the rs1426654*A lightening allele and then independently model each ancestry against melanin index values within the dosages (*Supplemental Figure 4.7 and Supplemental Table 4.4*). European ancestry consistently lightened skin color (*Supplemental Figure 4.7 and Supplemental Table 4.4*), with a dosage of 0 showing the steepest slope (-67.8 , $R^2 = 0.38$), then 1 copy (-42.3 , $R^2 = 0.43$) or 2 copies rs1426654*A (-33 , $R^2 = 0.38$). By contrast, WC ancestry showed a

modest darkening only in the subset of samples with 0 copies of the lightening allele (slope = 14.7, $R^2 = 0.06$) and yielded non-significant results among those with 1 or 2 copies ($p > 0.15$, $R^2 \leq 0.01$; *Supplemental Table 4.4*). Khoe-San ancestry was associated with increased melanin in all genotype groups, but proportion of Khoe-San ancestry among the subset of individuals with 0 rs1426654*A copies explains only 4% of the variation in skin pigmentation, as opposed to 24% and 30% for the 1 and 2 dosages, respectively (*Supplemental Table 4.4*).

4.3.4 Heritability is unevenly distributed across the genome

The two prior continental African-focused studies examining the genetic architecture of pigmentation have proposed a general model comprising a few loci with moderate to large effects, alongside numerous variants of smaller effect.¹⁶⁹ Because small effects are harder to map, we do not know how this variance is distributed across the genome. Drawing parallels to height, a trait with comparably high heritability estimates ($h^2 = 0.8$),¹⁷² we hypothesized that skin pigmentation has a similar distribution of heritability,¹⁷⁵ with small-effect loci distributed relatively evenly across the genome. Under this model we expect chromosomal heritability to scale with chromosome length, as previously observed in partitioned heritability analyses of height.¹⁷⁵

We partitioned the genetic data by chromosome and estimated heritability of melanin index independently for each chromosome using GCTA. We then modeled heritability estimates as a function of chromosome length to assess whether longer chromosomes contribute proportionally more to trait variance, as expected under a standard polygenic architecture. However, we found no relationship between chromosome length and heritability ($R^2 = 0.009$; Figure 4.3A). However, we identified four chromosomes with enriched heritability estimates that fall above the 95% confidence interval: chromosomes 15, 5, 6, and 9 ($h^2 = 0.213, 0.096, 0.093,$

and 0.073, respectively). In total, these chromosome's heritability estimates explain nearly half of the phenotypic variance in pigmentation (46%) and, unsurprisingly, contain canonical pigmentation genes, including *SLC24A5* and *OCA2* on chromosome 15, *SLC45A2* on chromosome 5,^{176,177} *IRF4* on chromosome 6^{178,179} and *TYRP1* on chromosome 9.¹⁸⁰ Other chromosomes fell below the 95% confidence interval, providing little to no significant contribution to the phenotypic variance (Figure 4.3A).

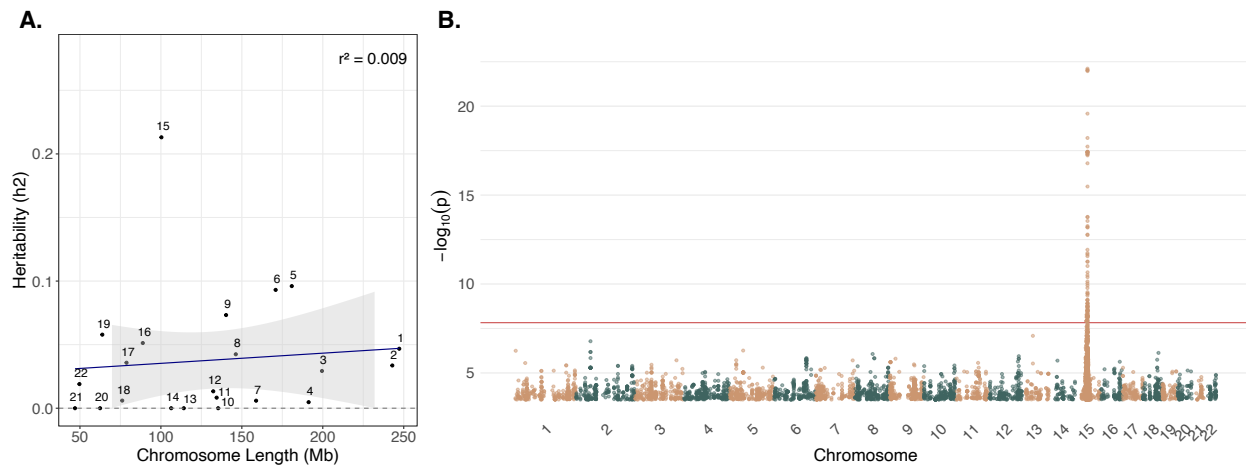


Figure 4.3) Skin pigmentation genetic architecture. A) Heritability by chromosome length. B) Standard GWAS results. The Manhattan plot depicts the $-\log_{10}(p\text{-value})$ results for the standard skin pigmentation GWAS. The model assessed 12,573,868 million markers across the genome and included a genetic-relatedness matrix and the top 5 within-population PCs.

4.3.5 Genome-wide association study

With our larger sample size, we tested whether we could fine-map genes associated with baseline skin pigmentation, with a particular focus on the four chromosomes with high heritability. First, each array type was independently imputed against the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA2) reference panel. This reference

panel includes 78 $\neq$ Khomani San and 49 Nama individuals, in addition to other continental African populations, making it the most representative reference panel available for the individuals sampled in this study (Methods). Our standard genome-wide association study (GWAS) assessed 12,573,868 markers and included a genetic relatedness matrix (GRM) and the top 5 principal components (PCs) as covariates. The model is well calibrated ($\lambda = 1.002$; *Supplemental Figure 4.8*) and included 2.5 times the sample size of previous Khoe-San-focused pigmentation studies.^{111,169,170}

We successfully replicated prior findings on chromosome 15 for *SLC24A5* (rs1426654), which served as our positive control ($p = 9.45 \times 10^{-23}$; Figure 4.3B). Interestingly, no additional variants reached genome-wide significance despite observing high heritability estimates from chromosomes 5, 6, and 9. We noted a potential secondary peak in our standard GWAS model that fell about 1.1 million base pairs (Mbp) downstream of *OCA2* and 3.9Mbp upstream of *SLC24A5*. The intronic variant rs2733210 (*FRMD5*; $p = 5.40 \times 10^{-09}$) is at high frequency among European populations (~95%) and moderate frequency within the study populations (37%). *FRMD5* plays a role in regulating cell mobility and strengthened cell-to-cell adhesion [<https://www.ncbi.nlm.nih.gov/gene/84978>]. This SNP is annotated as a cis-eQTL in both sun-exposed and not sun-exposed skin in GTEx, where it modulates expression of the upstream gene *ADAL* (*Supplemental Figure 4.9A – 4.9C*). *ADAL* is predicted to enable adenosine deaminase activity [<https://www.ncbi.nlm.nih.gov/gene/161823/>], and previous studies suggest adenosine may suppress melanogenesis by downregulating tyrosinase activity.¹⁸¹

We conducted a second GWAS conditional on rs1426654 to determine whether this signal was independent of *SLC24A5*. The association at rs2733210 was attenuated in the conditional model ($p = 0.04$), suggesting that this peak likely reflects long-range linkage

disequilibrium (LD) due to admixture.^{182,183} No other significant or suggestive loci were found in the conditional model (*Supplemental Figure 4.10*).

GWAS summary statistics were input into FUMA¹⁸⁴ to functionally annotate and map variants with p-values $\leq 1 \times 10^{-5}$ and to test for their enrichment in skin tissue. FUMA is a publicly available web-based platform that annotates SNPs, maps them to genes, and prioritizes biological pathways by integrating diverse functional genomic data.¹⁸⁴ We leveraged FUMA's Genotype-Tissue Expression (GTEx) integration to identify tissue-specific expression quantitative-trait loci (eQTLs), and incorporated Hi-C chromatin interaction fibroblasts data (ENCODE ENCSR220MNC) to functionally annotate regulatory variants (see Methods).

About 53% and 30.5% of variants across 59 FUMA-identified genomic loci – that is, windows defined by merging LD blocks around significant or near-significant SNPs – were intronic and intergenic, respectively (*Supplemental Figure 4.11*). SNPs were mapped to 176 genes, of which 131 were protein coding. Chromosome 15 contained the highest number of FUMA-identified genomic loci ($n = 13$; *Supplemental Figure 4.12*) relative to other chromosomes, likely a result of long-range LD surrounding a strongly associated locus (*SLC24A5*). Additional loci were identified across the genome, though these typically contained fewer genes, if any (*Supplemental Figure 4.12*).

Gene expression for the 176 mapped genes were assessed across 58 tissue types (derived from GTEx; *Supplemental Figure 4.13*). This analysis uses gene-based p-values while adjusting for average expression across all tissues (see FUMA tutorial for more detail).¹⁸⁴ Results identified skin (sun-exposed and not sun-exposed) tissues as one of the top enriched signals (*Supplemental Figure 4.13*), along with lung, salivary gland, and small intestine tissue. Enrichment in non-skin related tissues likely reflects the pleiotropic roles some of the identified

genes play, having shown high expression across multiple tissues (*Supplemental Figure 4.12*). Differentially expressed genes (DEGs) across various tissue types and found strong enrichment of genes among upregulated DEGs in skin and cell-cultured fibroblast tissue (*Supplemental Figure 4.15*). FUMA-derived gene-sets are reported in *Supplemental Table 4.5*.

4.3.6 Constraints with including PCs in GWAS

GWAS in admixed populations offer statistical advantages to gene discovery relative to their homogenous ancestral populations. Variation in allele frequencies between ancestral populations can increase power to detect novel loci within admixed populations, up to two-fold greater power than in either of their ancestral populations alone,¹⁸⁵ and improve transferability of results across populations.^{185,186} For instance, *SLC24A5* was first identified in zebrafish and subsequently studied in humans. The lightening allele's fixation among European populations resulted in little power to detect its effect on skin color through association studies,¹⁵² but variation at this locus eased its identification among admixed populations.^{133,169}

However, these advantages come with important caveats when the phenotype of interest is strongly correlated with ancestry. As demonstrated in *section 4.3.1.*, skin pigmentation is a highly diverged trait with a strong correlation to ancestry. Current standard GWAS practice include PCs to correct for population structure. PCs capture broad ancestral differences; however, when ancestry is strongly correlated with a phenotype, including PCs can inadvertently reduce power to detect true genetic associations.^{186,187} This is particularly important for ancestry-specific or causal variants with highly differentiated allele frequencies across populations, where the genetic contribution to a phenotype differs across ancestries. A recent study suggests incorporating a GRM may sufficiently control for structure, and additional covariates correcting for population structure are over-correcting the model.¹⁸⁷ This might be less problematic in large

biobank-derived samples, but it can have constraints within smaller sample sizes association studies that are limited in power to detect small-effect variants. Additionally, for a phenotype like skin pigmentation, removing PCs would increase type I error due to ancestry-related allele frequency differences, highlighting the inherent trade-off between false positive control and statistical power in ancestry-correlated traits.

To describe this further, we ran an additional GWAS that only included a GRM as a covariate (GWAS_{GRM}) and compared it to the standard base model that included a GRM and top 5 PCs (GWAS_{GRM+PCs}; *Supplemental Figures 4.16 and 4.17*). We found that the GWAS_{GRM} model reported more genome-wide significant associations exceeding the Bonferroni threshold; however, many of these were likely spurious, representing independent loci without strong linkage with neighboring SNPs (*Supplemental Figure 4.16*). Given the conservative nature of Bonferroni correction, we also evaluated false discovery rate (FDR) thresholds at 0.05, 0.10, and 0.20. The GWAS_{GRM} model yielded a greater number of loci surpassing all four significance thresholds (*Supplemental Figure 4.17*). Beta values remained consistent ($r = 0.95$) between models, but p-values varied ($r = 0.77$; *Supplemental Figure 4.17*).

We next focused our analysis on 169 previously characterized and newly associated pigmentation loci identified in Bajpai et al. (2023).¹⁸⁸ We attenuated the data to these loci and ran a similar protocol to the models described above, executing GWAS_{GRM+PCs} and GWAS_{GRM} models (*Supplemental Figures 4.18 and 4.19*). The attenuated GWAS_{GRM+PCs} model identified an additional gene beyond *SLC24A5*, *NOX4* on chromosome 11, which encodes an enzyme known to inhibit tyrosinase production.¹⁸⁹ Two additional genes had loci which surpassed Bonferroni threshold in the attenuated GWAS_{GRM} model, including the canonical locus *SLC45A2* on

chromosome 5 and *MTMR9* on chromosome 8, the latter believed to play a role in cell proliferation (NCBI Gene ID: 66036; *Supplemental Figure 4.18*).

4.3.7 Poor Transferability of European-Derived Polygenic Scores

We assessed the transferability of two recently developed polygenic score (PGS) panels (PGS001897 and PGS002110) following Privé et al. (2022)'s methods as closely as possible.¹⁹⁰ Both panels were generated using penalized logistic regression methods and trained on >300,000 Europeans from the UK Biobank, which uses self-reported skin pigmentation data categorized into 6 options: very fair, fair, light olive, dark olive, brown, black. Privé and colleagues assessed prediction of these European-derived PGS onto different global populations using partial correlation analyses, and found prediction drastically decreased with increase distance between the target population and European.¹⁹⁰

Although ADMIXTURE ($k=5$) estimates of European ancestry comprise 15–27.8% of the Cederberg, Orange River, Nama, and Karoo (<10 % of the ≠Khomani, !Xoo, and Ju|'hoansi) and is strongly associated with lighter skin ($R^2=0.51$), our partial correlation analyses find European-derived PGS panels fail to predict phenotypic variation in the study cohort, with 95% confidence intervals overlapping with 0 similarly to their performance among Chinese and Nigerian participants in the original study.¹⁹⁰ Only when we examine the partial correlation within populations do we see a signal among the ≠Khomani. The PGS here have a correlation of 0.14 and 0.15 to melanin (PGS001897 and PGS002110, respectfully; *Supplemental Figures 4.20 – 4.21*). All over populations under evaluation have a report no correlation ($r=0$; *Supplemental Figures 4.22–4.23*). Correlation values comparing models with PCs and without PCs are reported in *Supplemental Table 4.6*.

4.3.8 Prioritizing candidates for CRISPR-driven gene knockout in zebrafish

We prioritized candidate genes that met all of the following criteria: (1) they are protein-coding; (2) contain at least one SNP acting as an eQTL in fibroblasts or pigmented tissues (e.g., skin or eye); (3) they have a single ortholog in zebrafish; and (4) their zebrafish ortholog is expressed in chromatophores within the first week after fertilization.

Rationale: Advantages to using zebrafish as a model organism

Danio rerio, or zebrafish, are robust model organisms commonly used to understand the function of orthologous genes associated with a particular human characteristic or disease.^{191,192} The *D. rerio* genome contains 26,202 protein-coding genes and it totals 1.42 gigabases over 25 chromosomes.¹⁹² About 71% of the zebrafish genome comprises of genes orthologous to human genes.^{192–194} Zebrafish are capable of yielding >100 progenies in a single breeding session, have external fertilization, and are easy to house because of their relatively small size.¹⁹¹ During the pharyngula period, about 24 hours post-fertilization (hpf), pigmentation begins to develop in the retina and skin. By 36hpf, pigmentation noticeably develops in the tail and ventral stripes and begin differentiating wild types (WT) from experimental zebrafish.^{152,195} Their rapid development and high number of progeny allow for high-throughput pigmentation analyses to be conducted in as little as a week post-fertilization.

Zebrafish have three primary types of chromatophores that contribute to their characteristic stripe patterns. Melanophores produce darkly pigmented melanin and gives rise to the black stripes seen in adult zebrafish, analogous to mammalian melanocytes.^{196–198} Xanthophores contain yellow to orange pteridine pigments and are primarily responsible for the lighter stripe regions, while iridophores are reflective cells whose iridescent appearance comes from crystalline purine platelets, such as guanine, rather than traditional pigments.¹⁹⁶ Although

previous research has found physiological changes can rapidly alter pigmentation, these environmentally induced responses are reversible and distinct from the baseline pigmentation patterns established during development.^{197,199}

Pigmentation is well conserved in zebrafish and has successfully been used to study human pigmentation genetics.^{152,195,200} In fact, *SLC24A5* was first identified in zebrafish and later investigated in humans.¹⁵² Morpholino-mediated knock-out of *slc24a5* led to mosaic hypopigmentation and a *golden* phenotype in 94% of injected embryos.¹⁹⁵ *TYR* knockouts have previously been studied in zebrafish,¹⁹⁵ with nonsynonymous variants at this locus shown to lighten pigmentation in zebrafish and can contribute to oculocutaneous albinism in humans.^{195,201} Interrogating *IRF4*, a gene involved in the regulation of *TYR* and previously associated with skin pigmentation in humans, has also been successfully studied in zebrafish as well, with mutants exhibiting reduced melanin levels compared to WT zebrafish.²⁰⁰ These studies demonstrate strong conservation of pigmentation genes between zebrafish and humans, justifying the use of zebrafish as a high-throughput model organism for human pigmentation research.

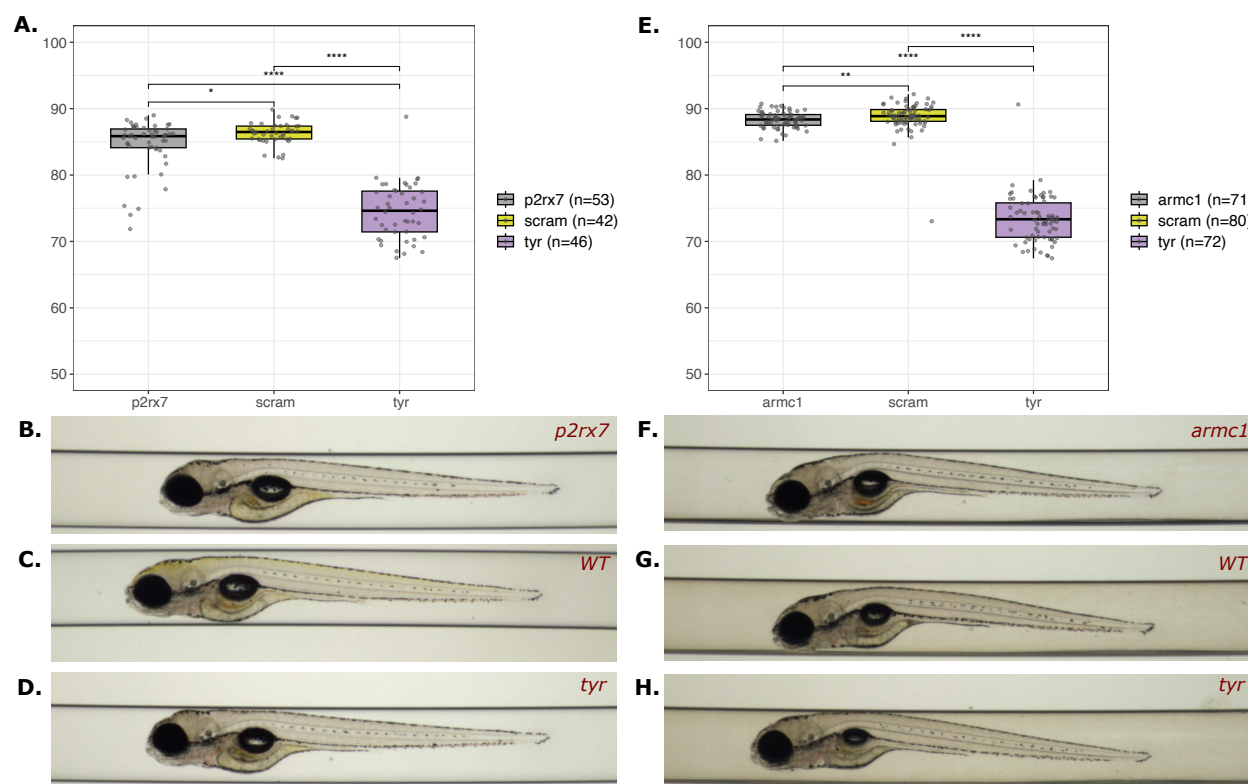


Figure 4.4) CRISPR-driven knockout results for *p2rx7* and *armc1*. Figures A and E depict boxplot results for *p2rx7* and *armc1*, respectively, comparing mutant, wildtype scramble-negative control (scram; WT), and positive control *tyr*-knock outs, with sample size for each in the legend. Figures B and F are 5dpf *p2rx7* and *armc1* mutant zebrafish. Figures C, D, G, and H are of 5dpf scramble and *tyr* knock-outs, with respect to the label on the top right of the images.

4.3.9 Interrogating causality of candidate genes

Several FUMA-derived genes (P2RX7, ARMC1, BLVRB) were prioritized for functional validation based on their co-expression with canonical pigmentation genes.²⁰² Mutant *p2rx7* and *armc1* zebrafish exhibited significantly altered pigmentation scores compared with negative control (scramble) zebrafish (Figure 4.4A and 4.4E; $p < 0.05$ and $p < 0.01$, respectively), though neither was as lightly pigmented as *tyr* knockouts. In contrast, *blvr*b mutant zebrafish did not differ significantly from negative controls (Supplemental Figure 4.27).

We also prioritized previously identified *SNX13*, *CYB561A3* and *FRMD5* and *ADAL*, the second peak upstream from *SLC24A5*. (*Supplemental Table 4.7*). Previously associated genes *SNX13* and *CYB561A3* were independently targeted using multiple guide RNAs, and pigmentation was quantified using our Pigmentation Score (see *section 4.6.11*). Knockout of *SNX13* and *CYB561A3* did not significantly reduce pigmentation relative to scrambled WT controls, and both showed elevated pigmentation compared to the *tyr* positive control ($p < 0.001$; *Supplemental Figure 4.24A*). Mutant *snx13* zebrafish were also raised into adulthood and qualitatively assessed alongside WT negative controls, but no differences were observed (*Supplemental Figures 4.24B and 4.24C*). Similar patterns were observed when we interrogated *FRMD5* and *ADAL* (*Supplemental Figures 4.25A – 4.25C*), the potential secondary peak identified upstream of *SLC24A5*, though we anticipated this hit was likely a result of long-range LD.^{182,183} Previously associated regions did not show significant differences in pigmentation compared to WT scramble controls, suggesting that if these genes play a role in pigmentation it is likely an indirect and minor effect, or zebrafish is not a suitable model for these genes.

4.4 Discussion

Recent studies in diverse continental African populations reveal that skin pigmentation has a far more complex genetic architecture than in Eurasian or admixed populations, with just a small number of large-effect loci and many more low-effect variants spread across the genome. These patterns largely reflect differences in demographic history and selection pressures across latitudes. For example, 4 loci explain 35% of the variance in skin color among admixed Cape Verdeans, whereas the top fifty GWAS hits in Khoe-San and Khoe-San descendants account for only about 23%. Here we (1) investigate the chromosomal distribution of Khoe-San heritability, (2) collectively examine its genetic architecture within new and published Khoe-San cohorts, and

(3) highlight challenges of standard GWAS approaches and (4) transferability of skin pigmentation PGS.

4.4.1 Partitioned heritability estimates highlight genomic regions omitted by GWAS

We modeled skin pigmentation heritability by chromosome and found chromosome length does not predict percent of phenotypic variance each chromosome explains, contrary to the proportional scaling seen for other highly polygenic traits like height.¹⁷² Instead, heritability is strikingly enriched on chromosomes 5, 6, 9, and 15. A strong heritability estimate on chromosome 15 is unsurprising given the robust association at our positive-control locus rs1426654 (*SLC24A5*). Chromosome 15 also contains other key pigmentation genes, *OCA2/HERC2*, who have had markers linked to variation in skin color across global populations,^{142,151,203} including eastern and southern African populations.^{111,170} Although these variants did not reach genome-wide significance in our GWAS, their contribution should not be overlooked when assessing the heritability attributable to chromosome 15. The concentration of nearly half the skin pigmentation heritability at these 4 chromosomes implies skin pigmentation architecture in the Khoe-San is much simpler than other complex traits, such as height.

Particularly surprising is the high heritability on chromosomes 5, 6, and 9 despite the absence of genome-wide significant hits in our GWAS. These chromosomes are home to canonical pigmentation loci, including *SLC45A2* (chromosome 5),^{176,177} *IRF4* (chromosome 6),^{178,179} and *TYRPI* (chromosome 9).¹⁸⁰ Variants within or near these genes have previously been replicated in pigmentation variation, associated with abnormal pigmentation, and/or found to be under selection.²⁰⁴ Despite the lack of GWAS hits, heritability enrichment along these chromosomes indicate the presence of markers with combined influence on skin color variation across the study populations, since alleles that do not segregate cannot drive phenotypic

variation. For example, the rs1426654*A allele is nearly fixed in Europeans and explains less than 0.5% of the variance in skin color,²⁰⁴ but explains 12.8% among eastern and southern Africans because of its strong effect and moderate frequency.¹¹¹ Meaning, variants along chromosomes 5, 6, and 9 have segregating alleles that collectively explain high proportions of the phenotypic variance but are not identifiable in GWAS at the current our sample size.

Two possible hypothesize could explain the enriched heritability and lack of GWAS hits on chromosomes 5, 6, 9, and 15: 1) Khoe-San skin pigmentation genetic architecture is primarily localized to those four chromosomes, with multiple low-effect alleles concentrated in those regions 2) numerous low-effect epistatic interactions among variants on Khoe-San haplotypes are influence pigmentation and masking individual GWAS signals. Although we more than doubled the sample size of previous studies on Khoe-San skin pigmentation genetics,^{169,170} we only replicate additional loci outside of *SLC24A5* when models were confined to previously identified pigmentation genes (GWAS_{GRM+PCs} and GWAS_{GRM}). This is likely a result of lowering the multi-hypothesis burden but given that we and others have demonstrated ancestry correlates with melanin index values,^{133,169} one might expect the GWAS_{GRM} models to uncover more than two regions. The scarcity of signals across previously associated pigmentation genes in GWAS_{GRM} and the conditional rs1426654 GWAS may indicate a genetic architecture predominantly driven by *SLC24A5* and small-effect variants, including eQTLs,¹⁷⁰ among Khoe-San and Khoe-San descendants.

Alternatively, many low effect epistatic interactions on Khoe-San haplotypes could be influencing skin pigmentation and remain undetected in GWAS. Epistatic interactions are not uncommon in pigmentation genetics, with the classic example of variants interacting between *OCA2/HERC2* impacting eye and skin pigmentation variation.^{142,205} When sufficiently strong,

epistatic interactions can appear additive and be identified in GWAS, but small-effect interactions are often missed and require large sample sizes to be detected.²⁰⁶ To identify these regions in a traditional GWAS, we would need to allow for interaction between loci in the model. Epistasis is easier to test with strong interactions and/or less complicated genetic architectures, but with relatively small sample sizes and polygenic traits you're left with little statistical power using standard GWAS methods. Regardless, efforts to identify loci contributing to skin pigmentation in Khoe-San descendants will require more sampling to increase power to detect small-effect loci.

4.4.2 Functional Annotation and Validation

FUMA's applied modules (SNP2GENE, GENE2FUNC, etc.) found over 80% of SNPs fall within intronic or intergenic regions, rather than in protein-coding exons. A subset of these variants colocalize with eQTL signals in cultured fibroblasts and/or both sun-exposed and non-sun-exposed skin, indicating their association with gene-expression differences in relevant cell types with bulk of these eQTL loci fall on chromosome 15 (*Supplemental Figure 4.26*). Consistent with these observations, SNP-to-gene property analyses showed significant enrichment for genes that are highly expressed in both sun-exposed and non-sun-exposed skin tissues. Tissue expression of prioritized genes and test for gene-set enrichment also support that lower p-value regions ($p \leq 1 \times 10^{-5}$) fall within genes expressed in skin; albeit GTEx data is biased toward European ancestry, with approximately 82% of samples derived from European descent individuals and only 12.4% from African Americans.²⁰⁷

CRISPR-driven knockouts of the previously identified (*SNX13* and *CYB561A3*) and newly found loci (*FRMD5* and *ADAL*) in zebrafish maintained normal pigmentation, which may reflect several factors. First, these genes likely have modest, indirect effects on pigmentation,

producing changes too subtle to be detected by day 5pft phenotyping pipeline or reach statistical significance between knockouts and negative controls. Especially in the case of knocking out inhibitors of known pigmentation genes, like *CYB561A3*, pigmentation might not elevate enough to be noticeable or reach statistical significance by 5dpf. One of these scenarios is likely the limiting factor for *CYB561A3* despite its validation as a regulator of *SLC24A5* and *TYR* in MNT-1 cell lines.¹⁷⁰ Secondly, FishInspector's current pigmentation models are not sensitive enough to capture the full breadth of melanophores, often undercounting when processing zebrafish images. As a result, we limited our pigmentation quantification to the eye, which may not catch effects of genes that act indirectly on melanogenesis or that influence chromatophores outside the eye. Future research may benefit from developing more sensitive imaging processing models catered to pigmentation.

4.4.3 Poor Transferability of European-Derived Polygenic Scores

We evaluated the transferability of two European-derived skin pigmentation PGS onto the seven Khoe-San populations included in this study and find both panels have zero predictive power (*Supplemental Figures 4.23–4.24*), with the exception of the ≠Khomani San when evaluated independently from the other populations. Interpreting these results is challenging because skin pigmentation is a highly structured phenotype that correlates with ancestry. Similar to GWAS, PGS require correcting for population structure to avoid false positive results. But when ancestry-informative markers are correlated with the phenotype, especially a highly structured phenotype like skin color, PCs adjustment can inadvertently remove genuine predictive power, because highly differentiated or ancestry-specific variants will be highly correlated with the genetic PCs. Our partial-correlation analysis cannot resolve whether the lack of PGS transferability reflects Khoe-San-specific pigmentation alleles missing from European

discovery panels or whether large allele-frequency differences combined with principal-component adjustment are masking any genuine predictive signal.

4.5 Conclusion

Skin pigmentation in Khoe-San–descendant populations follows a clear geographic gradient but shows little correlation with Khoe-San ancestry proportions once historical admixture and gene flow is accounted for. Partitioning heritability across chromosomes reveals strong enrichment on chromosomes 5, 6, 9, and 15, indicating that pigmentation loci are distinctly clustered. Despite high heritability estimates concentrated on a subset of chromosomes, standard GWAS approaches replicate only the canonical *SLC24A5* signal, while functional validation of additional candidate loci in zebrafish knockouts fails to confirm most of our novel genes. These results demonstrate how correlations between overall ancestry and phenotype can substantially attenuate power to detect associations in admixed populations when the phenotype is highly differentiated among ancestral groups. Fully characterizing the small-effect variants and potential epistatic interactions contributing to skin pigmentation in these populations will require larger sample sizes, improved statistical models, and functional assays optimized for subtle phenotypic effects.

4.6 Methods

4.6.1 Sample collections and quality control

Skin pigmentation was assessed using a handheld DSM II ColoriMeter spectrometer. The device was calibrated at each sampling location under consistent lighting conditions using the manufacturer-provided white blank tile, following standard protocol. Measurements were taken on areas of skin typically protected from sun exposure to estimate baseline pigmentation. Each

participant was measured five times on the upper inner surface of the right and left arms, and five times on the dorsal surface of one wrist. Areas with moles, scars, or tattoos were avoided. Erythema and melanin index values were recorded for all trials. Values more than two standard deviations from the per-arm mean were excluded, and the remaining values were averaged by individual. The final melanin index for each participant was computed by averaging the mean values from both arms, yielding a single estimate of constitutive pigmentation. See section 2.6.1 for genetic and ethnographic data collection, as well as methods used to incorporate the Botswanan San.

4.6.2 Merging within genotyping platforms

Samples were genotyped across several platforms in different batches, including 2 batches on H3Africa, 1 on Omni2.5, 2 on OmniExpress, 1 on Omni5, and 3 batches on Multi-Ethnic Genotyping Array Consortium (MEGA) Illumina arrays, with some individuals genotyped across multiple platforms. All but the H3Africa genotyped samples were lifted from HG19 to HG38 using `picard 'LiftoverVcf'`. Each data set was then normalized to HG38 to correct for reference and alternative allele swaps using `'bcftools norm -m -any -c s'` flags and annotated so that SNP IDs were `'chr:pos'` using `'bcftools annotate --set-id '%CHROM:%POS''` flag. Batches were merged within their respective arrays using `'bcftools merge -m id'` and then filtered for markers with greater than 5% missingness with `'bcftools filter -e 'F_MISSING > 0.05'`. After each merge, data sets were assessed using smartPCA and visualized with samples labeled by array. Coordinates for duplicate samples were evaluated against another along PC1 and PC2. Duplicate samples with different coordinates along PCs 1 and 2 were removed, while those which fell in proximity to one another were prioritized by density of array.

4.6.3 Phasing

All 5 data sets were then separated by chromosome using bcftools in preparation for phasing and imputation. Hg38 recombination map and SHAPEIT4 (v4.2) were used to phase samples against CAAPA reference panel. This panel best represents haplotypes within the datasets as it includes both $\neq$ Khomani San and Nama genomes. SHAPEIT4 window size was set to 2.5cM. We changed the default iteration sequence to prioritize accuracy over speed in the analysis by using the `--mcmc-iterations 10b,1p,1b,1p,1b,1p,1b,1p,10m` flag.

4.6.4 Imputation

All chromosomes across arrays were “chunked” into more computationally effective imputation windows using imp5Chunker and `--buffer-size 150000` flag. An internal script was then used to iterate through each segment and impute using impute5 (v1.1.5). Each chunk was imputed against the CAAPA reference panel using the hg38 recombination map. Accuracy was prioritized over computational speed, using a buffer size of 150kb, depth of PBWT index set to 8 (`--pwbt-depth`), and 0.01cM distance for selection is performed (`--pwbt-cm`). Each chromosome within an array then had their respective chunks consolidated using `bcftools concat` and sorted using `bcftools sort`. Monomorphic alleles were then filtered out using `bcftools filter -e 'AC==0 || AC==AN'` and filtered for imputation quality score of 0.80 or higher using vcftools (v0.1.16).

4.6.5 Post-imputation merging across platforms

We merged across arrays with similar platforms and similar Khoe-San ancestry profiles. We began by merging the Omni2.5 and Omni Express Plus arrays, then merged H3Africa with the MEGA genotyped samples together, while removing variants with at least 5% missingness after each merge. We then combined to two larger datasets which encompass individuals with

mostly southern Kalahari ancestry. Finally, we merge these samples with the Omni5 genotyped Botswanan San individuals and assess the merge using smartPCA.

We conducted two analyses to evaluate whether the clustering of Botswanan San samples reflects true genetic variation rather than a batch effect. First, we randomly retained 1.2 million markers from the Omni5 array to match the density of the H3Africa dataset, referred to as the test dataset, and reran phasing, imputation, filtering, and merging as previously described *section 2.6.2*. Second, we filtered the imputed data to include only markers shared with a previously assembled dataset used for ADMIXTURE, and local ancestry inference. Quality checks across both the test dataset and the original results showed no differences, supporting that the observed clustering reflects genuine genetic structure rather than batch-related artifacts.

4.6.6 Modeling Pigmentation Variation

We generated boxplots of melanin index values stratified by population (Figure 4.1A). This was done in R using the ggplot2 package, with population defined by project- and inferred-identifiers (see *Supplemental Information for Chapter 2: Inferring Botswanan Populations*). We integrated ancestry estimates derived from ADMIXTURE $k=5$ and rs1426654*A copy-number into a merged dataset. We assess the relationship between pigmentation and ancestry components using separate linear regressions for each continental ancestry (East Asian, South Asian, European, Western/Central African, and Khoe-San). This was implemented using base R's ``lm()`` function. Regression coefficients and p-values were extracted and annotated directly onto the resulting plots. We then extended these models by incorporating genotype dosage at rs1426654*A as an additional covariate to estimate the effect of ancestry while controlling for the locus. Predictions and confidence intervals were calculated using ``predict()`` and visualized across ancestries in a multi-panel layout.

Two additional approaches evaluated the relationship between ancestry components and melanin index values, while accounting for rs1426654*A genotype dosage. Both approaches aimed to clarify whether the observed ancestry associations with pigmentation were independent of this well-known pigmentation variant. First, we modeled melanin (baseM) as a function of each ancestry proportion and included rs1426654*A genotype dosage (coded 0, 1, or 2) as a categorical covariate using the formula ``baseM ~ ancestry + factor(rs1426654_dose)``. We visualized these models using ``geom_abline()`` in ggplot2, with separate lines corresponding to each dosage category (0, 1, and 2). The adjusted R^2 and slope coefficients were annotated for each ancestry panel. We then stratified the data by rs1426654*A dosage group and conducted independent linear models of the form ``baseM ~ ancestry`` within each subset. Regression fits and confidence intervals were visualized using ``geom_smooth(method="lm")``, with separate smoothers overlaid for each genotype subset.

4.6.7 Genome-wide association studies

We performed genotype data preparation, relatedness estimation, and genome-wide association analysis using PLINK¹²¹ (v1.9) and GCTA²⁰⁸ (v1.93.1beta). First, a VCF file containing quality-filtered genotype data was converted to binary PLINK format using ``--make-bed`` while preserving allele order. Variants included passed quality control thresholds for less than 5% genotype missingness, minor allele frequency greater than 1%, and Hardy-Weinberg equilibrium ($p > 1 \times 10^{-5}$). Samples without phenotype information were removed, and quality-controlled pigmentation data were merged using the `--pheno` option in PLINK.

We constructed the genomic relationship matrix (GRM) by randomly selecting 1.5 million variants from the post-imputed dataset. The selected variants were extracted using PLINK, and the resulting subset was used as input for GCTA to compute a binary GRM (``--`

make-grm-bin`). Principal components were computed from the GRM using GCTA's `--pca` function, retaining the top five components as quantitative covariates to account for population structure in downstream analysis.

Four primary GWAS models were implemented: (1) a model including only the GRM, (2) a model including both the GRM and the top five PCs to correct for population structure, (3) a model conditioning (`gcta64 --cojo-cond`) on rs1426654*A genetic information, which included the GRM, top five PCs, and (4) two targeted GWAS models restricted to loci identified in Pritchard et al. (2023), each of which included a GRM and were run with and without the top five PCs. For all models, the same post-imputation quality-controlled dataset and phenotype data were used. Multiple testing correction was applied using a Bonferroni threshold based on the number of variants tested in each analysis. Summary statistics were visualized with custom R scripts which generate Manhattan and quantile-quantile (QQ) plots using ggplot2 R-package.

4.6.8 Partitioned heritability

We estimate chromosome heritability using PLINK (v1.9) and GCTA (v1.93.1beta). We began by splitting the genotype dataset – previously filtered for quality control and limited to 1.5 million randomly selected SNPs used for GRM construction – into individual chromosomes using PLINK (`--chr`) and converting each subset into binary PLINK format (`--make-bed`). For each chromosome, a separate GRM was then constructed using GCTA's `--make-grm` function. A multi-GRM file was generated, listing all 22 chromosome-specific GRMs, and used as input for GCTA's restricted maximum likelihood analysis (`--mgrm` and `--reml`) to jointly estimate the variance explained by each chromosome.

We visualized the proportion of phenotypic variance explained by each autosome using ggplot2 in R to plot chromosome-specific heritability estimates (h^2) alongside their standard

errors. We generated a figure displaying heritability estimates ordered by chromosome number (Figure 3B), with each point representing a chromosome, with error bars indicating the standard error around the estimate, and a horizontal dashed line marking zero heritability to aid interpretation. To further explore the relationship between chromosome length and heritability, we merged the heritability estimates with chromosome length data from the GRCh38 reference genome. Chromosome lengths were converted to megabases (Mb) and modeled using a linear regression (`lm`) in R. The regression assessed whether longer chromosomes tend to explain a greater proportion of phenotypic variance. Model results, including the regression equation and coefficient of determination (R^2) annotated directly onto the resulting plot, and a fitted linear trendline and 95% confidence interval.

4.6.9 Transferability of Polygenic Scores

We evaluate the transferability of European-derived PGS for skin pigmentation to our study populations using two recently developed weight files (PGS001897 and PGS002110) from the PGS Catalog (Privé et al 2022). These scores were trained on 391,124 White British participants individuals in the UK Biobank, with models that adjusted for sex, age, birth date, deprivation index, and 16 genetic PCs. We applied these scores to our dataset, including sex, age, and 16 PCs, although birth date and deprivation index were unavailable. Using a standardized processing pipeline, we reformatted, lifted over, filtered, and harmonized the SNP weights to match our genotype data, discarding incongruent SNPs and resolving strand inconsistencies before calculating per-chromosome PGS values using PLINK2. The final number of markers assessed from PGS001897 and PGS002110 weight files were 14,598 and 261,878, respectively. We assessed the predictive performance using R^2 from linear models both with and without PCs. We estimated the partial correlation between residualized PGS and phenotype values—adjusting

for sex, age, and 16 PCs. Confidence intervals for the partial correlation were computed using Fisher's Z-transformation.

4.6.10 Functional Annotation of GWAS

We used the online platform FUMA (Functional Mapping and Annotation of Genome-Wide Association Studies, v1.5.2) to functionally annotate and prioritize genetic loci identified in our GWAS. FUMA integrates positional, expression quantitative trait loci (eQTL), and chromatin interaction data to map SNPs to candidate genes and assess their regulatory potential in relevant biological contexts. We implemented our standard GWAS using default FUMA settings with custom integration of Hi-C fibroblast chromatin interaction data, sourced from the Aiden Lab achieves.

Analyses were based on the 1000 Genomes Phase 3 AFR population reference panel, excluding the MHC region. Independent lead SNPs were defined at a genome-wide significance threshold of $P < 1 \times 10^{-05}$, with a secondary SNP-wise threshold ($p < 0.05$) and an LD threshold of 0.6 for clumping (default). Functional mapping included positional mapping within ± 10 kb of lead SNPs, CADD score filtering, and fibroblast chromatin data. eQTL mapping was performed using multiple datasets, including GTEx v8 eye, fibroblast, and unexposed- and sun exposed-skin tissues. Chromatin interaction mapping incorporated significant Hi-C interactions from cultured fibroblasts with an FDR threshold of 0.001 and promoter windows of 250–500 bp. MAGMA (v1.08) was also used within FUMA to perform gene-based association testing and tissue expression enrichment. Results were visualized through the FUMA web browser. Regions containing protein-coding genes with significant eQTL associations in skin, eye, or fibroblast tissues and/or evidence of chromatin interaction were prioritized for further functional validation. See FUMA's tutorial for more detail [<https://fuma.ctglab.nl/tutorial>].

4.6.11 Experimental gene-knockouts

Prioritized candidate genes were individually knocked out using 3 to 4 guide RNAs (gRNAs) targeting early the first set of exons, in order to maximize the likelihood of generating loss-of-function alleles.²⁰⁹ In cases involving paralogous, each paralogue was assessed both independently of one another. CRISPRScan, a genome specific gRNA design tool, was used for on-target prediction and off-target scoring of gRNAs.^{210,211} Each gene interrogated will have an accompanied positive and negative batch control, where *tyr* is our positive control and a scramble is our negative control. All control and experimental embryos were derived from the same parental mating pairs to prevent genetic background bias across injections. Microinjection took place during the single-cell embryo stage to increase the indel efficiency. Validation of CRISPR knockout was via PCR amplification followed by polyacrylamide gel electrophoresis (PAGE) to confirm indel presence at the targeted loci.

Zebrafish were imaged using a VAST BioImager from Union Biometrics at 5 days post-fertilization (dpf).^{152,195} The VAST BioImager automatically pulls larvae from 96-well plates through a tube and orients each larva for dorsal, ventral, and lateral imaging. All experiments were performed on larvae for high-throughput analyses. FishInspector, a software that produces measurements for a variety of zebrafish phenotypes using VAST BioImager images, was used to quantitatively characterize pigmentation. Pigmentation readings for “eye mean intensity” from mutant zebrafish larvae outputted from this software were analyzed and compared against “mock” controls, with the expectation that gene-knockout has altered pigmentation in mutant zebrafish. Since the absence of melanin pigment (e.g., in *tyr* knockouts) results in higher *mean* intensity values, due to increased light transmission from the white background, we transformed the raw output into a more intuitive metric termed **Pigmentation Score**. Raw pixel intensity values

range from FishInspector range from 0 to 255. Our method inverts the intensity scale and normalizes values to a 0 to 100, where 100 represents the highest pigmentation (i.e., darkest), and 0 represents no visible pigmentation. This approach allows us to compare pigmentation across experimental groups with differing melanin content.

Supplementary Information

Section 1: Supplemental Information for Chapter 2

Inferring Botswanan Populations

We began by running an unsupervised ADMIXTURE inferring 2 to 7 population clusters (k) with the entirety of the Crawford et al. (2017) data (Supplemental Figure 2.1), along with the GBR, BEB, STU, GIH, KHV, IND, YRI, HMB, LWK, Zulu, Gumuz, Somali, Wolayta, Omoro, KHM, and Nama, and query populations. From there, we subset the Crawford et al. (2017) data for the 314 individuals with the highest Khoe-San ancestry proportion (Supplemental Figure 2.1A), as that is the reported number of San individuals sampled in the study,¹¹¹ and then re-ran ADMIXTURE k2 through k7 with those individuals (Supplemental Figure 2.1B).

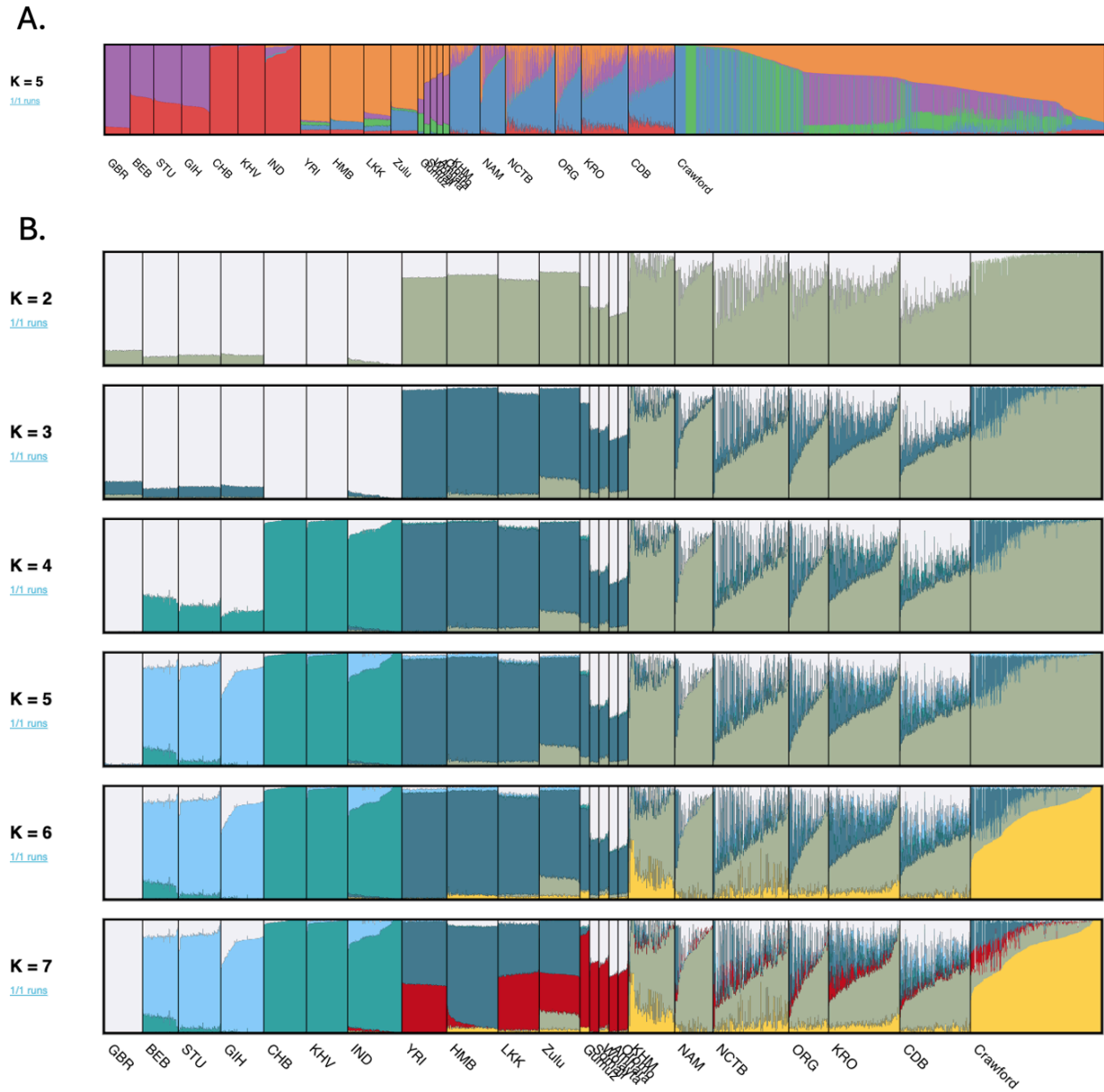
We then filtered the Botswanan San, ≠Khomani, and Nama individuals with 85% or greater Khoe-San ancestry as inferred by ADMIXTURE k5 and plotted them on a PCA (Supplemental Figure 2.2A). We observe a previously reported cline among the Kalahari San populations, ≠Khomani, and Nama. The Botswanan San split into two clusters, one of which anchored a corner of PC2. From there, we were able to infer the cluster of Botswanan San individuals driving PC1 likely have greater genetic variation relative to the other high Khoe-San populations. This leads us to infer that cluster of individuals to be the !Xoo, who were reported to have a higher population effective size relative to the Ju|'haonsi.²¹²

We continued to our KHS ancestry-specific MDS. We add ≠Khomani, Nama, and Botswanan San individuals who did not meet the 85% Khoe-San ancestry cut-off into the query. We project those haplotypes, along with all first- and second-degree relatives in the data set, onto an MDS with the 85% and greater KHS reference. Projecting these individuals minimizes the effect relatives and admixture can have on dimensionality reduction. We use the mclust²¹³ R

package to read in MDS2 data for the Botswanan San, the dimension where the Botswanan San split into two clusters. This package is designed for complex clustering and classification using a Gaussian finite mixture, meaning it assumes the data is made up of several Gaussian distributions. We use it to model MDS2 distribution with two clusters and identify the mean and standard deviation of both clusters (Supplemental Figure 2.2B-D). We filtered our model results to include only individuals with a probability of 0.9 or higher of belonging to a cluster. Upon examining samples across both clusters, we found no overlap between individuals assigned to each cluster. Most individuals who did not meet the ≥ 0.9 probability threshold are those positioned between the two cluster peaks, i.e., more than one standard deviation away from the modeled mean. These 10 individuals were labeled as BSAN_UNK (Botswanan San, Unknown). The individuals within the cluster positioned on the right side of MDS2 were labeled !Xoo, as they represent the samples contributing most to the genetic variation in MDS1. This classification leaves the remaining cluster of individuals to be identified as Ju|'hoansi (Supplemental Figure 2.3).

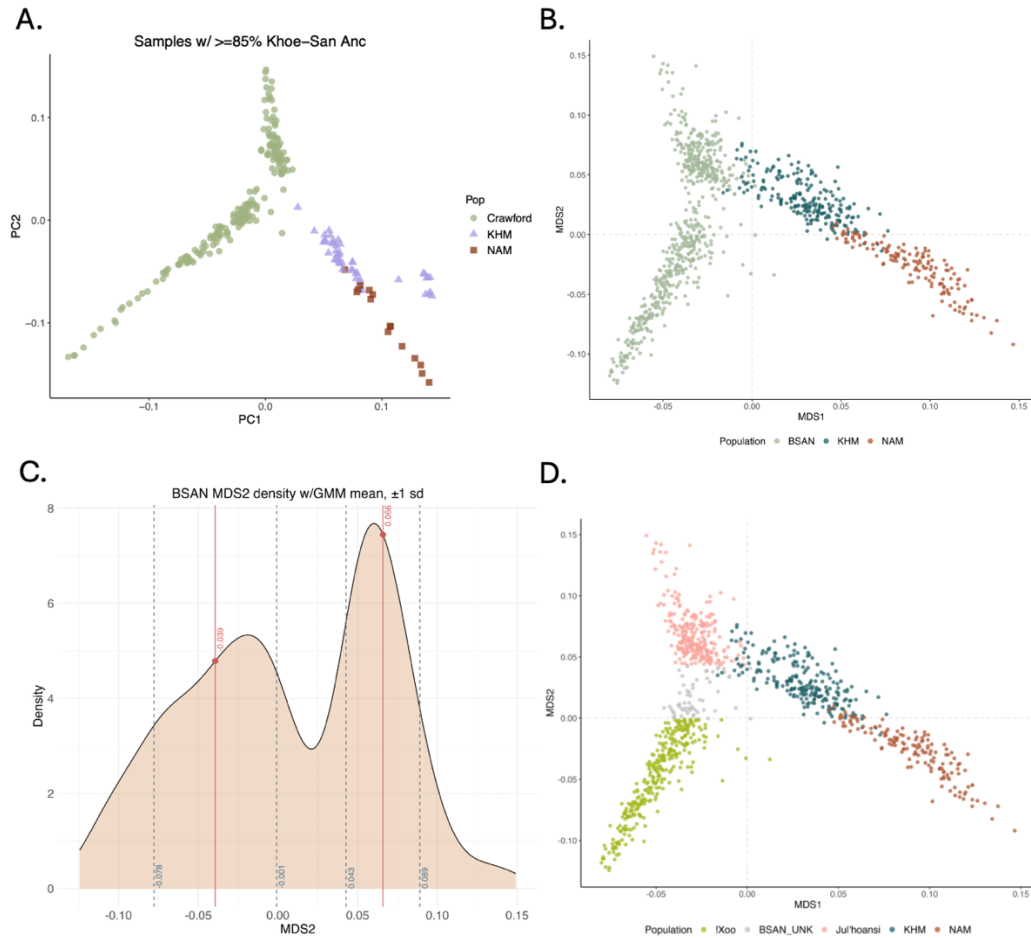
In order to refine the clustering of the assigned BSAN_UNK individuals, we performed a dimensionality reduction with uniform manifold approximation and projection (UMAP)²¹⁴ followed by genetic clustering with Hierarchical Density-Based Spatial Clustering of Applications with Noise implementation, HDBSCAN(ϵ).²¹⁵ We used as input the first 100 principal components (PCs) obtained with SmartPCA program from EIGENSOFT 7.2.1⁶⁹ and applied the same approach described in Diaz-Papkovich et al 2023.²¹⁶ For visualization purposes, UMAP was run with 25 number of nearest neighbors, a minimum distance of 0.3, and a dimension reduction of 2 components. For clustering, UMAP parameters included 50 number

of neighbors, minimum distance of 0.01 and 5 number of components. Next, HDBSCAN($\hat{\epsilon}$) was used to extract clusters using an $\hat{\epsilon}$ of 0.3 and a minimum number of points in a cluster of 50.

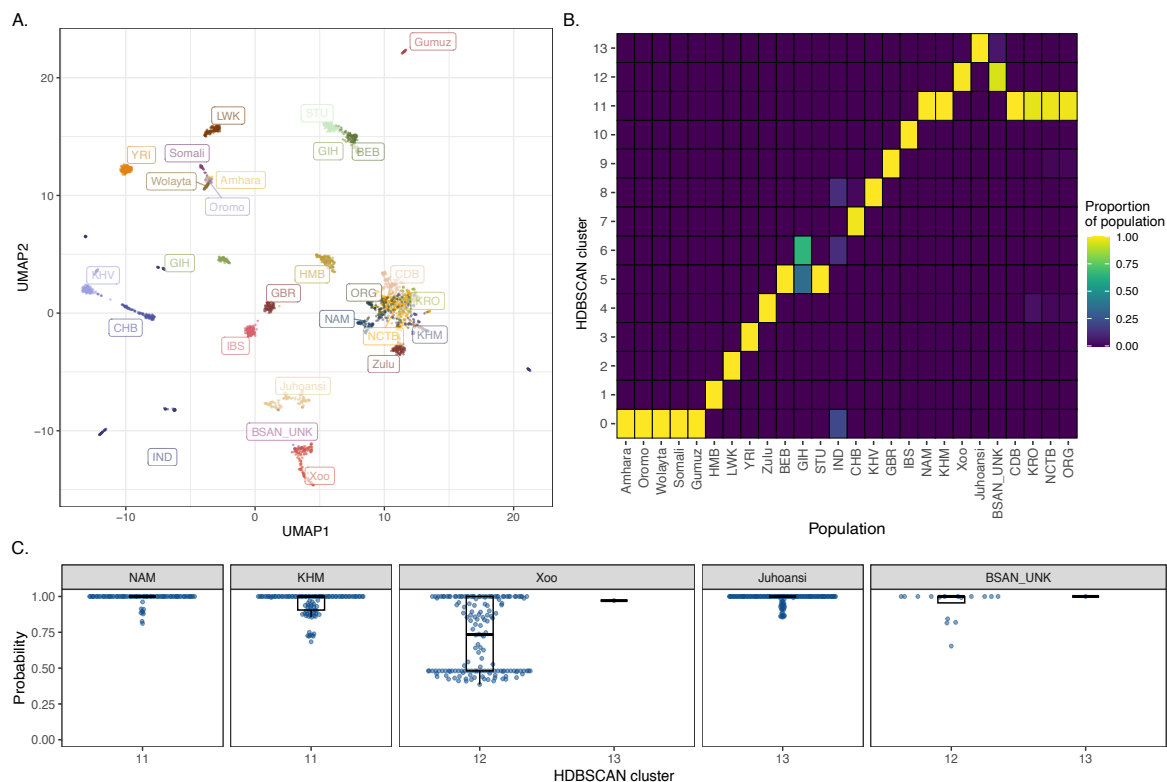


Supplemental Figure 2.1) Identifying the Botswanan San using ADMIXTURE. A)

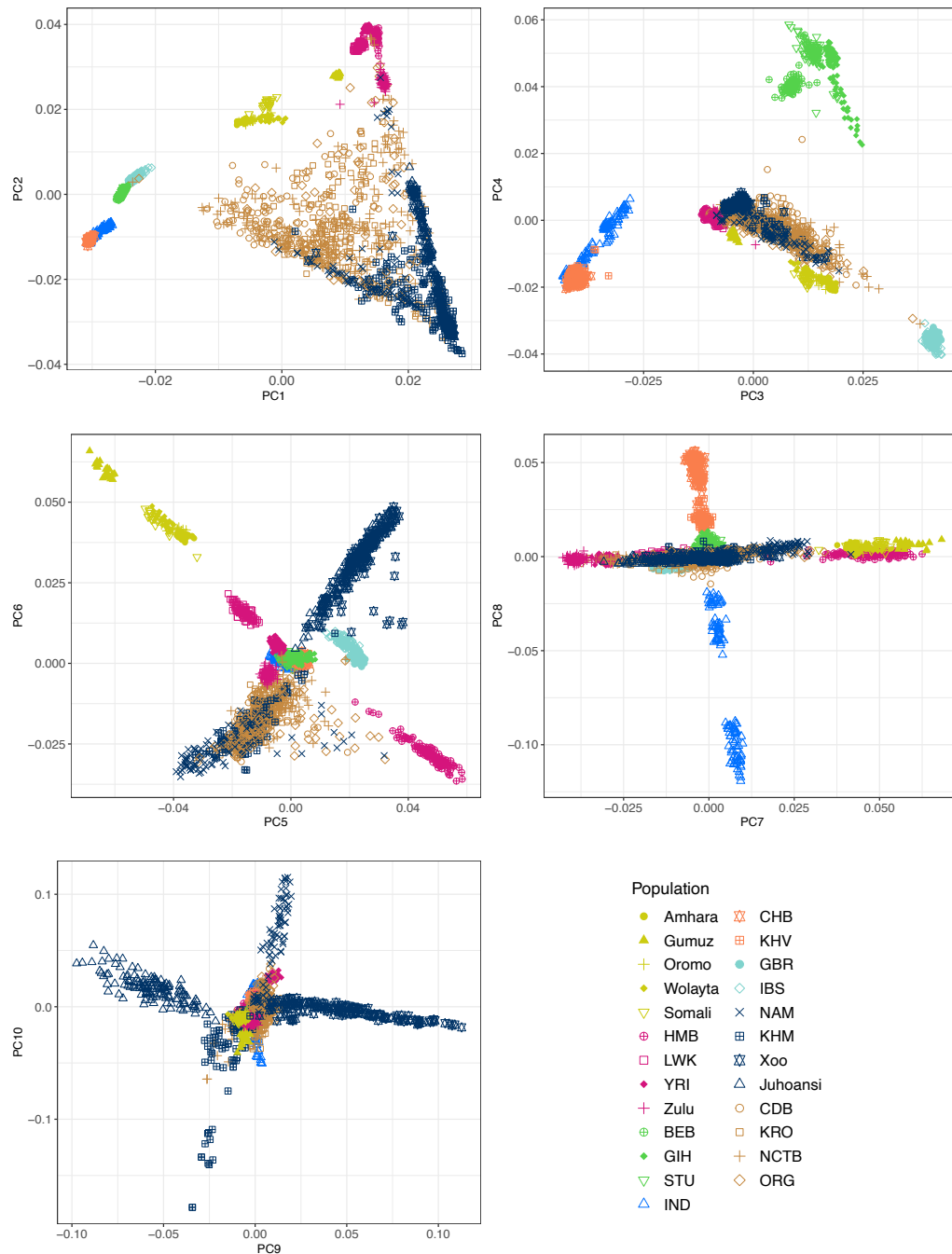
ADMIXTURE at k5 with the entire Crawford et al. (2017) data. **B)** ADMIXTURE k2 through k7 with only 314 Botswanan San with the highest Khoe-San ancestry as inferred by figure A's results.



Supplemental Figure 2.2) Differentiating Botswanan San Populations. **A)** Individuals with greater than 85% Khoe-San ancestry as inferred by ADMIXTURE k5 from Supplemental Figure 2.1A. **B)** Khoe-San (KHS) ancestry-specific multidimensional scaling (MDS) 1 and 2 of KHS reference populations. Individuals with less than 85% Khoe-San ancestry were added to the local ancestry inference query and projected into the MDS space, along with first- and second-degree relatives. **C)** mclust Gaussian mixture model identified mean and 1 standard deviation of MDS2 Botswanan San density. Red line specified averages while dashed blue lines highlight the 1 standard deviation cut offs. **D)** Cluster/population assignment as inferred by the Gaussian mixture model from mclust R package and previously published population effective size data.²¹²

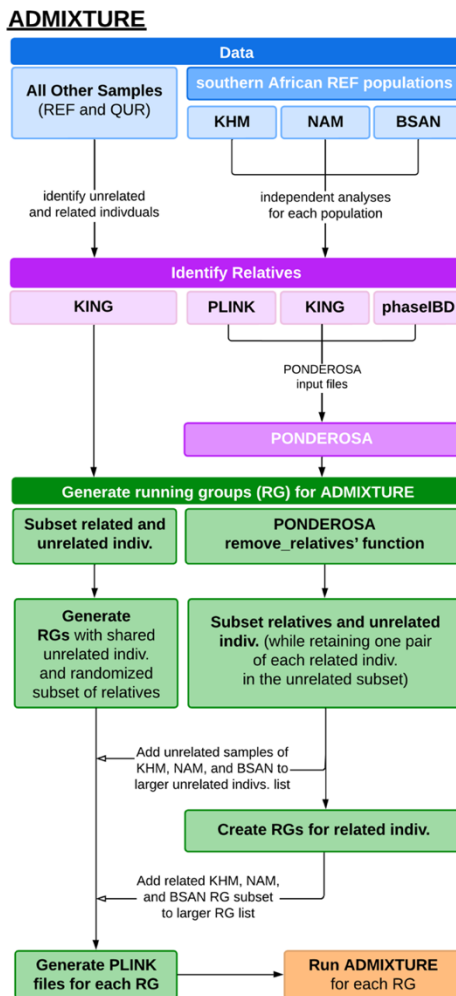


Supplemental Figure 2.3) UMAP and HBDscan of reference and query populations **A.** UMAP generated with 25 nearest neighbors, and minimum distance of 0.3. Labels refer to population labels. **B.** Proportion of individuals of each population contained within a given cluster inferred from HBDscan algorithm (see Methods for further details on the parameters used). **C.** Distribution probabilities per individual to be in the cluster inferred in B for each Khoe-San reference population.

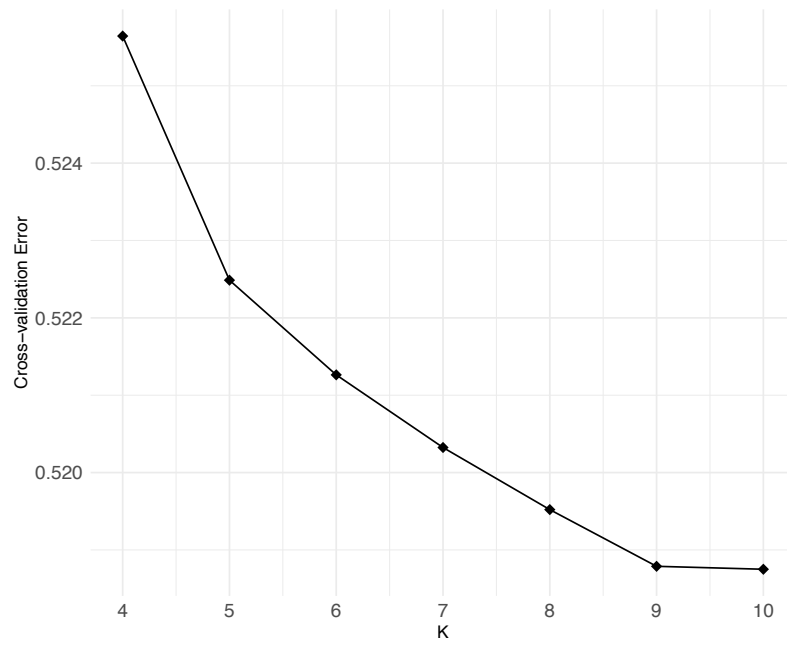


Supplemental Figure 2.4) PC1 through PC10. PCs 1 through 10 were generated using smartPCA and plotted using ggplot2 in R. Colors roughly represent broader ancestry, while shapes represent population.

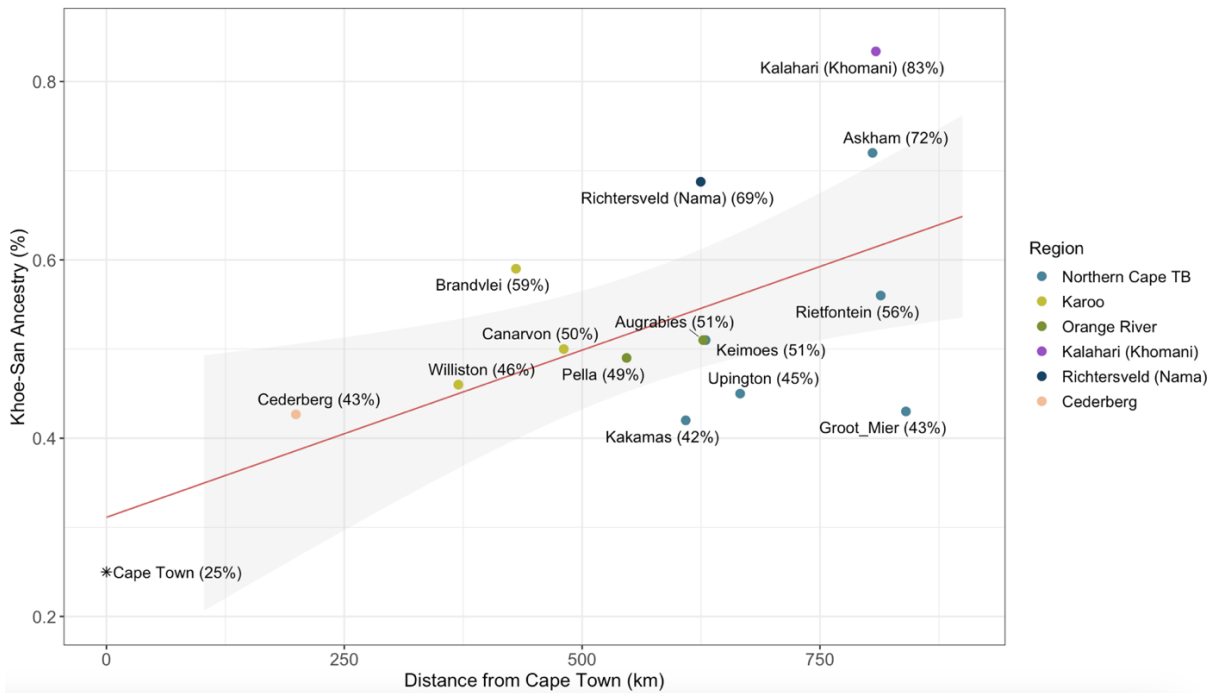
Characterizing sampled Coloured populations



Supplemental Figure 2.5) Identifying relatives' workflow. Visual of applied workflow used to identify relatives and subset samples into running groups (RG) for ADMIXTURE. Each RG contained the same set of unrelated individuals but a different subset of related individuals. They were all subject to 10 iterations of ADMIXTURE, all with a different initial starting seed and cross-validated five times. PLINK LD pruning was completed prior to sub-setting samples, but after identifying relatives. PLINK LD pruning not depicted in the workflow.

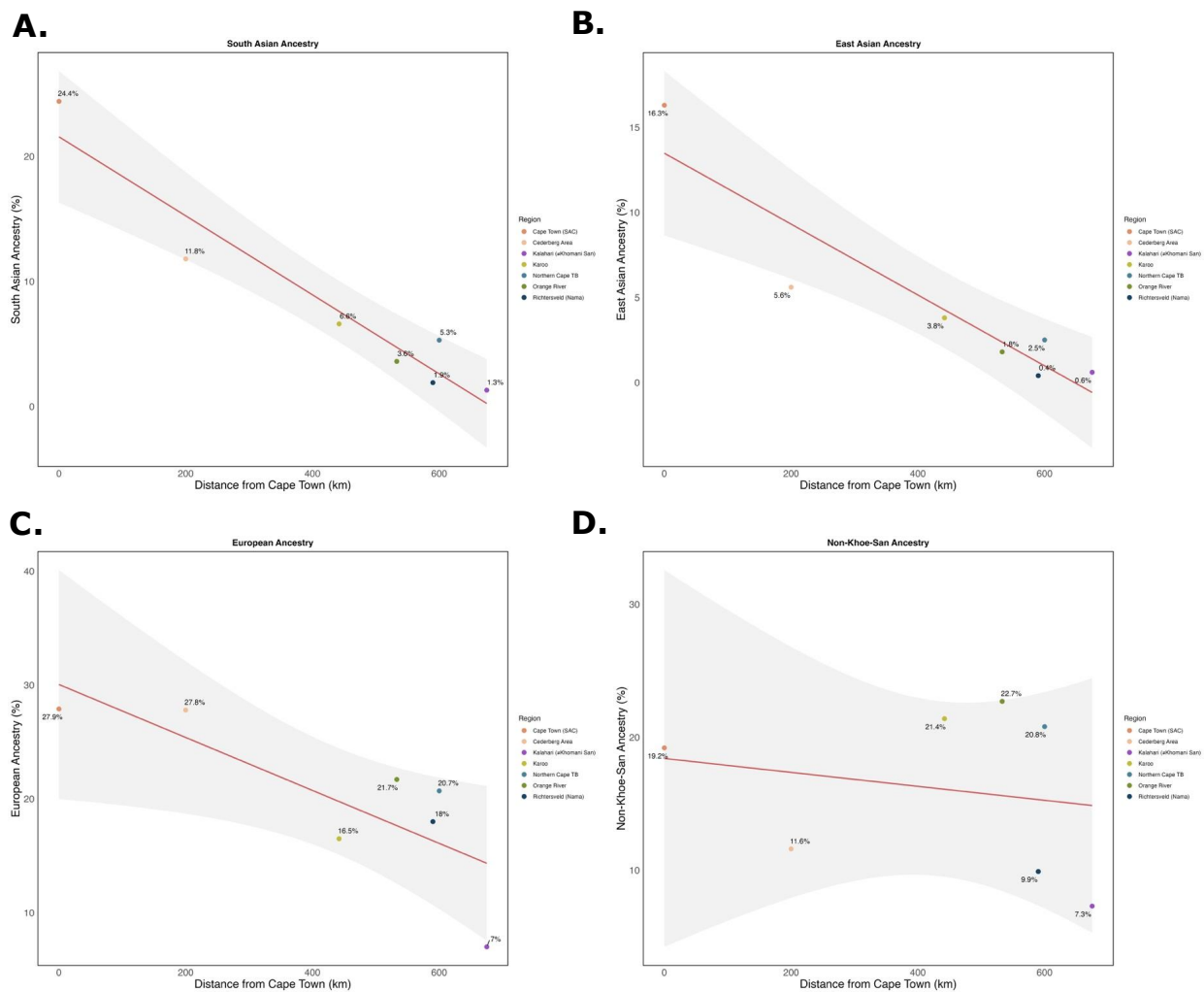


Supplemental Figure 2.7) Cross-validation error rate for ADMIXTURE visualized in



Supplemental Figure 2.8) Proportion of Khoe-San ancestry by distance from Cape Town.

Northern Cape Tuberculosis project participants are separated out by sampling site.



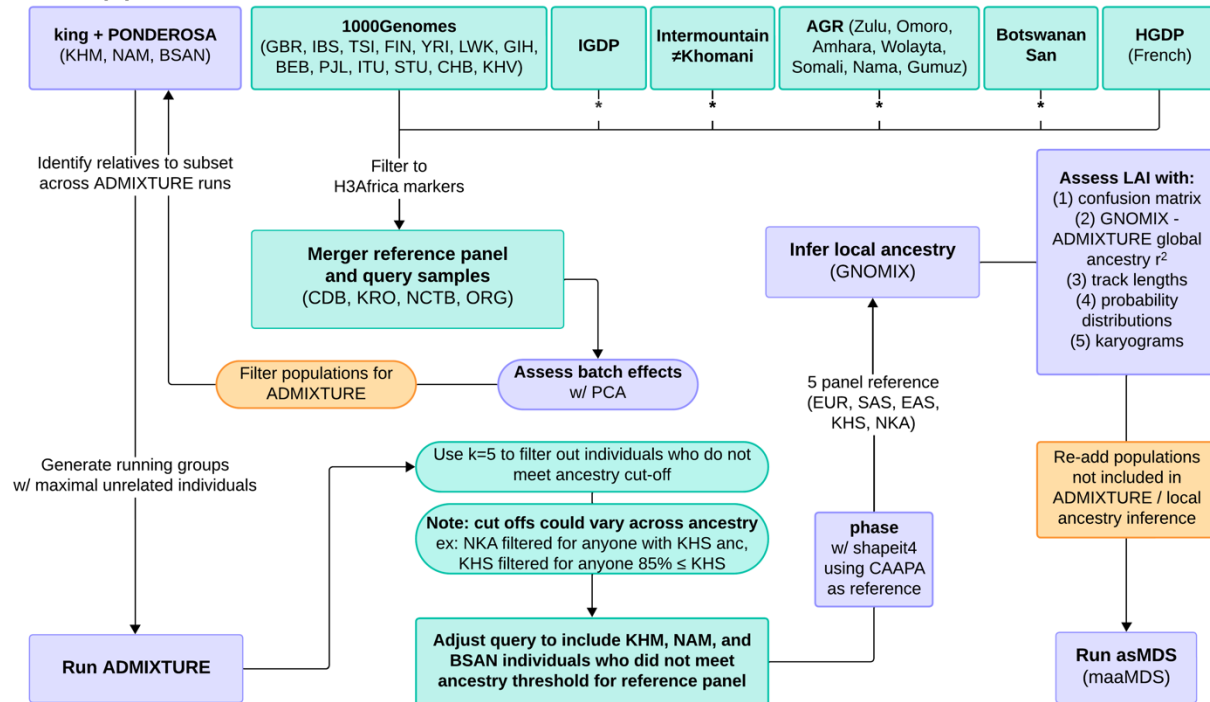
Supplemental Figure 2.9) Linear regression showing ancestry by distance for A) South Asian ($R^2 = 0.895$), B) East Asian ($R^2 = 0.847$), C) European ($R^2 = 0.0572$), and D) Non-Khoe-San Ancestry ($R^2 = 0.004$).

Local ancestry results

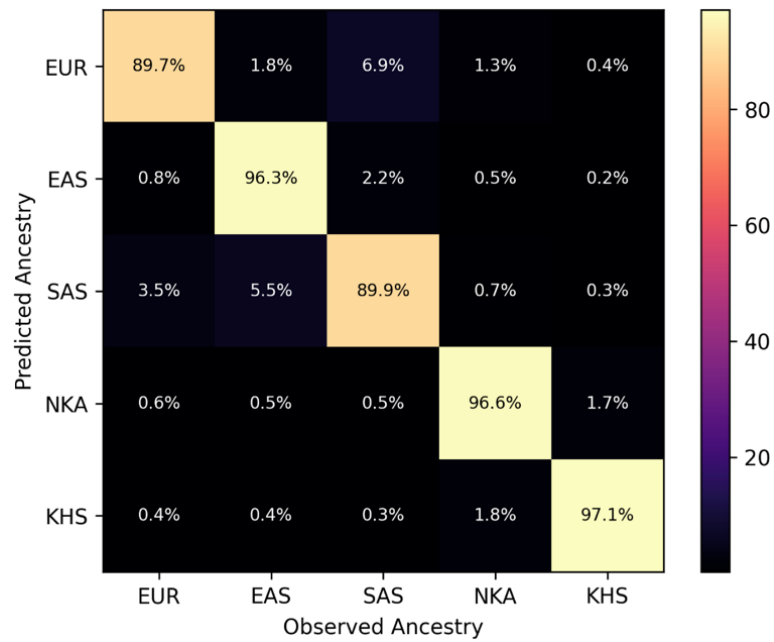
Our model appears well-calibrated, with the confusion matrix showing >96% accuracy in predicting EAS, NKA, and KHS haplotypes, and ~90% accuracy for EUR and SAS segments. Correlation between ADMIXTURE results for 5 population clusters and Gnomix global ancestry results is high for EUR, NKA, and KHS ($r^2 \geq 0.96$), albeit slightly lower for smaller segments like EAS ($r^2 = 0.94$) and SAS ($r^2 = 0.91$), still indicating a strong correlation.

The highest local ancestry misassignment occurs between SAS and EUR ancestries, with 6.9% of SAS segments being assigned to EUR and approximately 3.5% of EUR segments assigned to SAS. This misassignment also contributes to the larger variance in EUR and SAS ancestry probability distributions. We attributed this to the 'Indian cline,' referring to the geographical distribution of two primary ancestral groups across South Asia: the 'Ancestral North Indians' (ANI) and the 'Ancestral South Indians' (ASI)^{217–220}. The ANI share closer genetic similarities with populations from the Near Eastern and European populations, while the ASI are more closely related to East Asians^{217–220}, though many populations across South Asia retain a proportion of both ancestries. The ANI component among the BEB, STU, and GIH is visualized in our ADMIXTURE, highlighted in light green.

asMDS pipeline



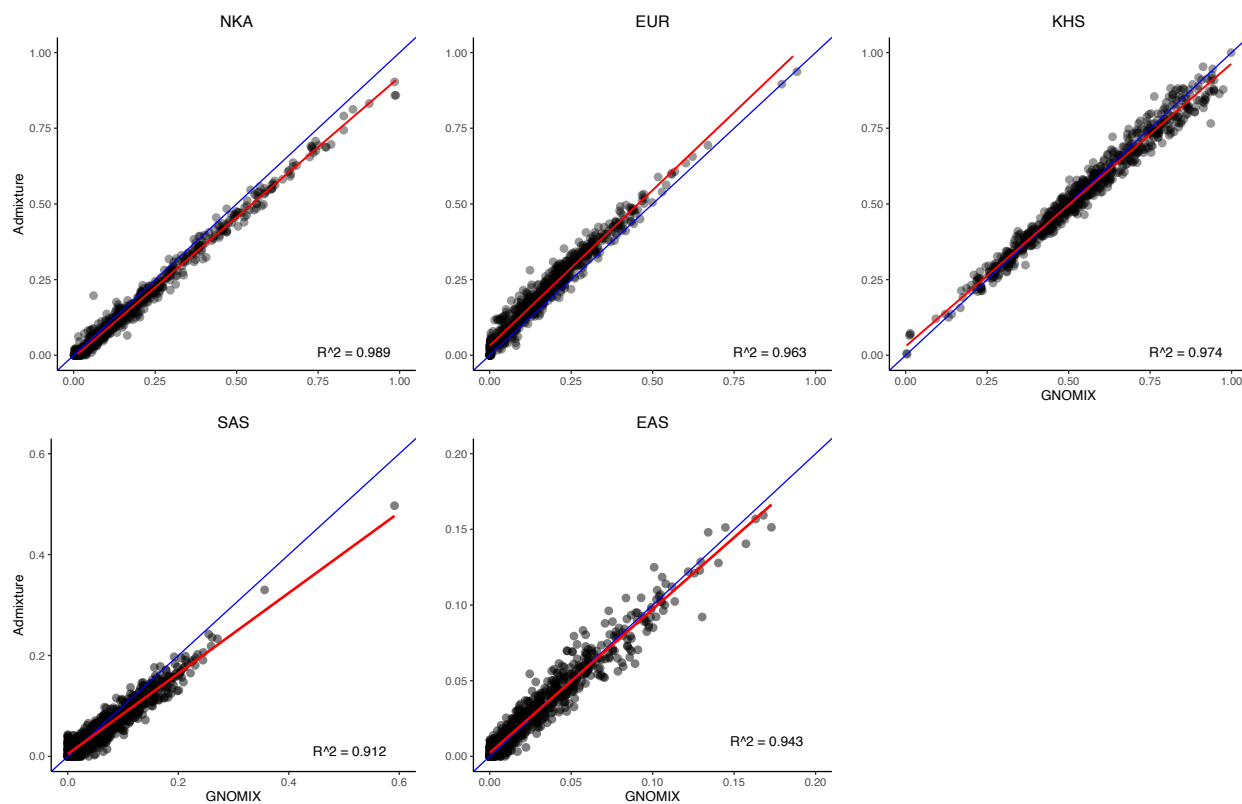
Supplemental Figure 2.10) Local Ancestry and ancestry-specific MDS pipeline. Shapes in turquoise highlight steps taken regarding sample consolidation and filtering, while orange shapes are steps taken to filter or add entire populations. Shapes in purple major analyses executed throughout the pipeline. The * within lines represent whole genomes that were lifted from HG19 to HG38.



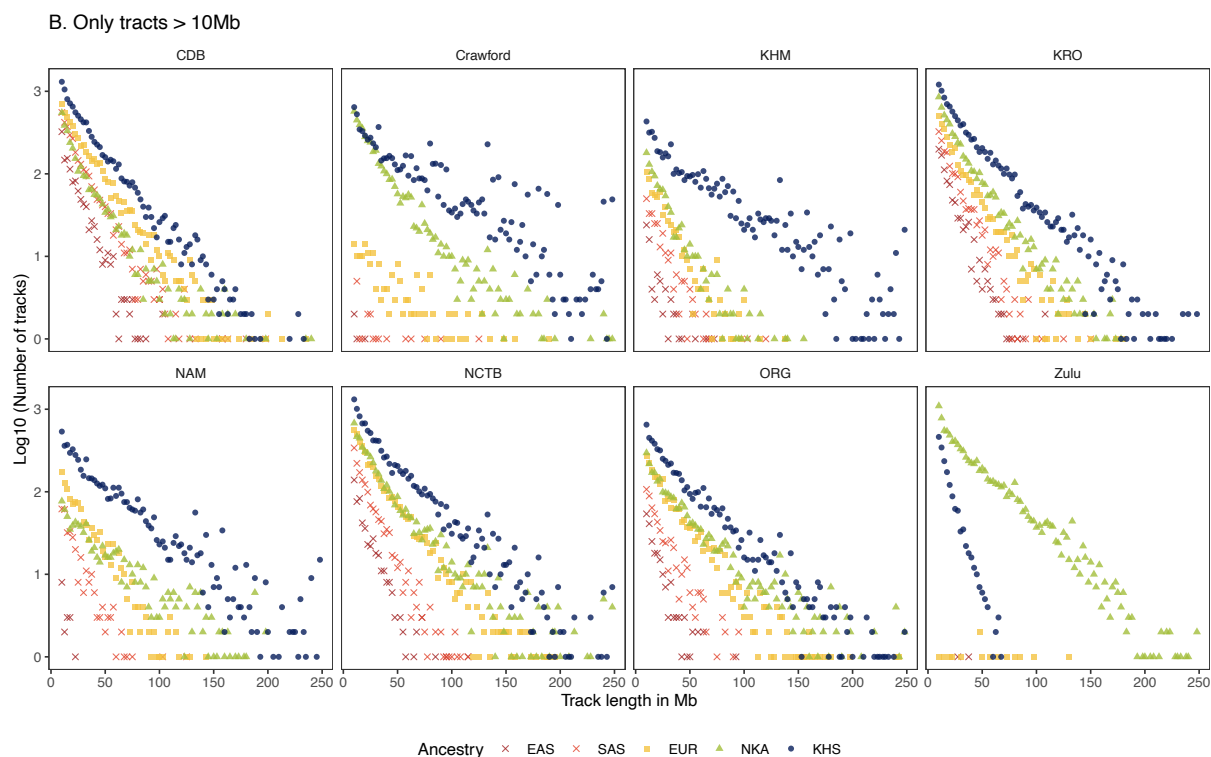
Supplemental Figure 2.11) Local ancestry inference confusion matrix normalized across all chromosomes, where the x-axis is the ancestry observed, and the y-axis is the model's predicted ancestry. The heat map highlights the accuracy of our model.

Pop	EAS	EUR	KHS	NKA	SAS
CDB	0.052	0.236	0.447	0.128	0.137
KRO	0.039	0.137	0.517	0.234	0.073
NCTB	0.023	0.174	0.517	0.224	0.062
ORG	0.017	0.176	0.523	0.238	0.045
KHM	0.008	0.067	0.789	0.110	0.026
NAM	0.003	0.116	0.729	0.121	0.032
BotswananSan	0.000	0.013	0.736	0.249	0.002
Zulu	0.000	0.002	0.156	0.841	0.000

Supplemental Figure 2.12) GNOMIX population-specific global ancestry average proportions.

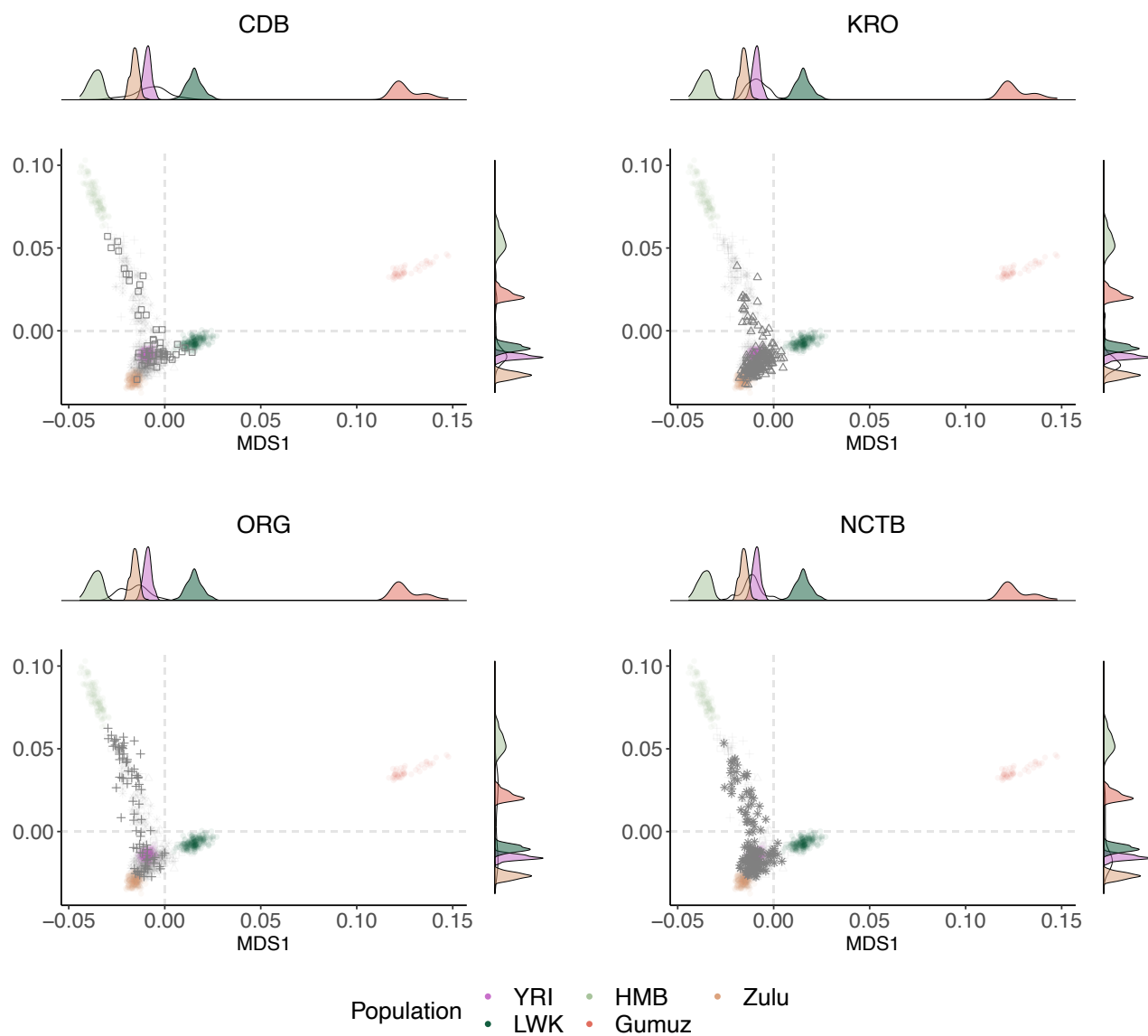


Supplemental Figure 2.13) Correlation between GNOMIX global ancestry results and ADMIXTURE, where the red line is the best fit and blue line is $x=y$. R-squared values for each ancestry is reported on the bottom right of the respective plot.



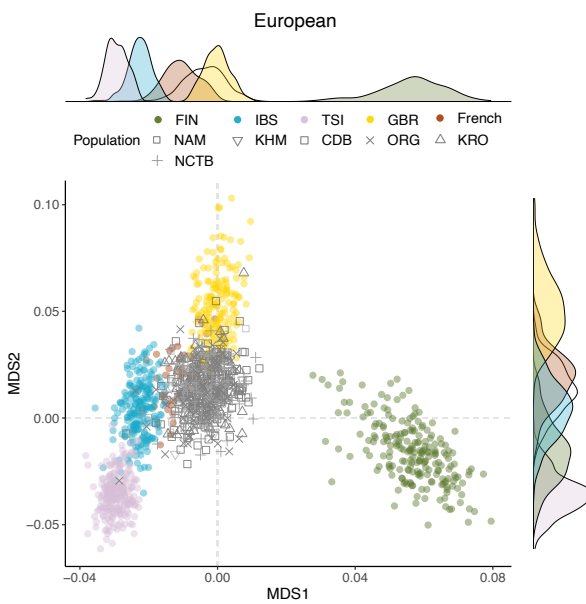
Supplemental Figure 2.14) Track lengths with at least 75% probability of the respective local ancestry and greater than 10Mb across query populations. The x-axis is the track length in Mb and the y-axis is the log10 of the number of tracks.

Ancestry-specific MDS results

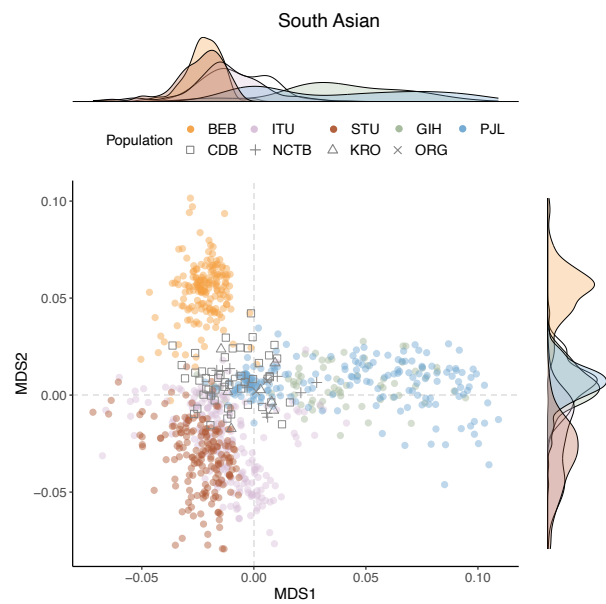


Supplemental Figure 2.15) NKA asMDS results by population. Density plots for MDS1 and MDS2 are provided with the respective query population density visualized using a transparent line.

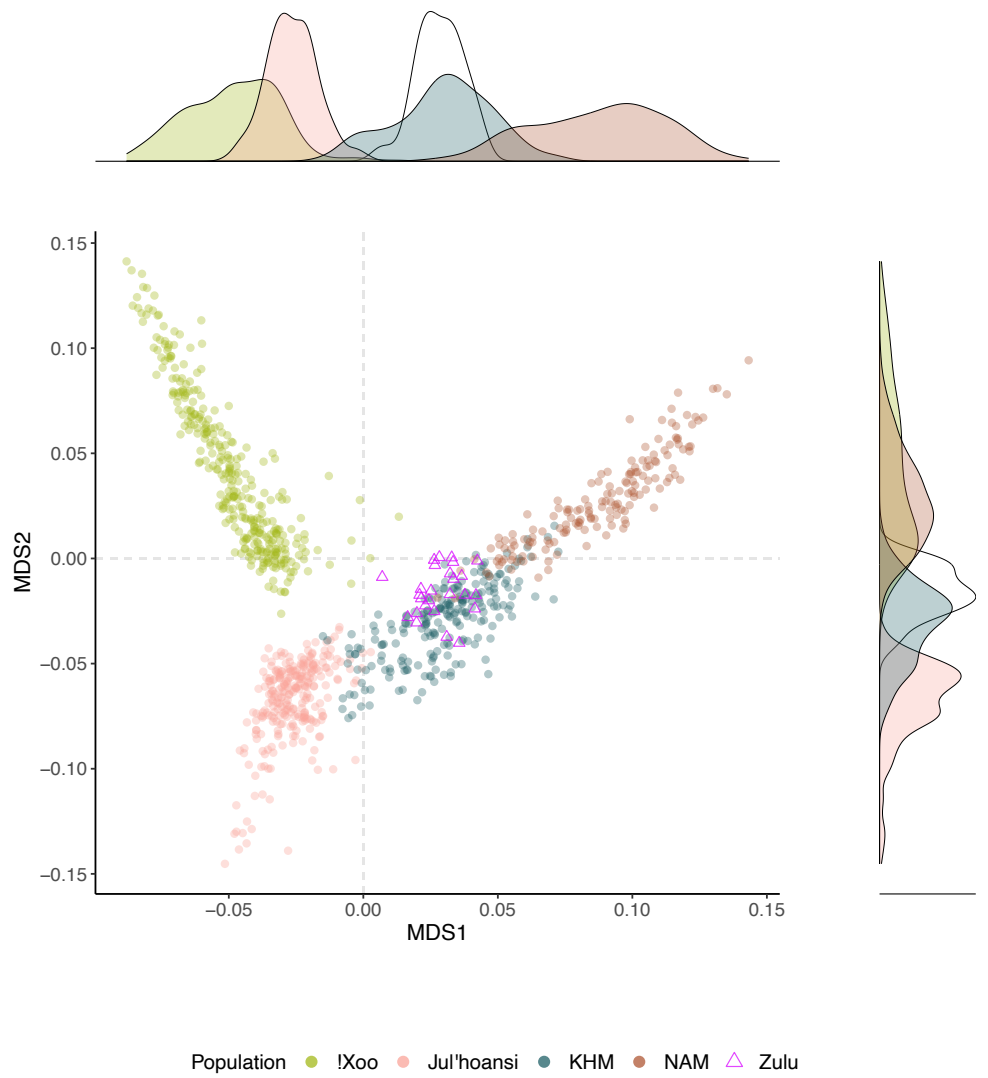
A.



B.



Supplemental Figure 2.16) European and South Asian asMDS. Local ancestry results were filtered for haplotypes with 75% probability for the respective ancestry and individuals were filtered for at least 20% for the ancestry under investigation.



Supplemental Figure 2.17) Zulu derived Khoe-San asMDS.

Admixture Proportions

Broad_Anc	Population	n
Query	Cederberg	170
Query	Karoo	171
Query	Northern Cape TB	183
Query	Middle Orange River	96
East African	Amhara	22
East African	Gumuz	23
East African	Oromo	24
East African	Somali	23
East African	Wolayta	24
East Asian	Han Chinese	103
East Asian	Vietnamese	99
European	British	91
European	Iberian	106
Southeast Asian	Indonesian	130
Khoe-San	Botswanan San	313
Khoe-San	≠Khomani San	115
Khoe-San	Nama	92
Niger-Congo	Himba	123
Niger-Congo	Luhya	99
Niger-Congo	Yoruba	108
Niger-Congo	Zulu	98
South Asian	Bengali	86
South Asian	Gujarati	103
South Asian	Tamil	102

Supplemental Table 2.1) Sample Size for each population included in ADMIXTURE analysis, as well as their respective broader ancestry (Broad_Anc) category. Newly collected samples have the broader ancestry “Query”.

Pop	SAS	EAS	NKA	KHS	EUR
CDB	0.118	0.056	0.116	0.432	0.278
KRO	0.066	0.038	0.214	0.518	0.165
ORG	0.036	0.018	0.227	0.502	0.217
NCTB	0.053	0.025	0.208	0.507	0.207
KHM	0.013	0.006	0.073	0.838	0.070
NAM	0.019	0.004	0.099	0.699	0.180
Crawford	0.002	0.001	0.129	0.859	0.009
Amhara	0.033	0.014	0.395	0.040	0.517
BEB	0.753	0.136	0.000	0.000	0.111
CHB	0.001	0.992	0.000	0.000	0.007
GBR	0.002	0.000	0.000	0.000	0.998
GIH	0.818	0.004	0.000	0.000	0.178
Gumuz	0.032	0.036	0.742	0.084	0.106
HMB	0.005	0.006	0.951	0.031	0.007
IBS	0.002	0.000	0.014	0.000	0.983
IND	0.116	0.863	0.011	0.009	0.001
KHV	0.012	0.986	0.000	0.000	0.002
LWK	0.015	0.011	0.887	0.036	0.050
Oromo	0.029	0.017	0.420	0.051	0.483
STU	0.864	0.040	0.000	0.000	0.096
Somali	0.013	0.018	0.507	0.036	0.425
Wolayta	0.035	0.020	0.460	0.081	0.404
YRI	0.015	0.008	0.956	0.005	0.016
Zulu	0.007	0.004	0.794	0.182	0.012

Supplemental Table 2.2) ADMIXTURE k=5

Pop	EAFR	EUR	SAS	IND	WBNT	EBNT	KHS	EAS
CDB	0.024	0.266	0.117	0.031	0.015	0.097	0.420	0.031
KRO	0.010	0.158	0.063	0.021	0.010	0.212	0.505	0.021
ORG	0.034	0.207	0.034	0.010	0.095	0.124	0.486	0.010
NCTB	0.020	0.199	0.051	0.013	0.034	0.175	0.494	0.014
KHM	0.015	0.066	0.013	0.003	0.006	0.075	0.819	0.004
NAM	0.065	0.160	0.017	0.002	0.056	0.031	0.668	0.001
Crawford	0.015	0.007	0.001	0.000	0.039	0.100	0.838	0.000
Amhara	0.574	0.413	0.010	0.001	0.000	0.000	0.002	0.000
BEB	0.000	0.064	0.793	0.030	0.000	0.000	0.000	0.112
CHB	0.000	0.005	0.004	0.002	0.000	0.000	0.000	0.989
GBR	0.000	0.997	0.003	0.000	0.000	0.000	0.000	0.000
GIH	0.000	0.126	0.868	0.002	0.000	0.000	0.000	0.003
Gumuz	0.833	0.000	0.000	0.000	0.007	0.127	0.032	0.001
HMB	0.012	0.003	0.003	0.002	0.872	0.092	0.014	0.002
IBS	0.023	0.974	0.002	0.000	0.000	0.001	0.000	0.000
IND	0.001	0.001	0.011	0.617	0.000	0.000	0.000	0.369
KHV	0.000	0.001	0.006	0.086	0.000	0.000	0.000	0.907
LWK	0.248	0.008	0.002	0.002	0.075	0.640	0.022	0.002
Oromo	0.603	0.376	0.008	0.001	0.000	0.002	0.010	0.000
STU	0.001	0.044	0.908	0.029	0.000	0.000	0.000	0.018
Somali	0.693	0.292	0.001	0.000	0.001	0.012	0.000	0.000
Wolayta	0.647	0.293	0.012	0.002	0.002	0.005	0.038	0.001
YRI	0.108	0.001	0.002	0.002	0.121	0.764	0.001	0.002
Zulu	0.011	0.007	0.002	0.001	0.010	0.787	0.181	0.002

Supplemental Table 2.3) ADMIXTURE k=8

Pop	S_SAN	WBNT	EAS	EAFR	IND1	SAS	N_SAN	EUR	EBNT	IND2
CDB	0.420	0.018	0.014	0.020	0.030	0.121	0.026	0.241	0.093	0.017
KRO	0.466	0.012	0.008	0.008	0.023	0.065	0.068	0.133	0.207	0.011
ORG	0.468	0.098	0.007	0.029	0.008	0.036	0.047	0.180	0.120	0.006
NCTB	0.431	0.036	0.006	0.017	0.014	0.053	0.090	0.177	0.169	0.007
KHM	0.600	0.006	0.002	0.011	0.003	0.013	0.255	0.042	0.066	0.002
NAM	0.696	0.057	0.002	0.054	0.001	0.019	0.022	0.119	0.030	0.001
Crawford	0.100	0.016	0.000	0.013	0.000	0.001	0.764	0.007	0.099	0.000
Amhara	0.003	0.000	0.000	0.575	0.000	0.010	0.000	0.410	0.000	0.000
BEB	0.000	0.000	0.090	0.000	0.034	0.793	0.000	0.064	0.000	0.019
CHB	0.000	0.000	0.951	0.000	0.045	0.001	0.000	0.002	0.000	0.000
GBR	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.997	0.000	0.000
GIH	0.000	0.000	0.002	0.000	0.002	0.868	0.000	0.127	0.000	0.002
Gumuz	0.000	0.003	0.000	0.830	0.000	0.000	0.038	0.000	0.128	0.000
HMB	0.009	0.871	0.002	0.012	0.002	0.002	0.007	0.002	0.090	0.002
IBS	0.000	0.000	0.000	0.023	0.000	0.001	0.000	0.974	0.001	0.000
IND	0.000	0.000	0.077	0.000	0.459	0.012	0.000	0.001	0.000	0.450
KHV	0.000	0.000	0.651	0.000	0.335	0.008	0.000	0.001	0.000	0.005
LWK	0.008	0.068	0.001	0.253	0.002	0.002	0.016	0.011	0.638	0.002
Oromo	0.009	0.000	0.000	0.603	0.001	0.008	0.004	0.373	0.002	0.001
STU	0.000	0.000	0.013	0.001	0.013	0.908	0.000	0.044	0.000	0.022
Somali	0.001	0.001	0.000	0.696	0.000	0.001	0.000	0.289	0.012	0.000
Wolayta	0.021	0.001	0.001	0.647	0.001	0.012	0.021	0.289	0.005	0.002
YRI	0.001	0.112	0.002	0.114	0.001	0.003	0.001	0.002	0.761	0.003
Zulu	0.167	0.008	0.001	0.013	0.001	0.001	0.023	0.004	0.780	0.001

Supplemental Table 2.4) ADMIXTURE k=10

Population Centroids and Euclidean Distances

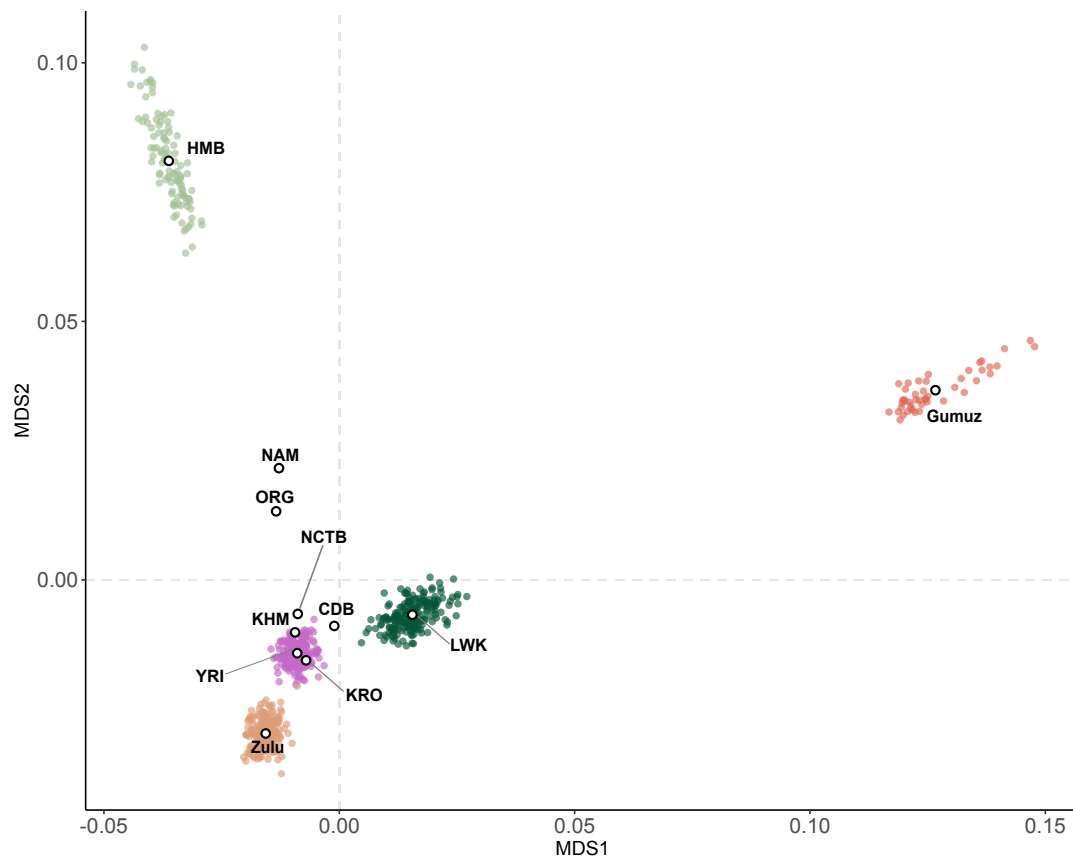
We aim to best characterize specific source populations for the five primary ancestries within the admixed study populations. We further investigated the relationships among populations by computing the Euclidean distances between each population's centroid (i.e., a population's mean coordinates along MDS1 and MDS2; see *Supplemental Methods*). For instance, the distances from the ≠Khomani San to the Cederberg (CDB), Karoo (KRO), and Orange River (ORG) populations are 0.036, 0.022, and 0.034, respectively, whereas the corresponding distances from NAM to these populations are 0.043, 0.059, and 0.046. These results suggest that, although Cederberg, Karoo, and Orange River are relatively similar in their overall MDS distribution, they share a more similar Khoe-San ancestry profile with ≠Khomani San than with the Nama. The Karoo (0.022) and NCTB (0.021) show the smallest distance to ≠Khomani San, indicating particularly close clustering. Additionally, Zulu haplotypes fall with the ≠Khomani San (0.010), suggesting that the Khoe-San haplotypes present in Zulu individuals are genetically similar to ≠Khomani San.

Among the query populations, we observe generally small distances to Yoruba (YRI), Luhya (LWK), and Zulu, compared to the Himba (HMB). Haplotypes from individuals sampled in the Karoo (0.002) demonstrate close similarities to Yoruba, while the Cederberg centroid appears slightly further from the Yoruba (0.009) and Luhya (0.017) (*Supplemental Table 2.2*). Participants sampled along Middle Orange River are more distant from the Yoruba (0.028) and Zulu (0.043), but still more similar to these groups than to Himba (0.072). Overall, these patterns indicate that the Cederberg, Karoo, and Orange River populations share closer ancestry profiles with Yoruba, Luhya, and Zulu than with Himba. Similarly, ≠Khomani San NKA haplotype centroid is closest to the Yoruba (0.004) and moderately more distant to the Zulu (0.021) and

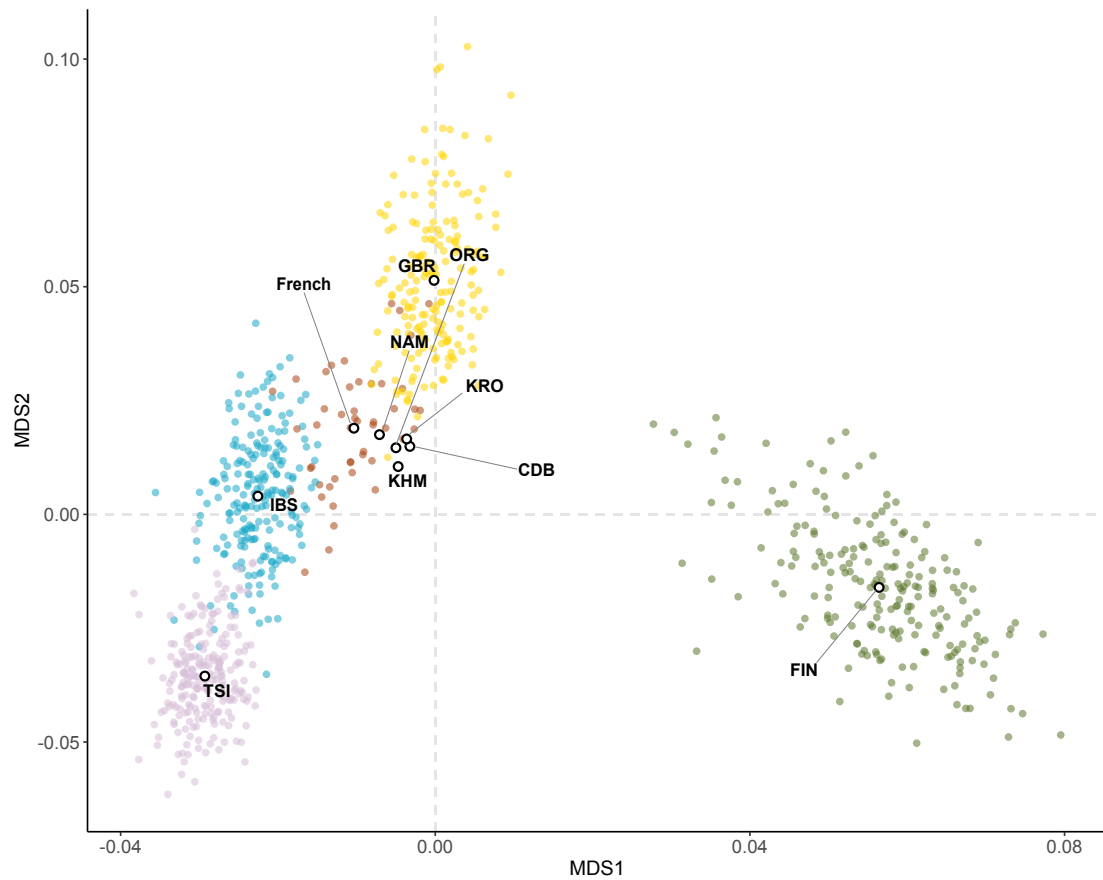
Luhya (0.025), with increasing distance to Himba (0.095). In contrast, the Nama NKA haplotype centroid is moderately distant to all NKA reference groups: Yoruba (0.036), Zulu (0.051), Luhya (0.040), Himba (0.064).



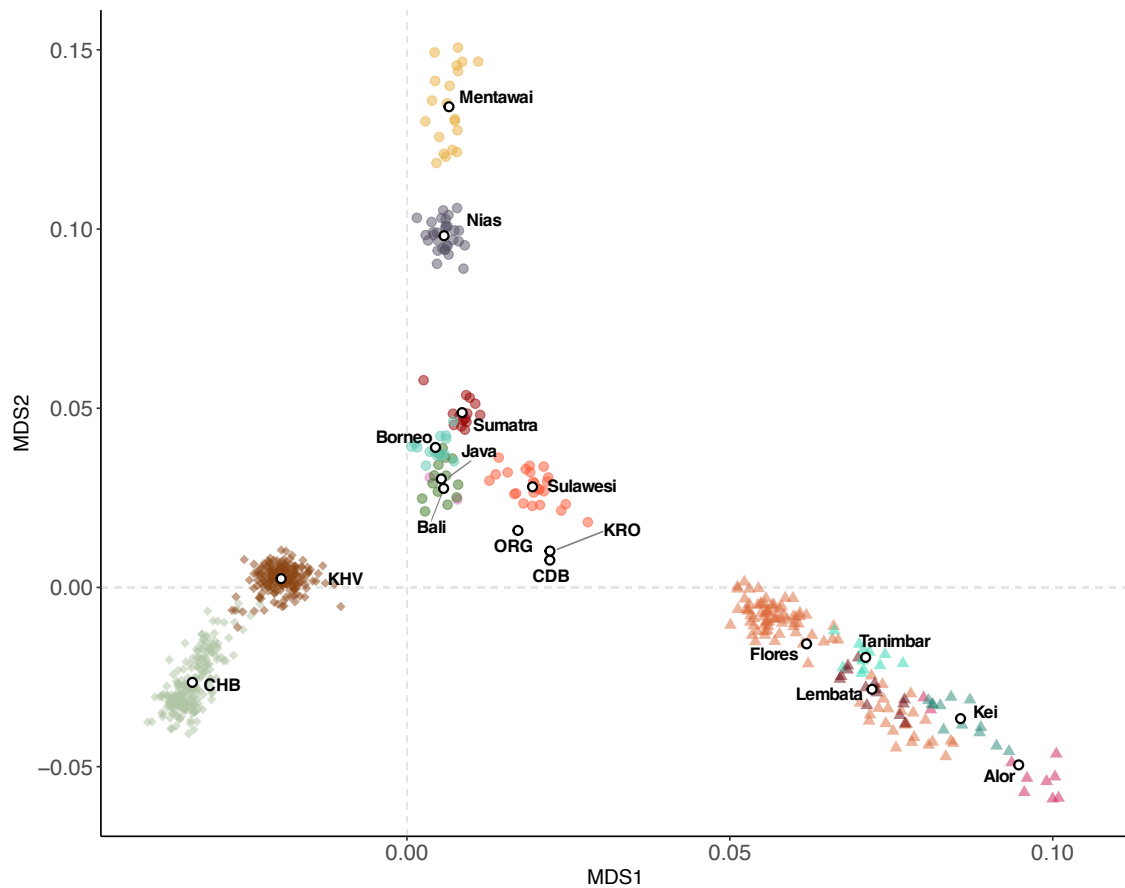
Supplemental Figure 2.18) Khoisan derived population centroids plotted against reference populations. Query sample haplotypes are not plotted for clearer visualization of reference populations distributions.



Supplemental Figure 2.19) Non-Khoe-San derived population centroids plotted against non-Khoe-San reference populations. Query sample haplotypes are not plotted for clearer visualization of reference populations distributions.

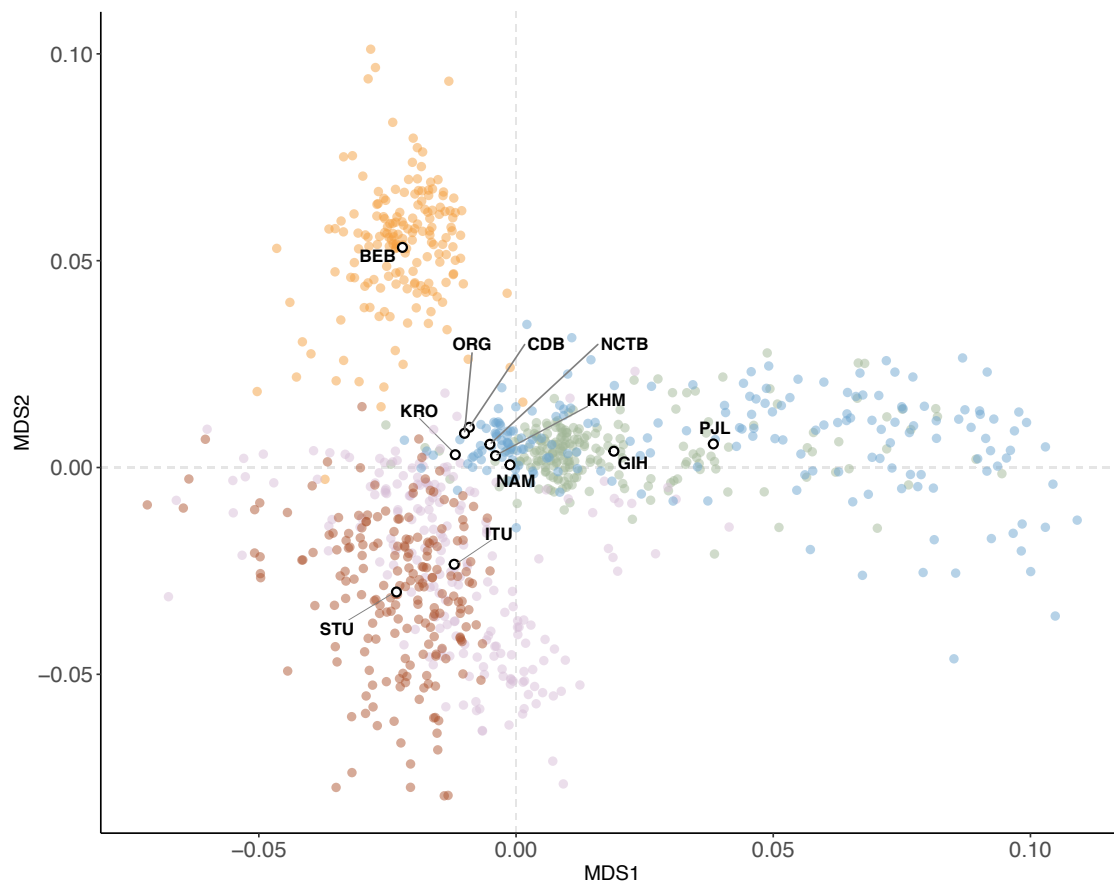


Supplemental Figure 2.20) European derived centroids plotted against reference populations. Query sample haplotypes are not plotted for clearer visualization of reference populations distributions.



Supplemental Figure 2.21) Other Asian derived centroids plotted against reference populations.

Query sample haplotypes are not plotted for clearer visualization of reference populations distributions.



Supplemental Figure 2.22) South Asian derived centroids plotted against reference populations. Query sample haplotypes are not plotted for clearer visualization of reference populations distributions.

	KHM	NAM	CDB	KRO	ORG	NCTB	!Xoo	Jul'hoansi	Zulu
KHM	0.000	0.080	0.036	0.022	0.034	0.023	0.100	0.067	0.010
NAM	0.080	0.000	0.043	0.059	0.046	0.058	0.131	0.146	0.073
CDB	0.036	0.043	0.000	0.016	0.005	0.014	0.108	0.103	0.030
KRO	0.022	0.059	0.016	0.000	0.012	0.002	0.098	0.087	0.014
ORG	0.034	0.046	0.005	0.012	0.000	0.011	0.103	0.099	0.027
NCTB	0.023	0.058	0.014	0.002	0.011	0.000	0.100	0.088	0.015
!Xoo	0.100	0.131	0.108	0.098	0.103	0.100	0.000	0.104	0.094
Jul'hoansi	0.067	0.146	0.103	0.087	0.099	0.088	0.104	0.000	0.073
Zulu	0.010	0.073	0.030	0.014	0.027	0.015	0.094	0.073	0.000

Supplemental Table 2.5) Khoe-San_Euclidean distances between population centroids using Khoe-San ancestry-specific MDS1 and MDS2 coordinates.

	KHM	NAM	CDB	KRO	ORG	NCTB	Gumuz	HMB	LWK	YRI	Zulu
KHM	0.000	0.032	0.008	0.006	0.024	0.004	0.144	0.095	0.025	0.004	0.021
NAM	0.032	0.000	0.033	0.038	0.008	0.028	0.140	0.064	0.040	0.036	0.051
CDB	0.008	0.033	0.000	0.009	0.025	0.008	0.136	0.097	0.017	0.009	0.025
KRO	0.006	0.038	0.009	0.000	0.030	0.009	0.144	0.101	0.024	0.002	0.017
ORG	0.024	0.008	0.025	0.030	0.000	0.020	0.142	0.072	0.035	0.028	0.043
NCTB	0.004	0.028	0.008	0.009	0.020	0.000	0.142	0.092	0.024	0.008	0.024
Gumuz	0.144	0.140	0.136	0.144	0.142	0.142	0.000	0.169	0.119	0.145	0.157
HMB	0.095	0.064	0.097	0.101	0.072	0.092	0.169	0.000	0.102	0.099	0.113
LWK	0.025	0.040	0.017	0.024	0.035	0.024	0.119	0.102	0.000	0.026	0.039
YRI	0.004	0.036	0.009	0.002	0.028	0.008	0.145	0.099	0.026	0.000	0.017
Zulu	0.021	0.051	0.025	0.017	0.043	0.024	0.157	0.113	0.039	0.017	0.000

Supplemental Table 2.6) Non-Khoe-San African Euclidean distances between population centroids using non-Khoe-San African ancestry-specific MDS1 and MDS2 coordinates.

	KHM	NAM	CDB	KRO	ORG	NCTB	FIN	French	GBR	IBS	TSI
KHM	0.000	0.005	0.003	0.004	0.002	0.003	0.066	0.011	0.042	0.020	0.052
NAM	0.005	0.000	0.003	0.003	0.003	0.003	0.070	0.006	0.037	0.020	0.055
CDB	0.003	0.003	0.000	0.001	0.002	0.000	0.067	0.009	0.038	0.021	0.055
KRO	0.004	0.003	0.001	0.000	0.002	0.001	0.067	0.008	0.037	0.022	0.056
ORG	0.002	0.003	0.002	0.002	0.000	0.001	0.067	0.009	0.040	0.020	0.054
NCTB	0.003	0.003	0.000	0.001	0.001	0.000	0.067	0.009	0.039	0.021	0.055
FIN	0.066	0.070	0.067	0.067	0.067	0.067	0.000	0.075	0.088	0.082	0.088
French	0.011	0.006	0.009	0.008	0.009	0.009	0.075	0.000	0.034	0.019	0.058
GBR	0.042	0.037	0.038	0.037	0.040	0.039	0.088	0.034	0.000	0.052	0.092
IBS	0.020	0.020	0.021	0.022	0.020	0.021	0.082	0.019	0.052	0.000	0.040
TSI	0.052	0.055	0.055	0.056	0.054	0.055	0.088	0.058	0.092	0.040	0.000

Supplemental Table 2.7) European Euclidean distances between population centroids using European ancestry-specific MDS1 and MDS2 coordinates.

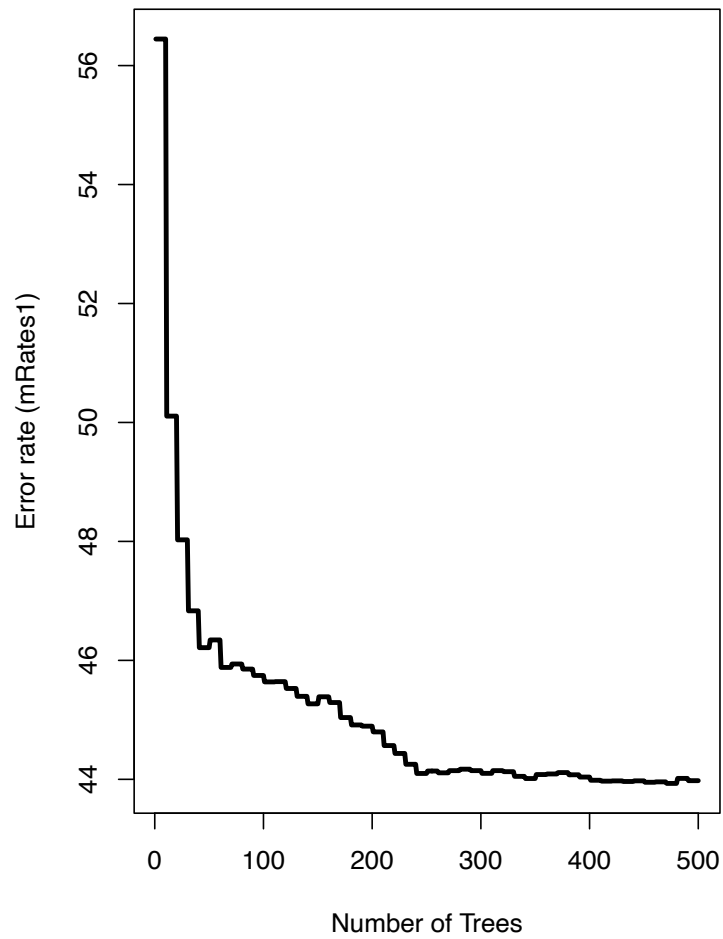
	KHM	NAM	CDB	KRO	ORG	NCTB	BEB	GIH	ITU	PJL	STU
KHM	0.000	0.004	0.009	0.008	0.008	0.003	0.054	0.023	0.027	0.042	0.038
NAM	0.004	0.000	0.012	0.011	0.012	0.006	0.057	0.020	0.026	0.040	0.038
CDB	0.009	0.012	0.000	0.007	0.002	0.006	0.045	0.029	0.033	0.048	0.042
KRO	0.008	0.011	0.007	0.000	0.005	0.007	0.051	0.031	0.026	0.050	0.035
ORG	0.008	0.012	0.002	0.005	0.000	0.006	0.047	0.029	0.032	0.048	0.041
NCTB	0.003	0.006	0.006	0.007	0.006	0.000	0.051	0.024	0.030	0.043	0.040
BEB	0.054	0.057	0.045	0.051	0.047	0.051	0.000	0.064	0.077	0.077	0.083
GIH	0.023	0.020	0.029	0.031	0.029	0.024	0.064	0.000	0.041	0.019	0.054
ITU	0.027	0.026	0.033	0.026	0.032	0.030	0.077	0.041	0.000	0.058	0.013
PJL	0.042	0.040	0.048	0.050	0.048	0.043	0.077	0.019	0.058	0.000	0.071
STU	0.038	0.038	0.042	0.035	0.041	0.040	0.083	0.054	0.013	0.071	0.000

Supplemental Table 2.8) South Asian Euclidean distances between population centroids using South Asian ancestry-specific MDS1 and MDS2 coordinates.

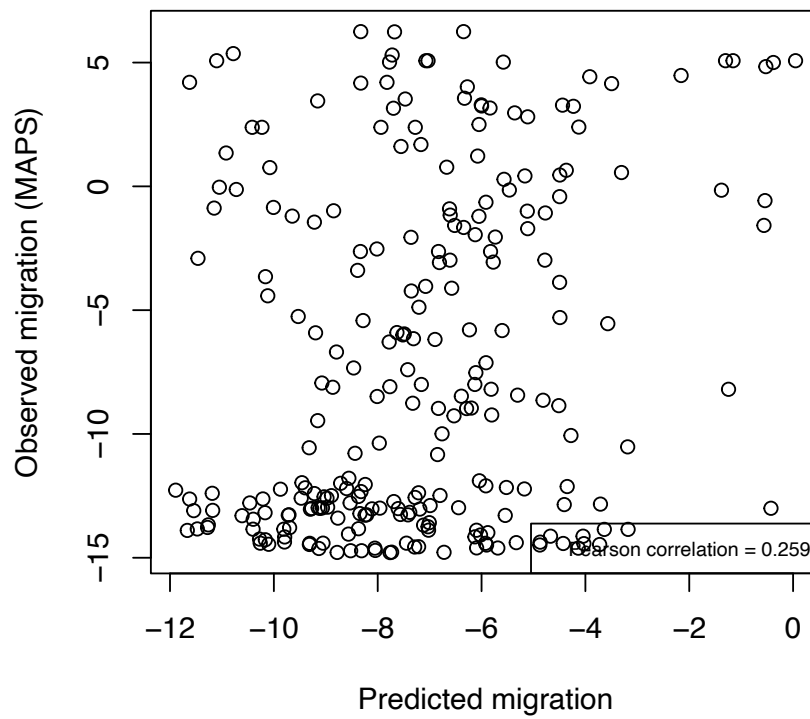
	CDB	KRO	ORG	NCTB	Sulawesi	Java	Flores	Sumatra	Alor	Bali	Borneo	CHB	KHV	Kei	Lembata	Mentawai	Nias	Tanimbar
CDB	0.000	0.003	0.010	0.010	0.021	0.028	0.046	0.043	0.092	0.026	0.036	0.065	0.042	0.077	0.062	0.127	0.092	0.056
KRO	0.003	0.000	0.008	0.008	0.018	0.026	0.047	0.041	0.094	0.024	0.034	0.066	0.042	0.079	0.063	0.125	0.089	0.057
ORG	0.010	0.008	0.000	0.003	0.012	0.019	0.055	0.034	0.101	0.016	0.026	0.066	0.039	0.086	0.070	0.119	0.083	0.064
NCTB	0.010	0.008	0.003	0.000	0.011	0.019	0.054	0.033	0.101	0.017	0.026	0.068	0.041	0.086	0.070	0.117	0.082	0.064
Sulawesi	0.021	0.018	0.012	0.011	0.000	0.014	0.061	0.023	0.108	0.014	0.019	0.076	0.047	0.093	0.077	0.107	0.071	0.070
Java	0.028	0.026	0.019	0.019	0.014	0.000	0.073	0.019	0.120	0.003	0.009	0.069	0.037	0.105	0.089	0.104	0.068	0.082
Flores	0.046	0.047	0.055	0.054	0.061	0.073	0.000	0.084	0.047	0.071	0.079	0.096	0.083	0.032	0.016	0.160	0.127	0.010
Sumatra	0.043	0.041	0.034	0.033	0.023	0.019	0.084	0.000	0.131	0.021	0.011	0.086	0.054	0.115	0.100	0.085	0.049	0.093
Alor	0.092	0.094	0.101	0.101	0.108	0.120	0.047	0.131	0.000	0.118	0.126	0.130	0.125	0.016	0.031	0.204	0.172	0.038
Bali	0.026	0.024	0.016	0.017	0.014	0.003	0.071	0.021	0.118	0.000	0.012	0.067	0.036	0.103	0.087	0.107	0.071	0.081
Borneo	0.036	0.034	0.026	0.026	0.019	0.009	0.079	0.011	0.126	0.012	0.000	0.076	0.044	0.111	0.096	0.095	0.059	0.089
CHB	0.065	0.066	0.066	0.068	0.076	0.069	0.096	0.086	0.130	0.067	0.076	0.000	0.032	0.119	0.105	0.165	0.131	0.104
KHV	0.042	0.042	0.039	0.041	0.047	0.037	0.083	0.054	0.125	0.036	0.044	0.032	0.000	0.112	0.097	0.134	0.099	0.093
Kei	0.077	0.079	0.086	0.086	0.093	0.105	0.032	0.115	0.016	0.103	0.111	0.119	0.112	0.000	0.016	0.188	0.157	0.023
Lembata	0.062	0.063	0.070	0.070	0.077	0.089	0.016	0.100	0.031	0.087	0.096	0.105	0.097	0.016	0.000	0.175	0.143	0.009
Mentawai	0.127	0.125	0.119	0.117	0.107	0.104	0.160	0.085	0.204	0.107	0.095	0.165	0.134	0.188	0.175	0.000	0.036	0.167
Nias	0.092	0.089	0.083	0.082	0.071	0.068	0.127	0.049	0.172	0.071	0.059	0.131	0.099	0.157	0.143	0.036	0.000	0.135
Tanimbar	0.056	0.057	0.064	0.064	0.070	0.082	0.010	0.093	0.038	0.081	0.089	0.104	0.093	0.023	0.009	0.167	0.135	0.000

Supplemental Table 2.9) Southeast and East Asian distances. Euclidean distances between population centroids using Southeast and East Asian ancestry-specific MDS1 and MDS2 coordinates.

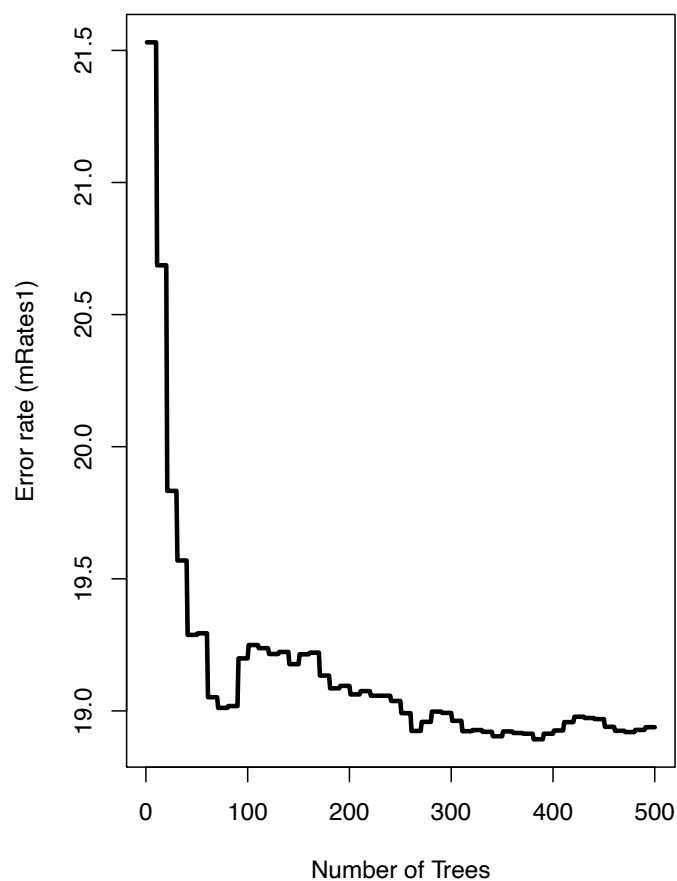
SPRUCÉ



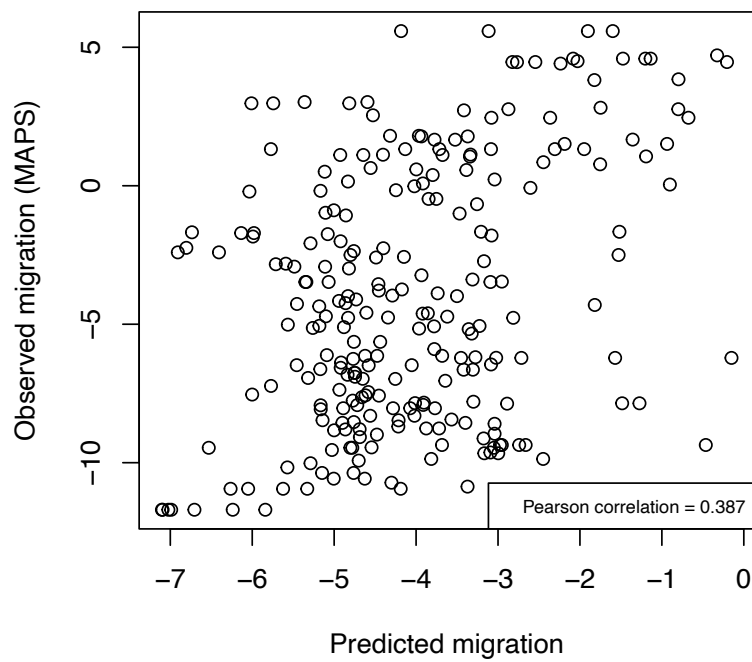
Supplemental Figure 2.23) Error rate plot showing the error rate decreasing as the number of trees in the random forest increases for IBD segments 2-6 cM representing 56-13 generations ago.



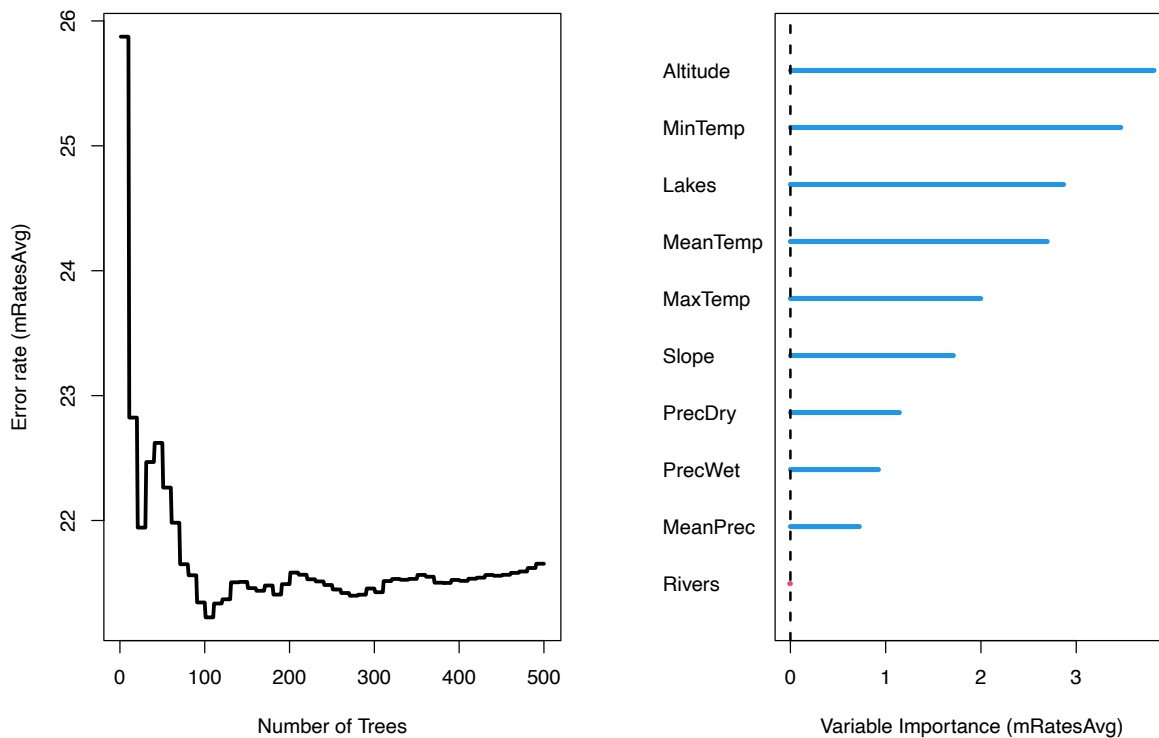
Supplemental Figure 2.24) Scatterplot showing predicted migration plotted against observed migration (Pearson correlations = 0.259) for IBD segments 2-6 cM/56-13 generations ago.



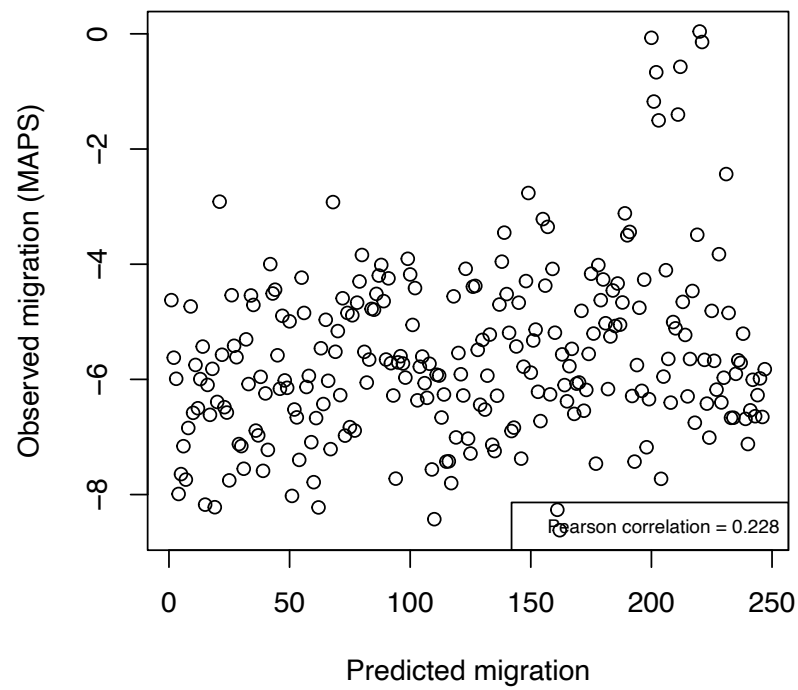
Supplemental Figure 2.25) Error rate plot showing the error rate decreasing as the number of trees in the random forest increases for IBD segments greater than 6 cM representing less than 13 generations ago.



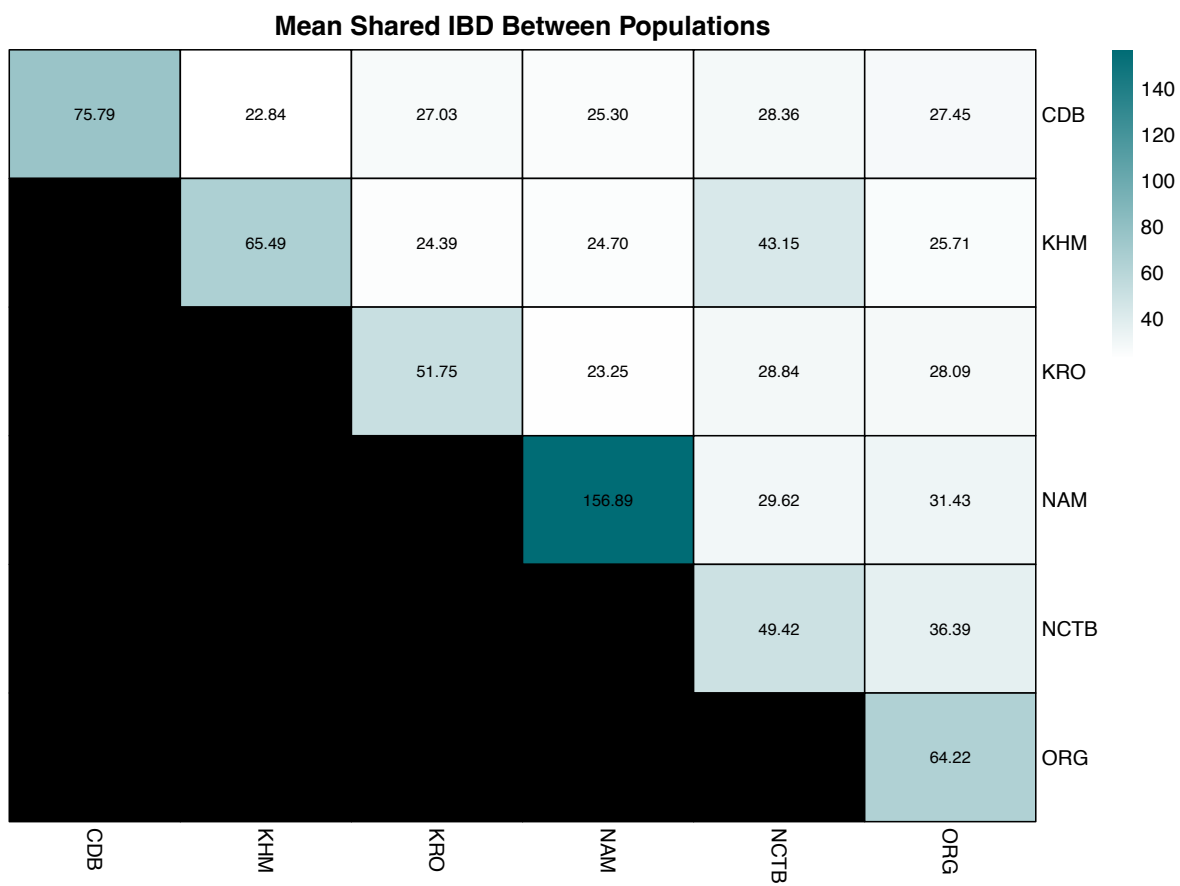
Supplemental Figure 2.26) Scatterplot showing predicted migration plotted against observed migration (Pearson correlations = 0.387) for IBD segments greater than 6 cM/ less than 13 generations ago.



Supplemental Figure 2.27) SPRUCE Random forest importance table average across time periods (2-6cM and greater than 6cM), 56> generations ago showing altitude and minimum temperature during the coldest month being the most important variables explaining the variation in estimated migration rate.

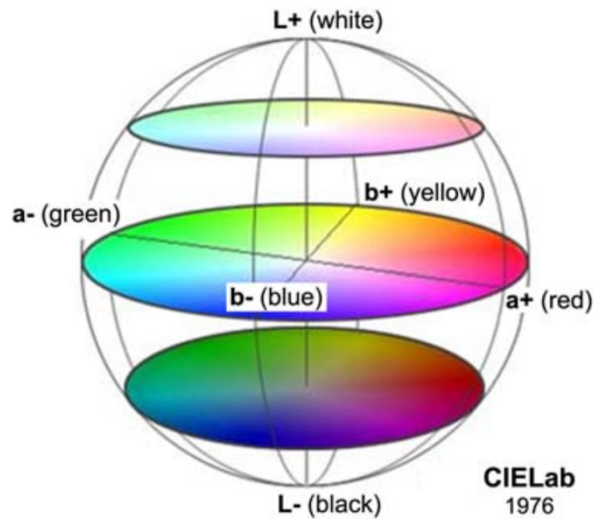


Supplemental Figure 2.28) Scatterplot showing predicted migration plotted against observed migration (Pearson correlations = 0.228) on average.

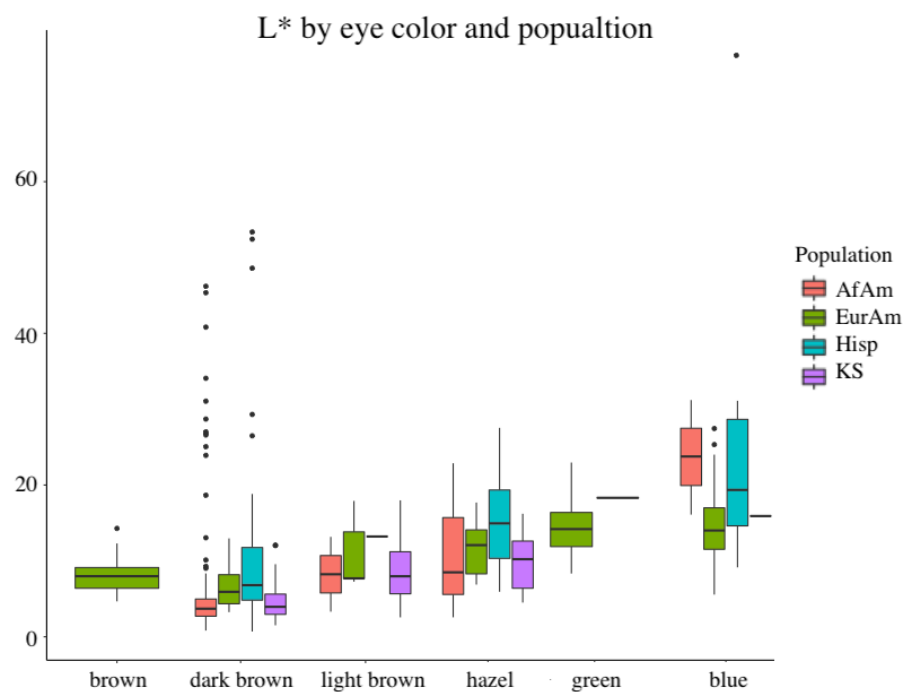


Supplemental Figure 2.29) Mean shared IBD length between and within Khoe-San and Khoe-San descendant communities.

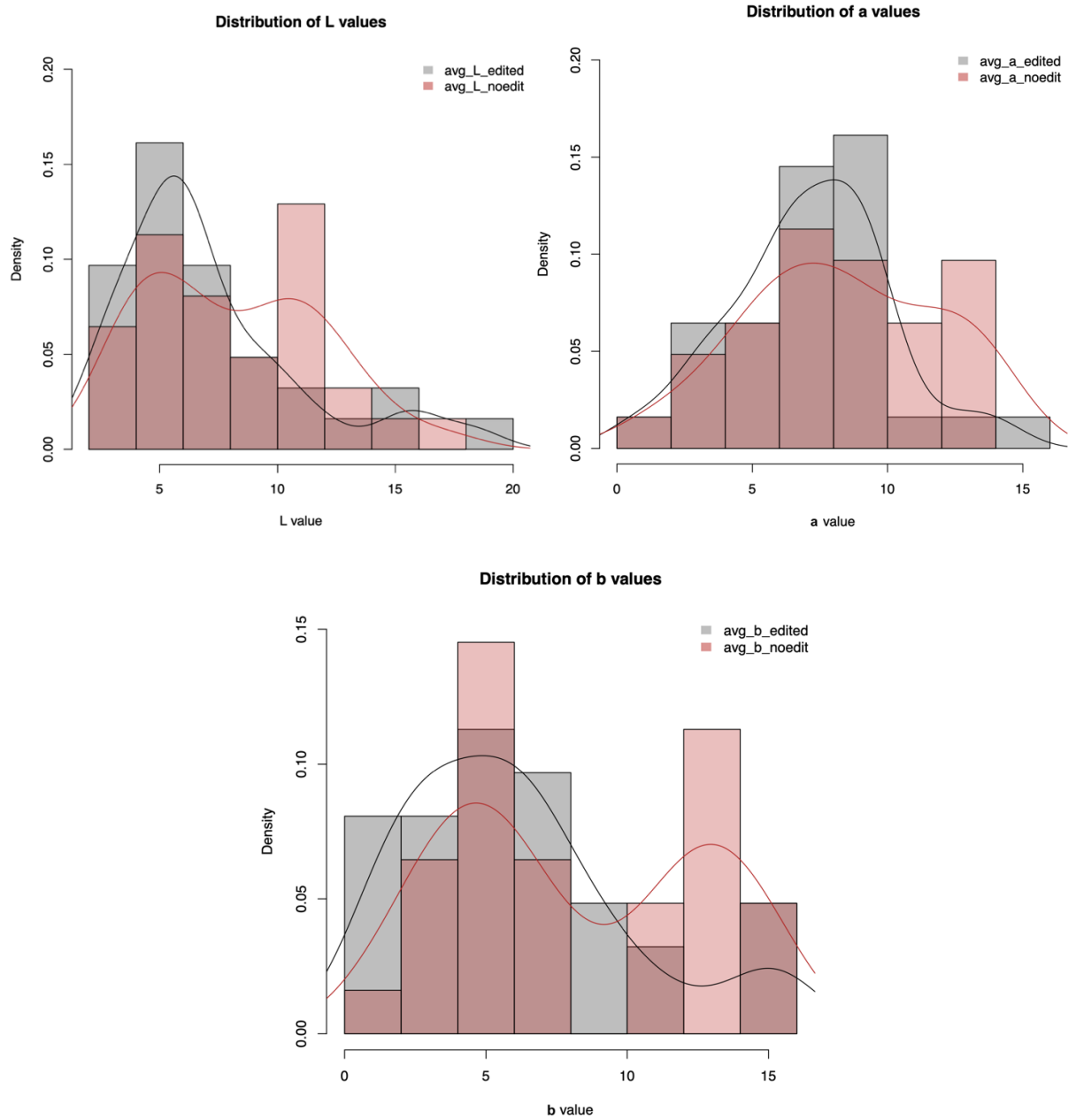
Section 2: Supplemental Information for Chapter 3



Supplemental Figure 3.1) A schematic of the CIE Lab color space. Lightness is represented by L^* ($L = 0$ yields black and $L = 100$ indicates white), a^* values describe the range between green ($-a^*$ values) and magenta ($+a^*$ values), and b^* describes the range between blue ($-b^*$ values) and Yellow ($+b^*$ values). Image by (author, add to reference list)

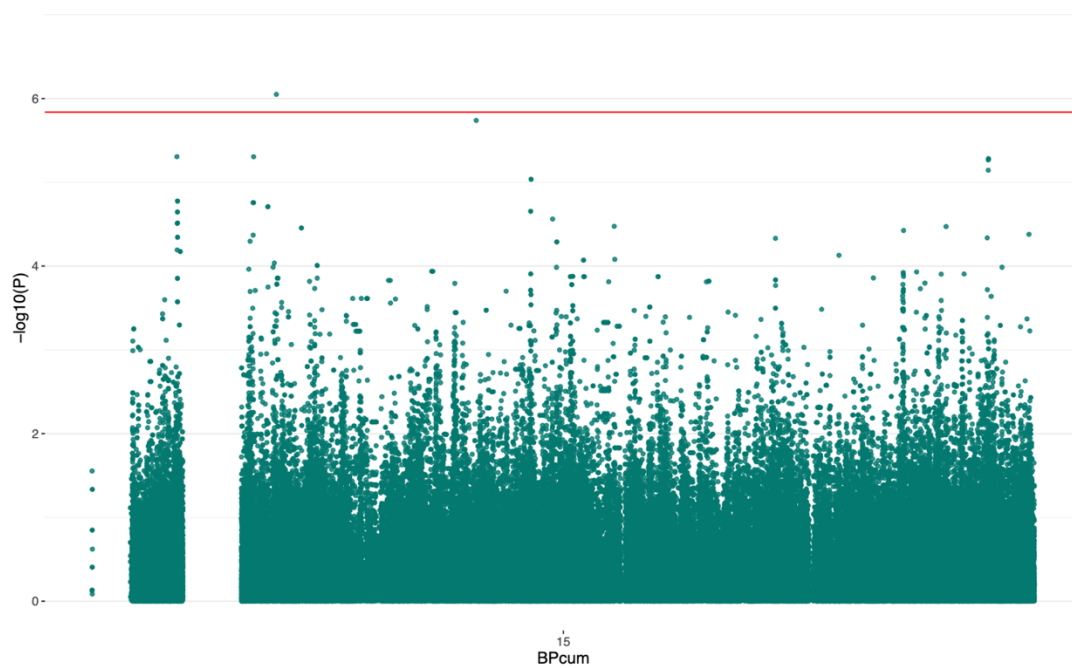


Supplemental Figure 3.2) Distribution of L* values among African Americans (red), European American (green), Hispanic (blue), and post-edited Khoe-San images (purple). Results from non-Khoe-San populations are unpublished and provided by Heather Norton for comparison.

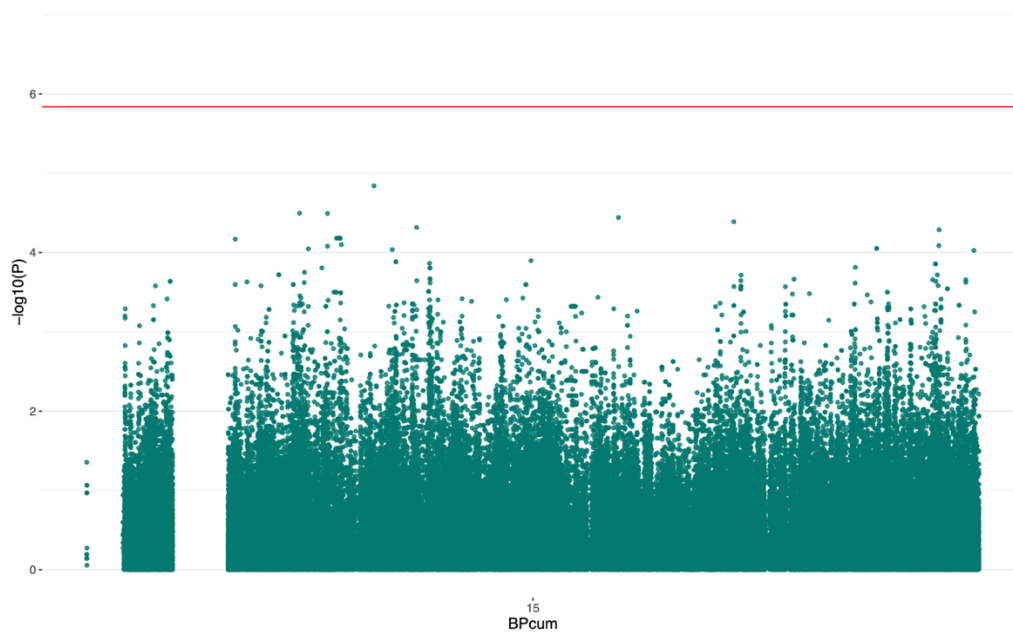


Supplemental Figure 3.3) Distribution of L^* , a^* , and b^* values before and after editing.

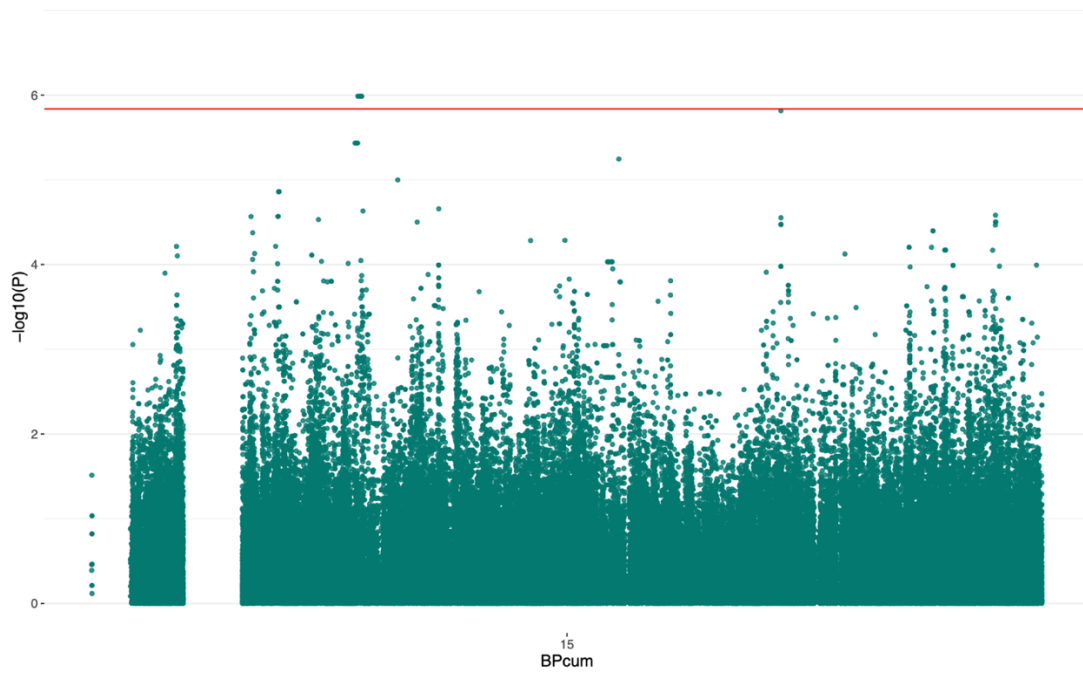
chr15.imputed.iris_data_allsamples_no_edits.mypcair.mypcrelateMat.5covs.L.gwasstats lambda= 1.02517



chr15.imputed.iris_data_allsamples_no_edits.mypcair.mypcrelateMat.5covs.a_avg.gwasstats lambda= 1.03158

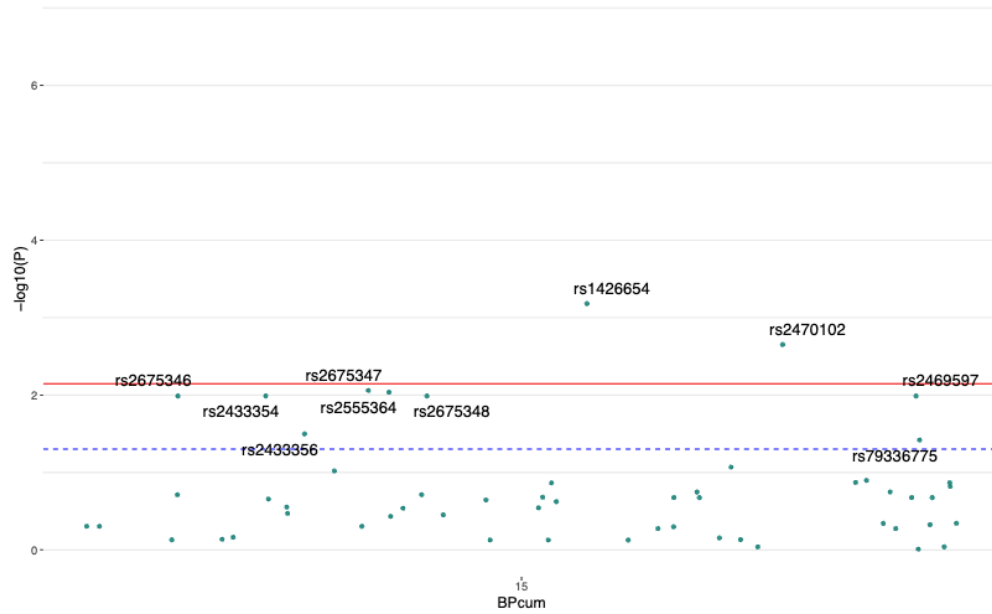


chr15.imputed.iris_data_allsamples_no_edits.mypcair.mypcrelateMat.5covs.b_avg.gwasstats lambda= 1.06391

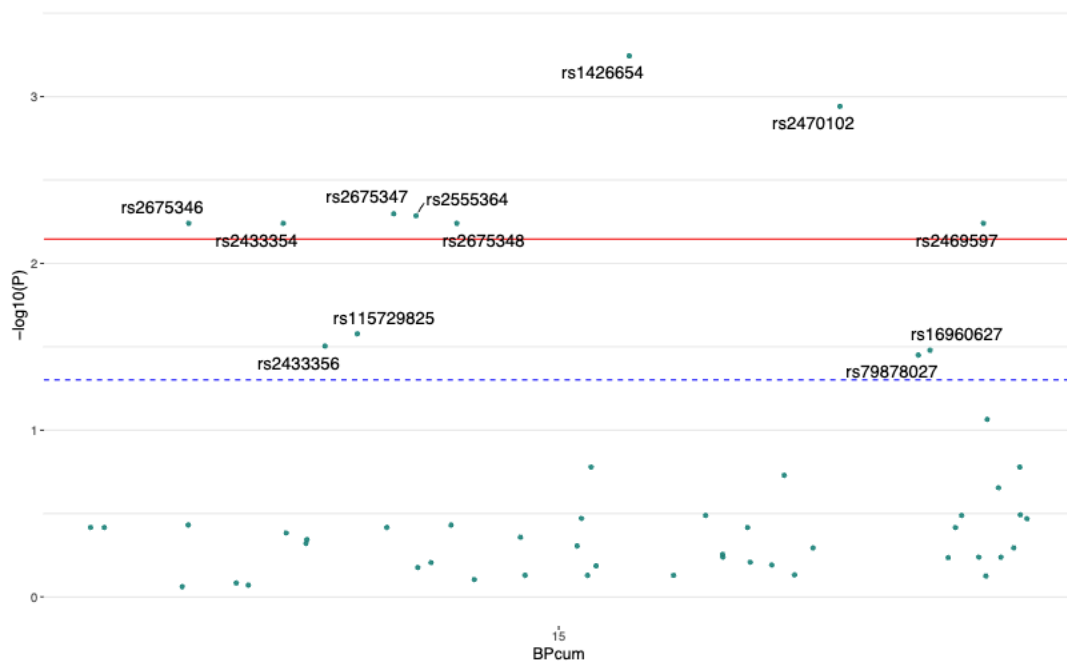


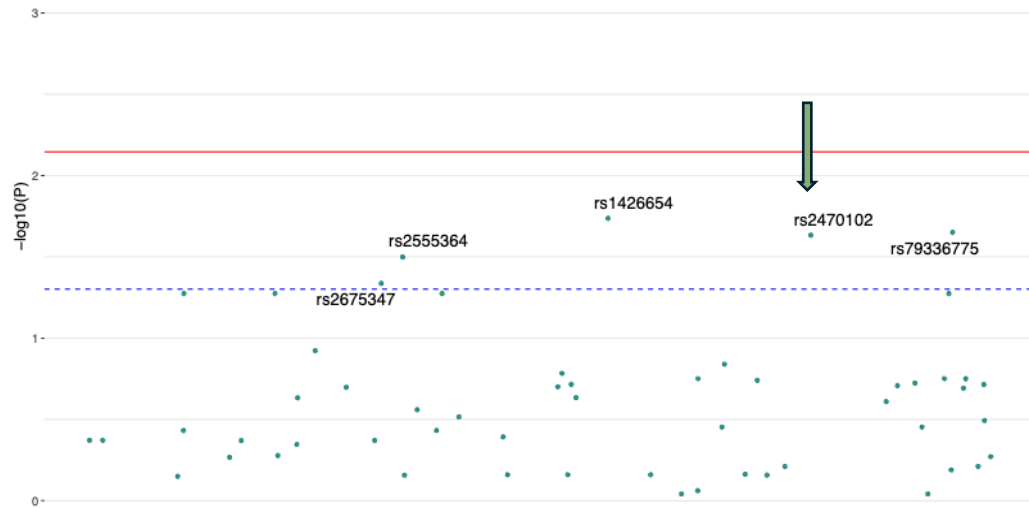
Supplemental Figure 3.4) The best fit model for each pigmentation value (L^* , a^* , and b , respectively from top to bottom) across chromosome 15.

SLC24A5.imputed.iris_data_allsamples_no_edits.mypcair.mypcrelateMat.5covs.a_avg.gwasstats



SLC24A5.imputed.iris_data_allsamples_no_edits.mypcair.mypcrelateMat.5covs.b_avg.gwasstats





Supplemental figure 3.5) Raw L* association results on all individuals. Red line represents the Bonferroni threshold. Blue line is the FDR (0.05) cut off. The green arrow highlights our SNP of interest, rs2470102.

Shapiro-Wilk Test Results			
Variable	p_value	Distribution	Significant
avg_L_edited	0.0010517	Not Normal	*
avg_a_edited	0.9296356	Normal	
avg_b_edited	0.0205202	Not Normal	*
avg_L_noedit	0.1351280	Normal	
avg_a_noedit	0.4639860	Normal	
avg_b_noedit	0.0043848	Not Normal	*

Supplemental Table 3.1) Shapiro-Wilk test was used to test for normalcy between un-edited and edited images.

Section 3: Supplemental Information for Chapter 4

Skin pigmentation, ancestry, and SLC24A5

Population	mean_baseM	n
Cederberg	48.1	171
OrangeRiver	49.7	90
Nama	52.3	191
Karoo	55.9	98
KhomaniSan	57.8	255
!Xoo	67.8	178
Jul'hoansi	70.9	135

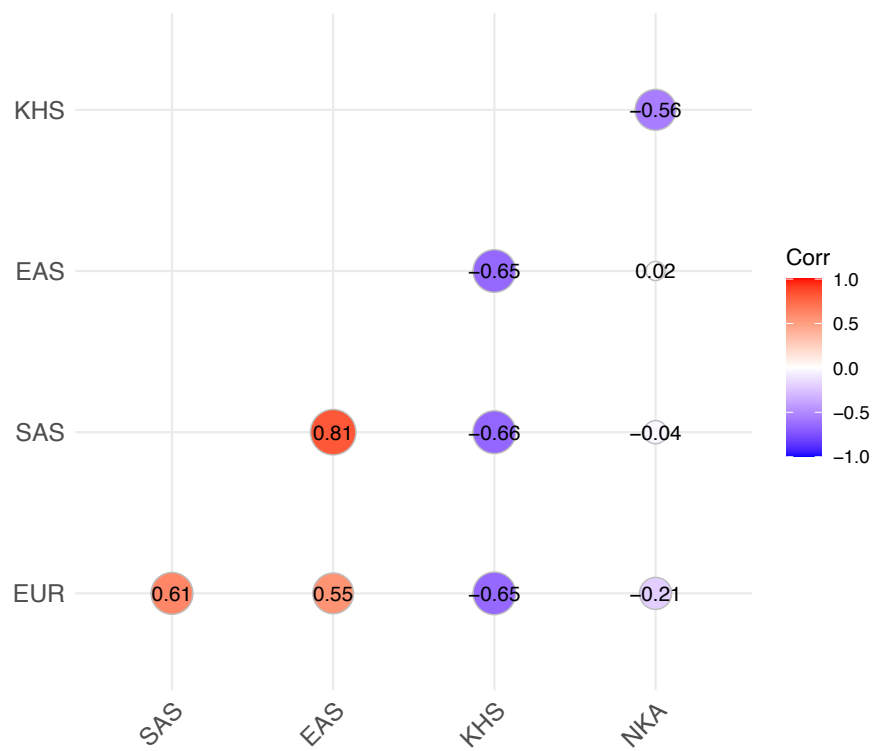
Supplemental Table 4.1) Population average melanin index values

Pop1	Pop2	Diff	95% CI	p-adj
!Xoo	Cederberg	-19.7	-22.4 to -17	6.22e-15
!Xoo	KhomaniSan	-10.0	-12.5 to -7.5	6.22e-15
!Xoo	Nama	-15.4	-18.1 to -12.8	6.22e-15
!Xoo	OrangeRiver	-18.1	-21.3 to -14.8	6.22e-15
Cederberg	Jul'hoansi	22.8	19.9 to 25.7	6.22e-15
Jul'hoansi	Karoo	-15.0	-18.4 to -11.7	6.22e-15
Jul'hoansi	KhomaniSan	-13.1	-15.8 to -10.4	6.22e-15
Jul'hoansi	Nama	-18.5	-21.4 to -15.7	6.22e-15
Jul'hoansi	OrangeRiver	-21.2	-24.6 to -17.7	6.22e-15
Cederberg	KhomaniSan	9.7	7.2 to 12.2	1.29e-14
!Xoo	Karoo	-11.9	-15.1 to -8.7	4.40e-14
KhomaniSan	OrangeRiver	-8.1	-11.2 to -5	6.98e-13
Cederberg	Karoo	7.8	4.6 to 11	2.99e-11
KhomaniSan	Nama	-5.5	-7.9 to -3	8.61e-10
Karoo	OrangeRiver	-6.1	-9.8 to -2.5	2.09e-05
Cederberg	Nama	4.2	1.6 to 6.9	5.76e-05
Karoo	Nama	-3.5	-6.7 to -0.4	1.64e-02
!Xoo	Jul'hoansi	3.1	0.2 to 6	2.56e-02
Nama	OrangeRiver	-2.6	-5.9 to 0.6	2.03e-01
Karoo	KhomaniSan	1.9	-1.1 to 4.9	4.80e-01
Cederberg	OrangeRiver	1.6	-1.7 to 4.9	7.70e-01

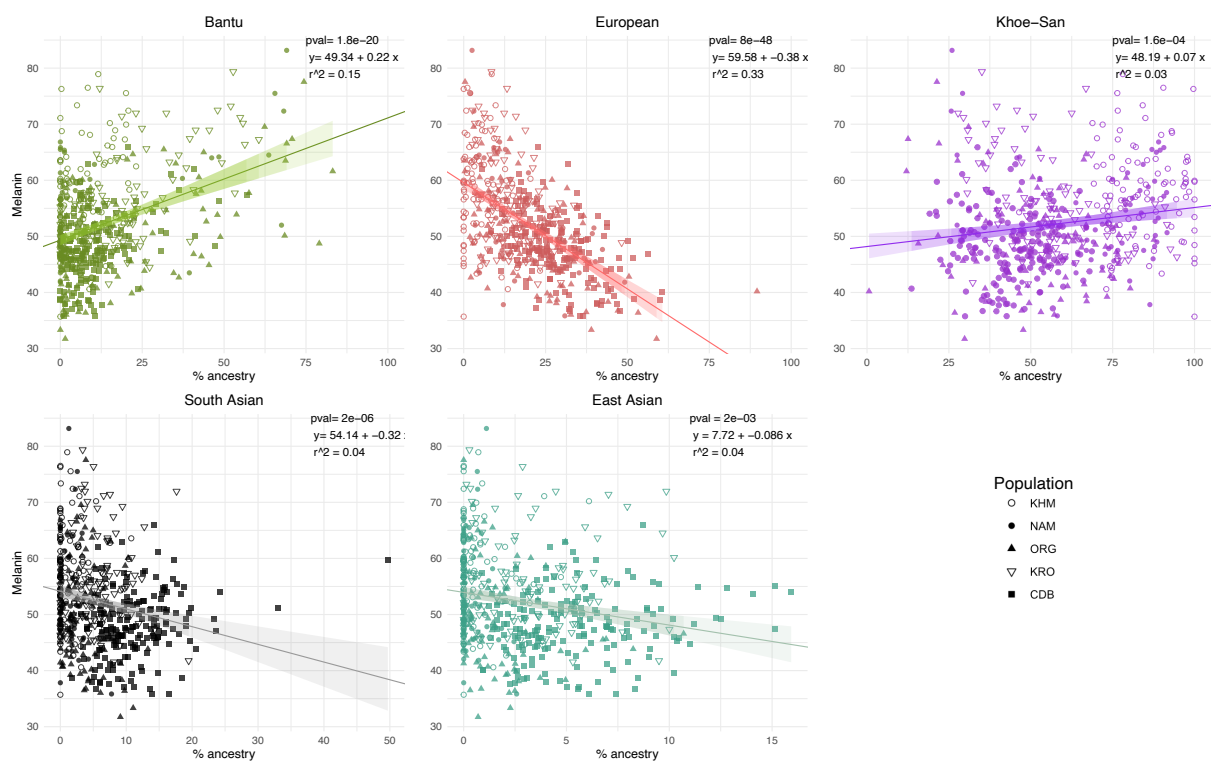
Supplemental Table 4.2) Post-hoc pairwise comparisons of average melanin index.

Population	n	count	frequency
!Xoo	178	39	11.0
Jul'hoansi	135	37	13.7
Karoo	98	62	31.6
KhomaniSan	255	168	32.9
OrangeRiver	90	75	41.7
Cederberg	171	166	48.5
Nama	191	196	51.3

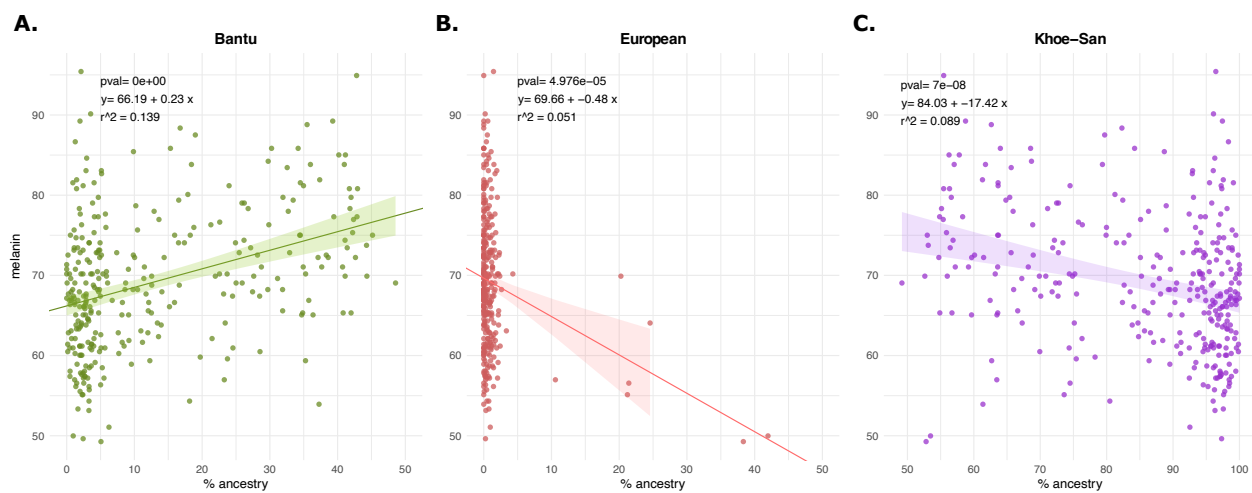
Supplemental Table 4.3) Frequency of SLC24A5*A lightening allele.



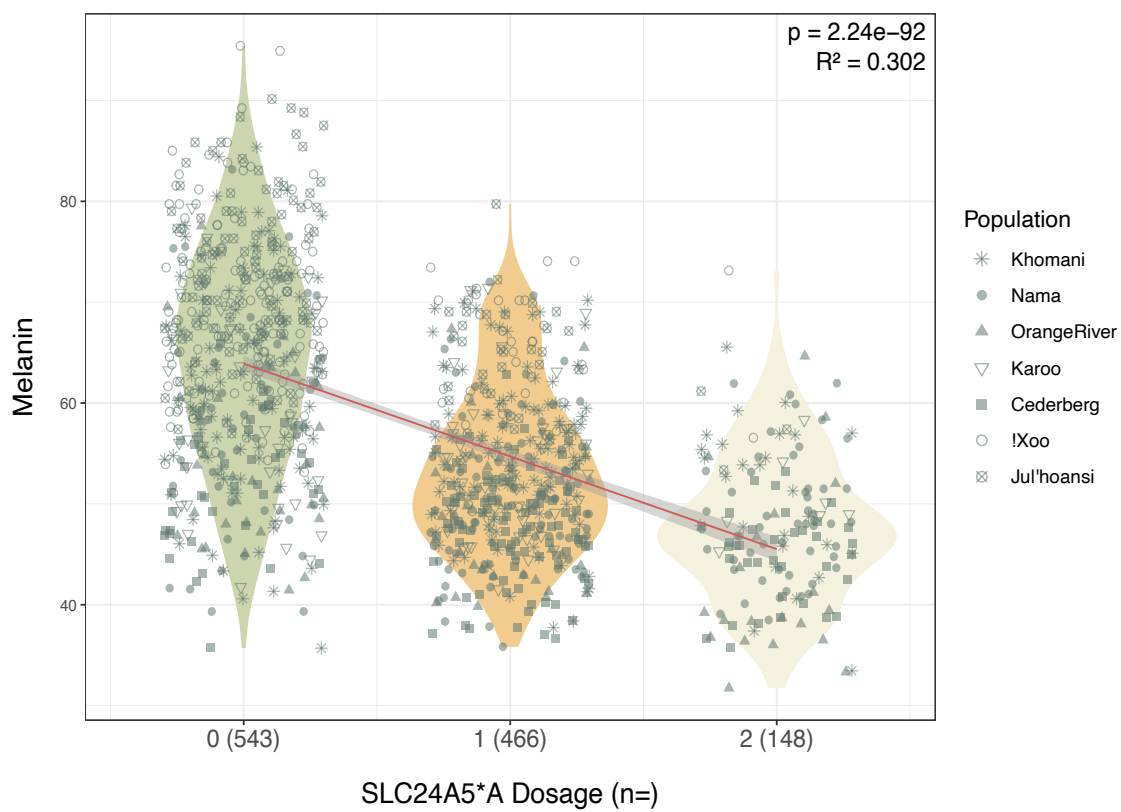
Supplemental Figure 4.1) Correlation of ancestry components.



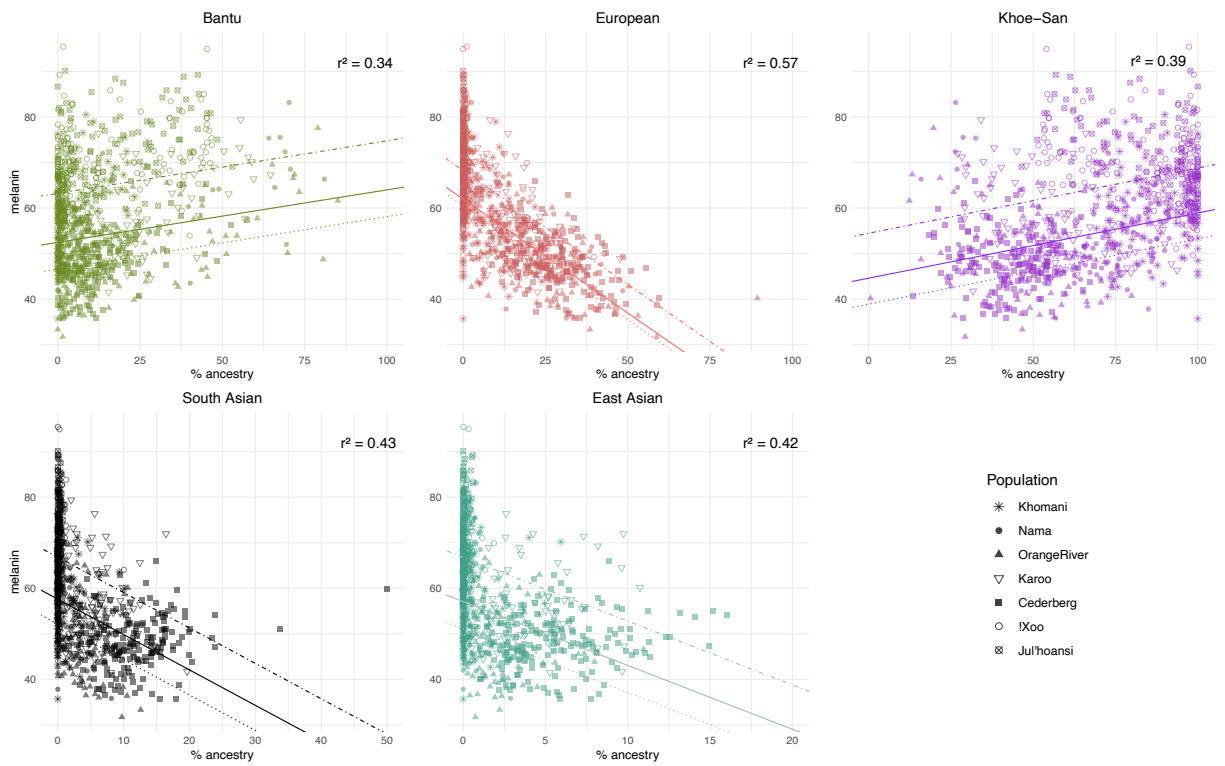
Supplemental Figure 4.2) Modeling ancestry against melanin excluding the Botswanan San.



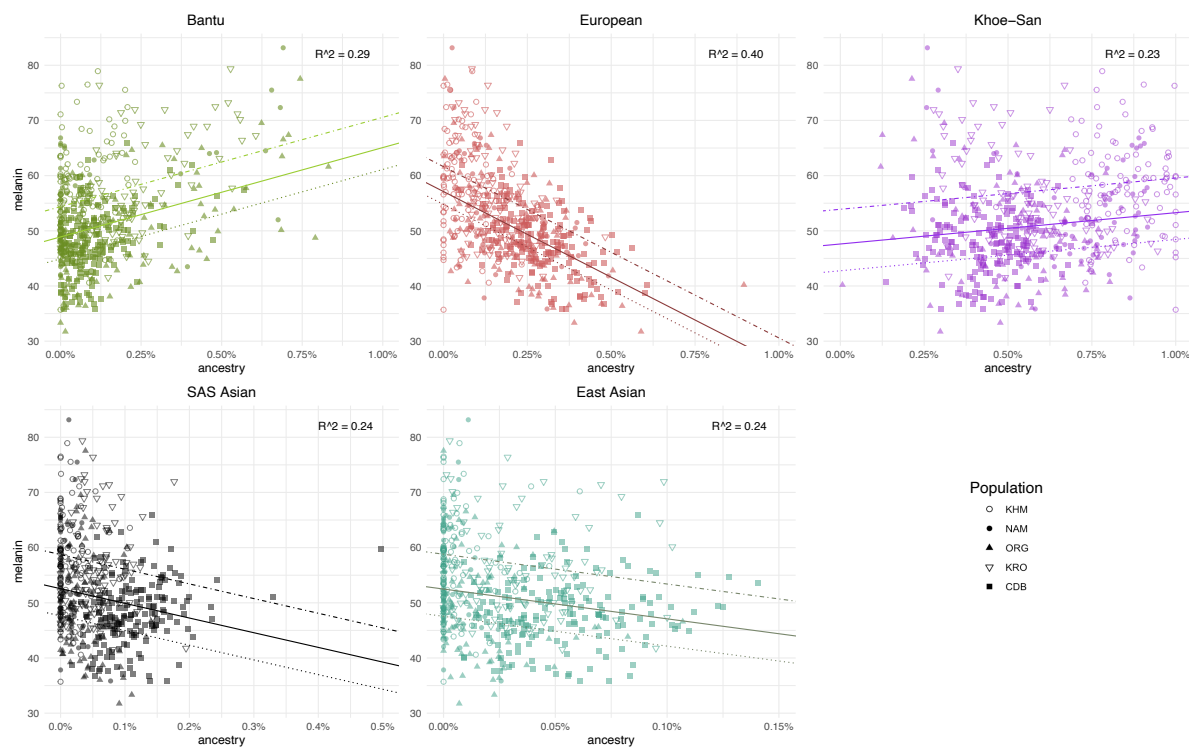
Supplemental Figure 4.3) Modeling ancestry against melanin within the Botswanan San.



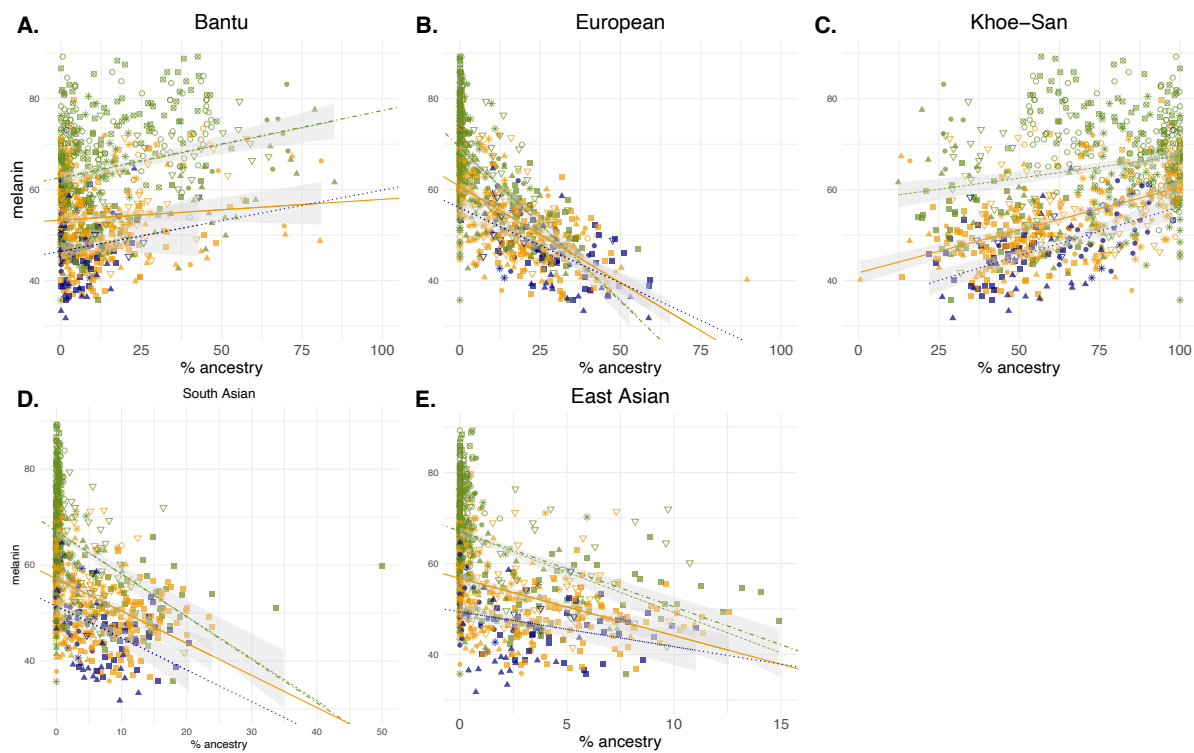
Supplemental Figure 4.4) Modeling melanin by rs1426654* A dosage.



Supplemental Figure 4.5) rs1426654*A as a fixed effect in ancestry by melanin models.



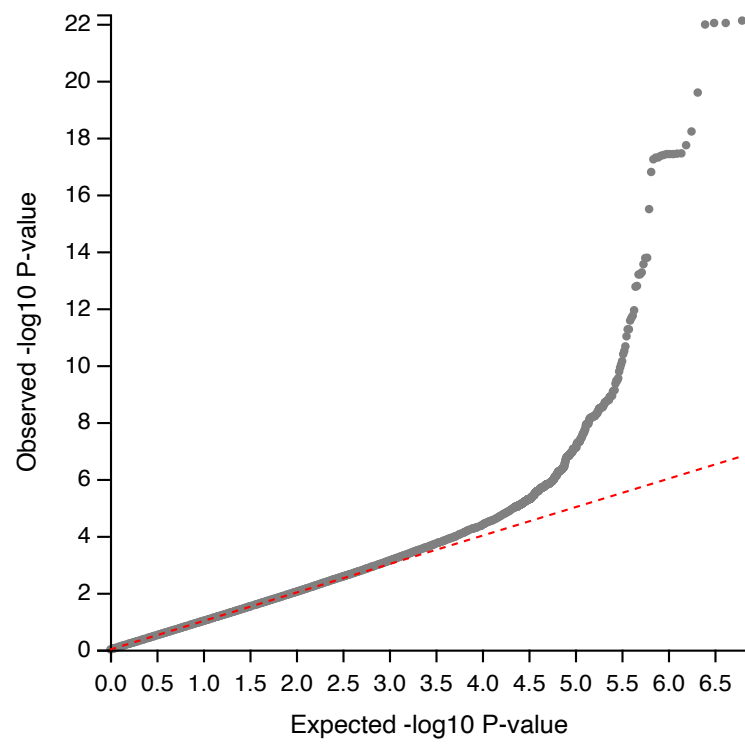
Supplemental Figure 4.6) rs1426654*A as a fixed effect excluding the Botswanan San.



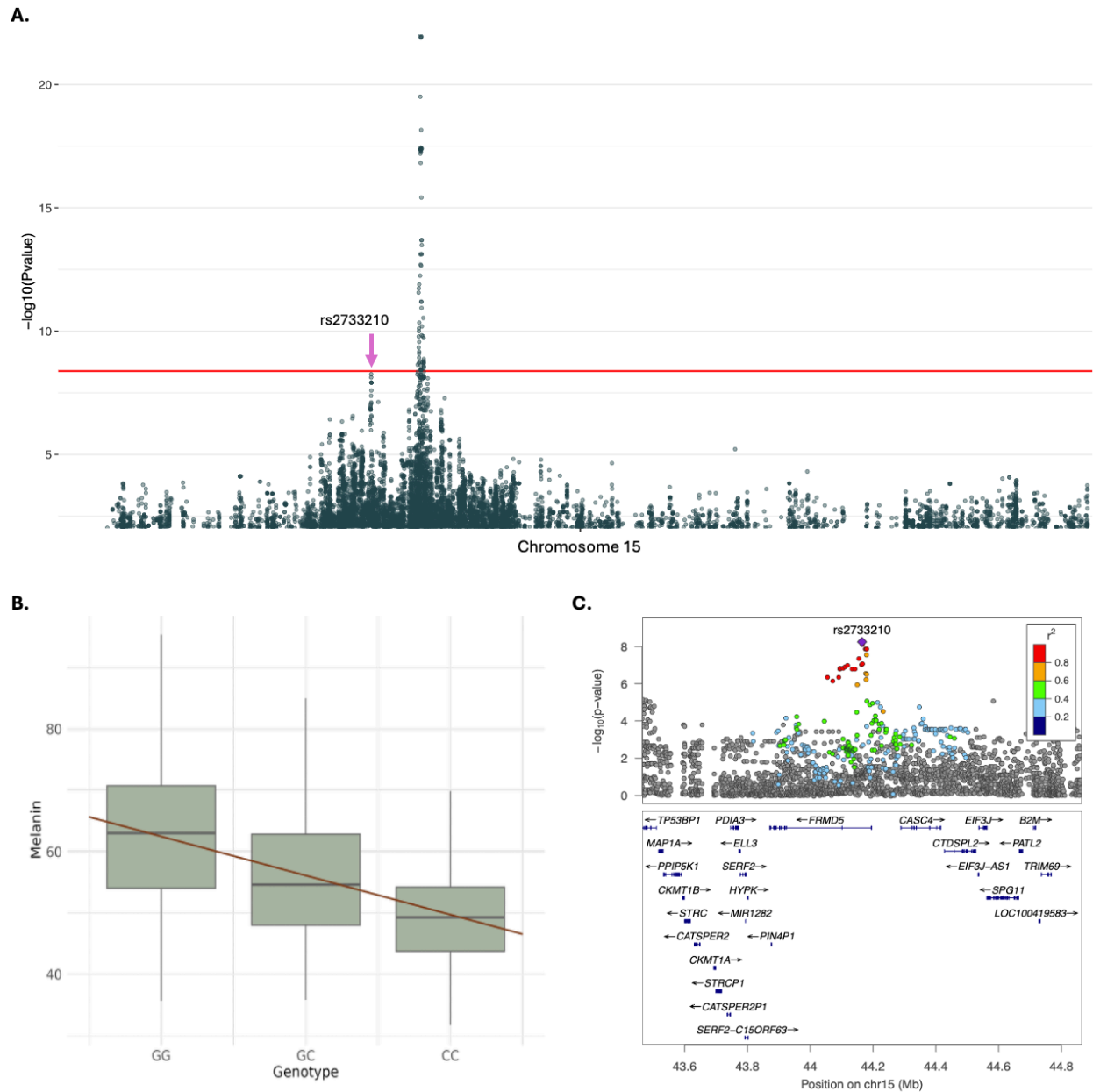
Supplemental Figure 4.7) Modeling ancestry by melanin while sub-setting by rs1426654*A dosage.

Ancestry	rs1426654_dose	$y = m(x) + b$	p-value	r ²
EAS	0	$y = -166.61x + 67.11$	1.15e-18	0.16
	1	$y = -126.03x + 56.74$	2.65e-14	0.16
	2	$y = -76.75x + 49.38$	4.43e-03	0.07
EUR	0	$y = -67.78x + 69.31$	2.11e-48	0.38
	1	$y = -42.34x + 60.70$	5.19e-42	0.43
	2	$y = -33.01x + 55.98$	4.80e-12	0.38
KHS	0	$y = 10.03x + 57.65$	6.22e-06	0.04
	1	$y = 18.74x + 41.74$	2.26e-21	0.24
	2	$y = 21.60x + 34.65$	2.20e-09	0.30
NKA	0	$y = 14.70x + 62.70$	2.22e-07	0.06
	1	$y = 4.58x + 53.29$	1.76e-01	0.00
	2	$y = 13.44x + 46.46$	1.51e-01	0.01
SAS	0	$y = -89.50x + 67.08$	2.18e-19	0.16
	1	$y = -67.05x + 57.09$	2.76e-15	0.17
	2	$y = -66.40x + 51.41$	3.06e-06	0.19

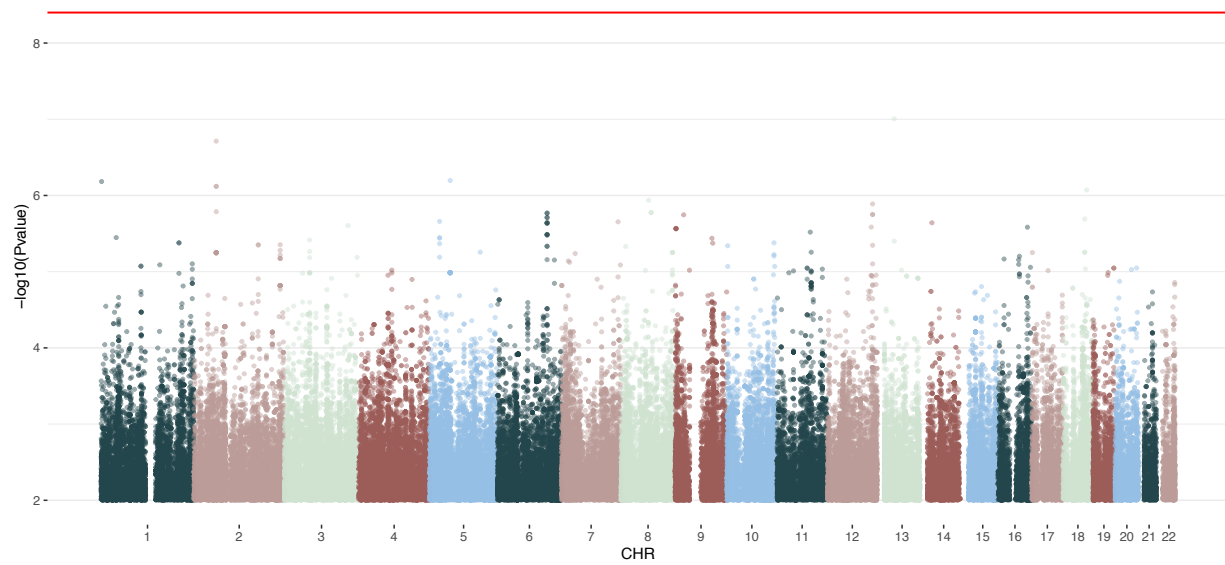
Supplemental Table 4.4) Ancestry by melanin results for models where individuals were subset based on their rs1426654*A dosage.



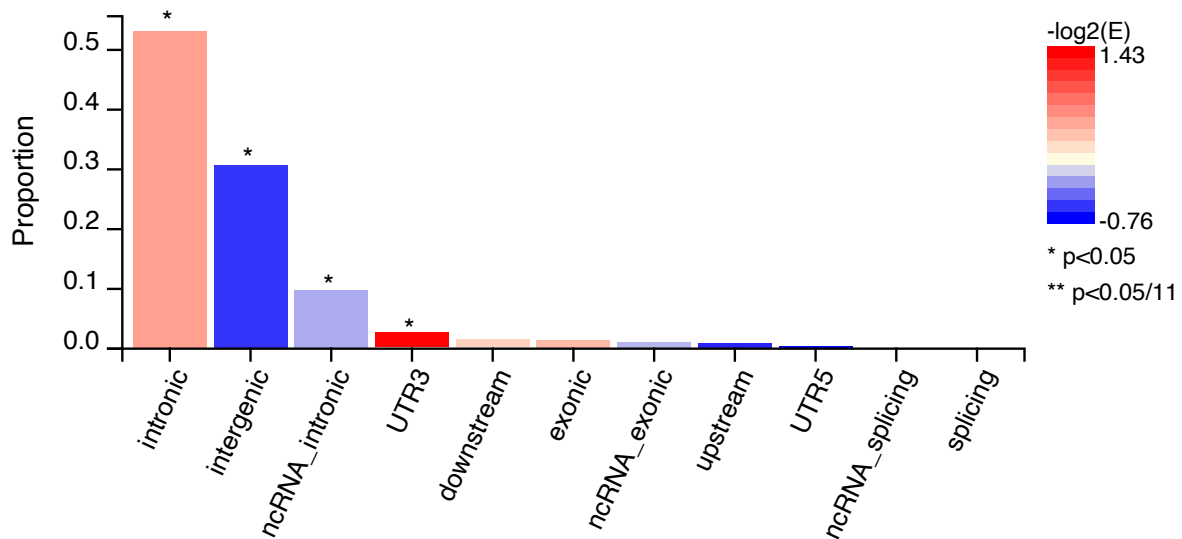
Supplemental Figure 4.8) Standard GWAS QQ-plot.



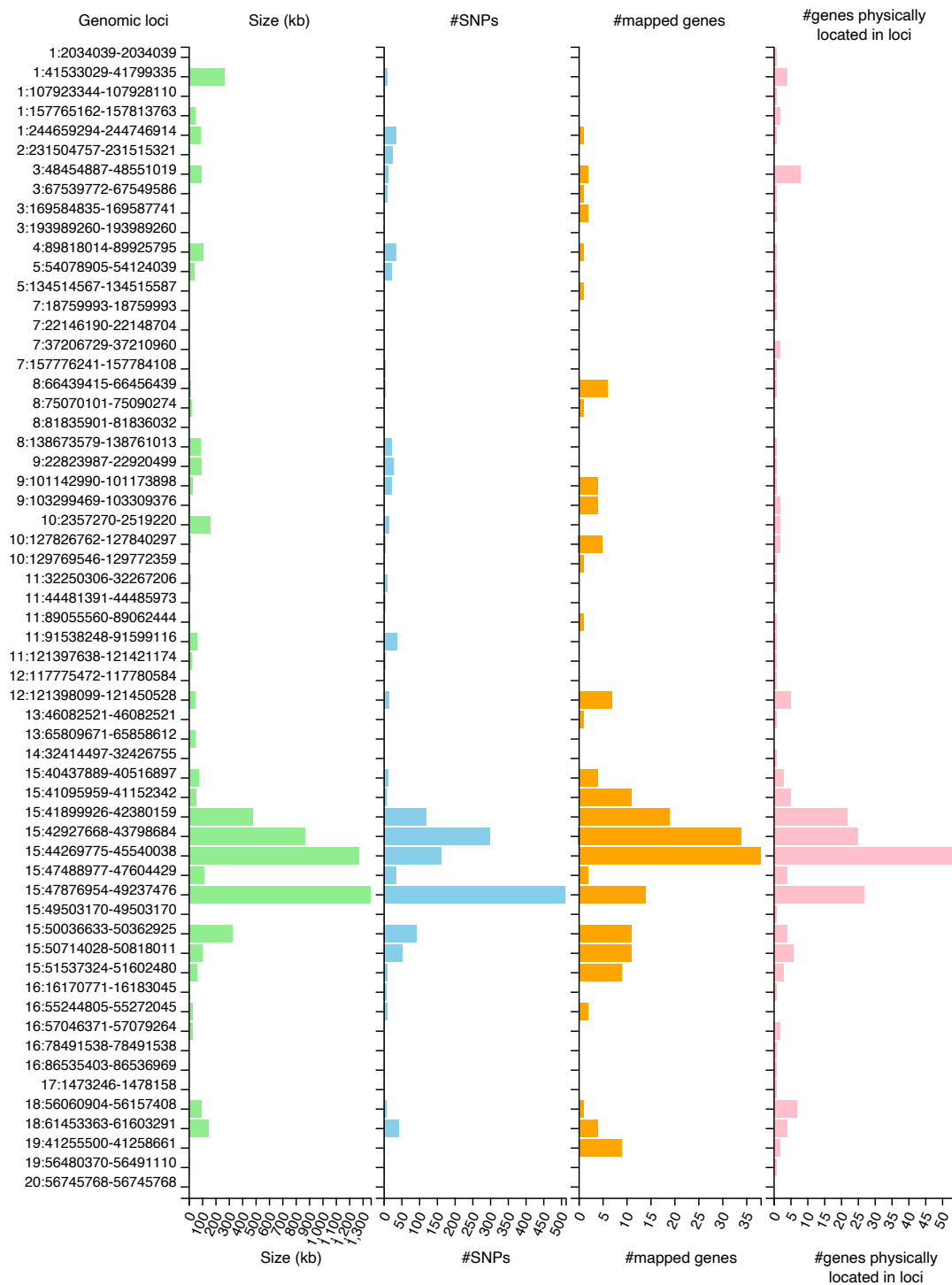
Supplemental Figure 4.9) *FRMD5*. A) A Manhattan plot highlighting the secondary peak in our standard GWAS, rs2733210 (*FRMD5*), upstream of *SLC24A5*. B) Box plot of melanin index values against rs2733210 genotype information. Beta value for this locus is -2.21. C) Locus zoom plot of rs2733210 (*FRMD5*).



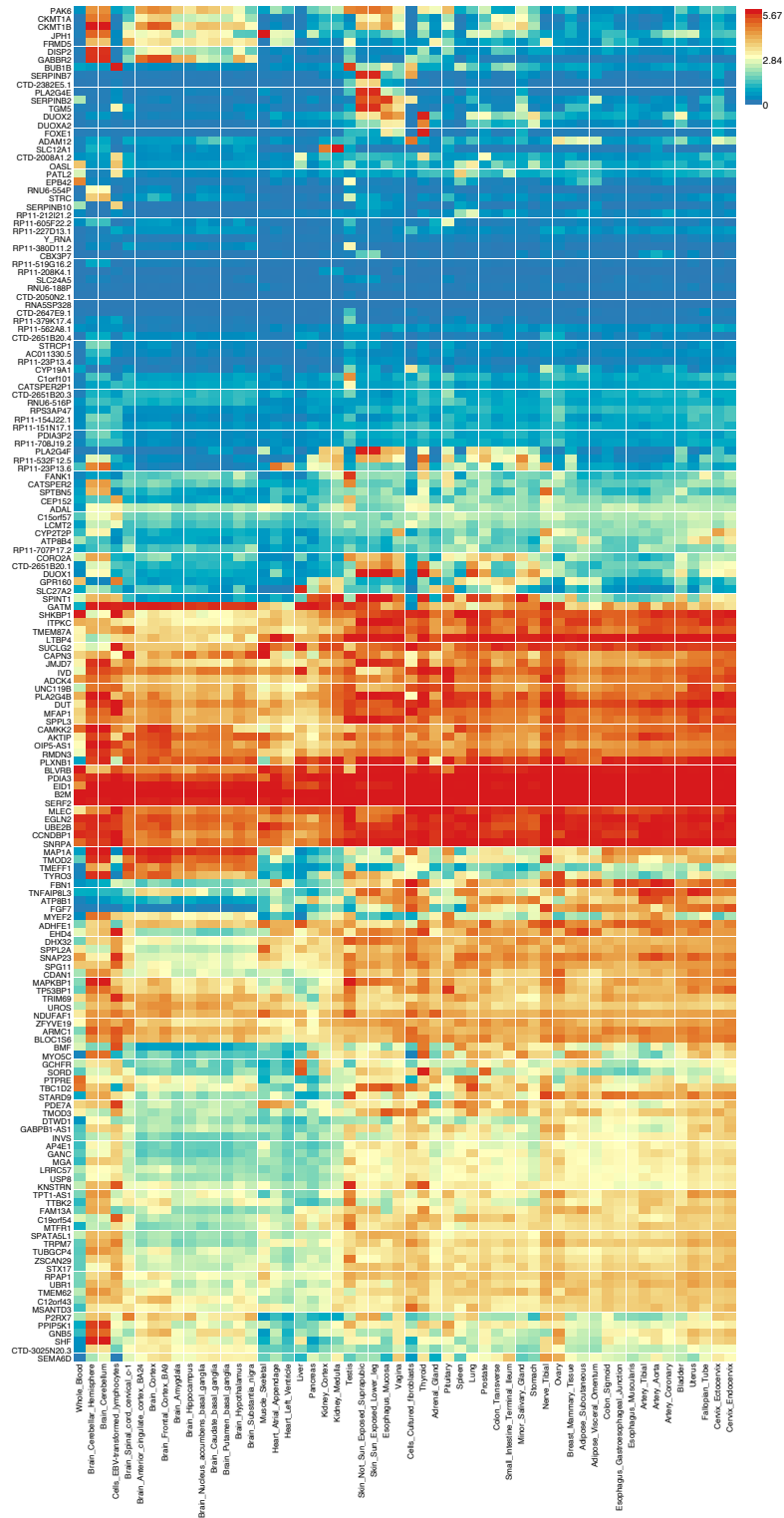
Supplemental Figure 4.10) Conditional GWAS. GWAS was conditioned on a single locus in *SLC24A5*, rs1426654.



Supplemental Figure 4.11) Number of SNPs per functional consequences on genes. Histogram displays the proportion of SNPs, including SNPs in LD of significant SNPs, which have corresponding functional annotations in ANNOVAR. Bars colors correspond their $\log_2(\text{enrichment})$ relative to all SNPs in the reference panel. When SNPs have more than one (different) annotations, they are counted for each annotation.



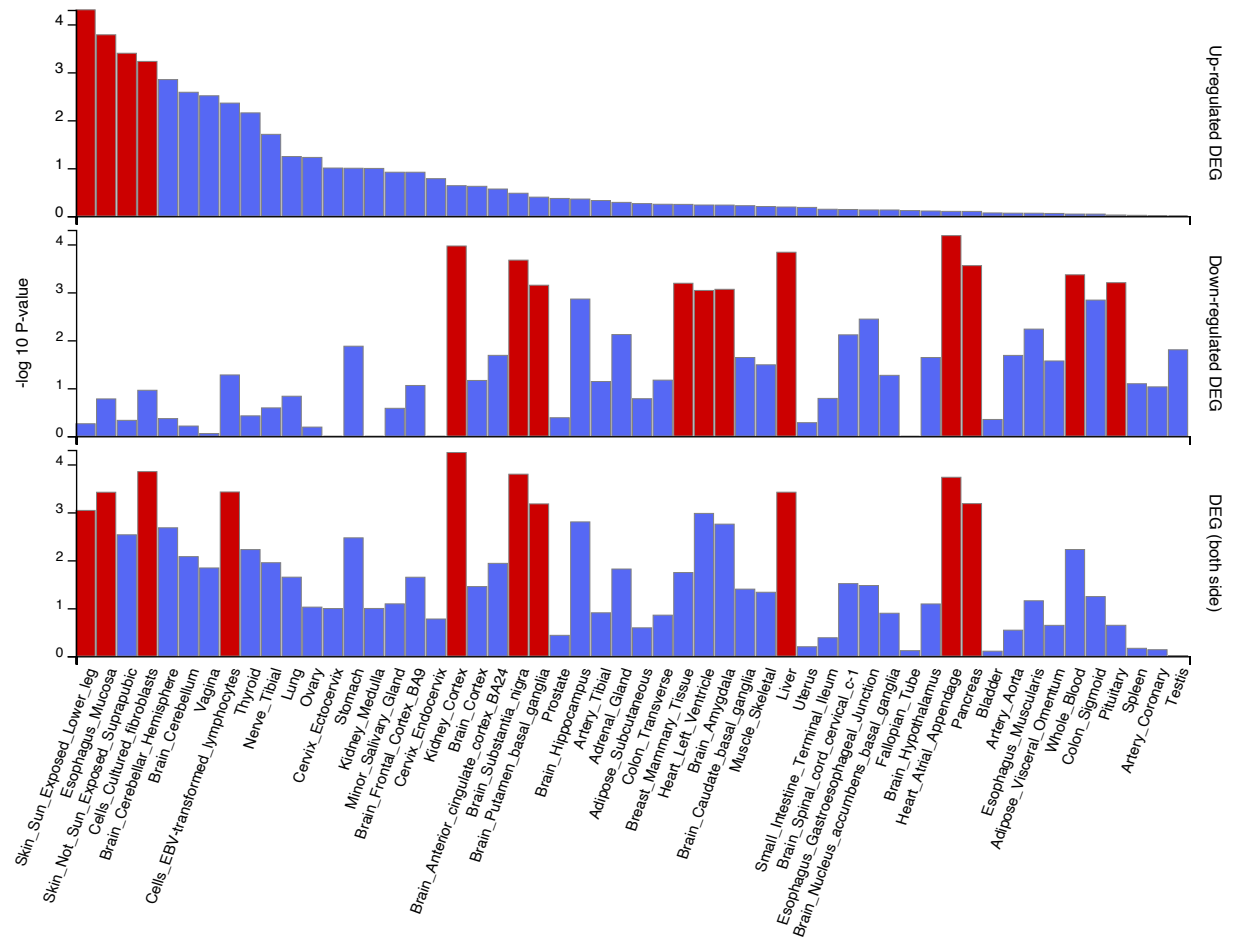
Supplemental Figure 4.12) 59 genomic windows with loci of interest. Chromosomes and positions of windows are specified on the y-axis. Title and x-axis labels begin with “#” indicate the count of the respective feature.



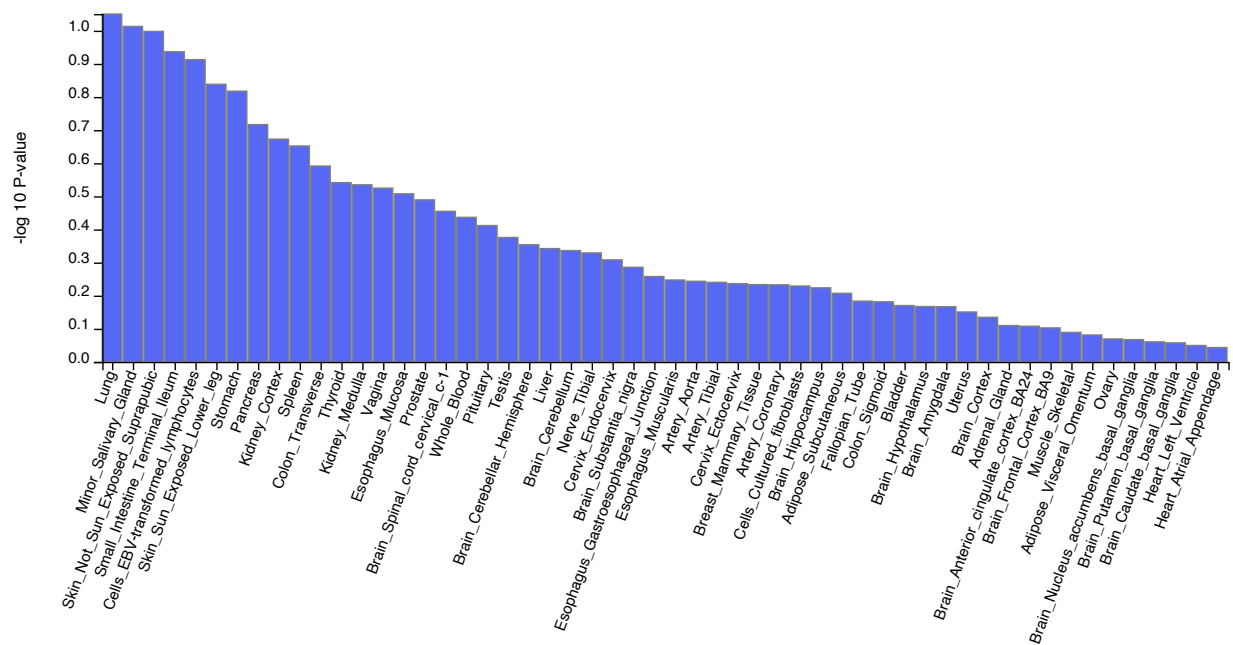
Supplemental Figure 4.13) Average expression across tissue (avg. RPKM log2 transformed) for each gene across GTEx 58 tissue types (see FUMA GENE2FUNC tutorial for details).

GeneSet	N	P-value	adjusted P	genes
chr15q15	50	5.45E-90	1.63E-87	BMF, BUB1B, PAK6, DISP2, KNSTRN, IVD, C15orf57, RMDN3, GCHFR, ZFYVE19, SPINT1, OIP5-AS1, NDUJAF1, RPAP1, TYRO3, MGA, MAPKBP1, PLA2G4B, JMJD7, SPTBN5, EHD4, CTD-2382E5.1, PLA2G4E, PLA2G4F, TMEM87A, GANC, CAPN3, SNAP23, LRRC57, STARD9, CDAN1, TTBK2, UBR1, EPB42, TMEM62, CCNDBP1, TGM5, LCMT2, ADAL, ZSCAN29, TUBGCP4, TP53BP1, MAP1A, PPIP5K1, CKMT1B, CKMT1A, CATSPER2P1, PDIA3, SERF2, MIR1282, MFAP1, FRMD5
chr15q21	37	1.03E-58	1.55E-56	SPG11, PATL2, B2M, TRIM69, SORD, DUOX2, DUOX2, DUOX1, SHF, CTD-2651B20.3, GATM, SPATA5L1, BLOC1S6, SEMA6D, SLC24A5, MYEF2, SLC12A1, DUT, FBN1, CEP152, EID1, FGF7, DTWD1, ATP8B4, SLC27A2, GABPB1-AS1, MIR4712, USP8, TRPM7, SPPL2A, AP4E1, TNFAIP8L3, CYP19A1, TMOD2, TMOD3, GNB5, MYO5C
chr8q13	4	7.13E-05	7.13E-03	ARMC1, MTFR1, PDE7A, ADHFE1
chr12q24	7	7.01E-04	4.29E-02	MLEC, UNC119B, SPPL3, C12orf43, OASL, P2RX7, CAMKK2
chr10q26	5	8.07E-04	4.29E-02	UROS, DHX32, FANK1, ADAM12, PTPRE
chr9q31	4	8.57E-04	4.29E-02	STX17, INVS, MSANTD3, TMEFF1

Supplemental Table 4.5) FUMA-derived gene-sets.

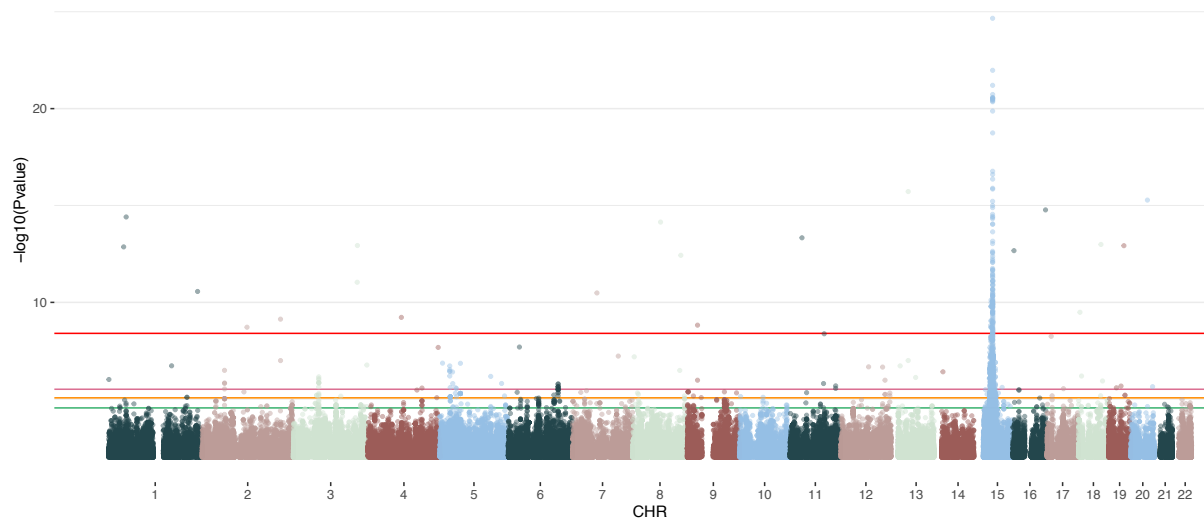


Supplemental Figure 4.14) Differentially expressed genes (DEG) across tissue types. Red bars highlight tissues with statistically significant enrichment (after multiple testing correction). Please see FUMA tutorial for more detail.

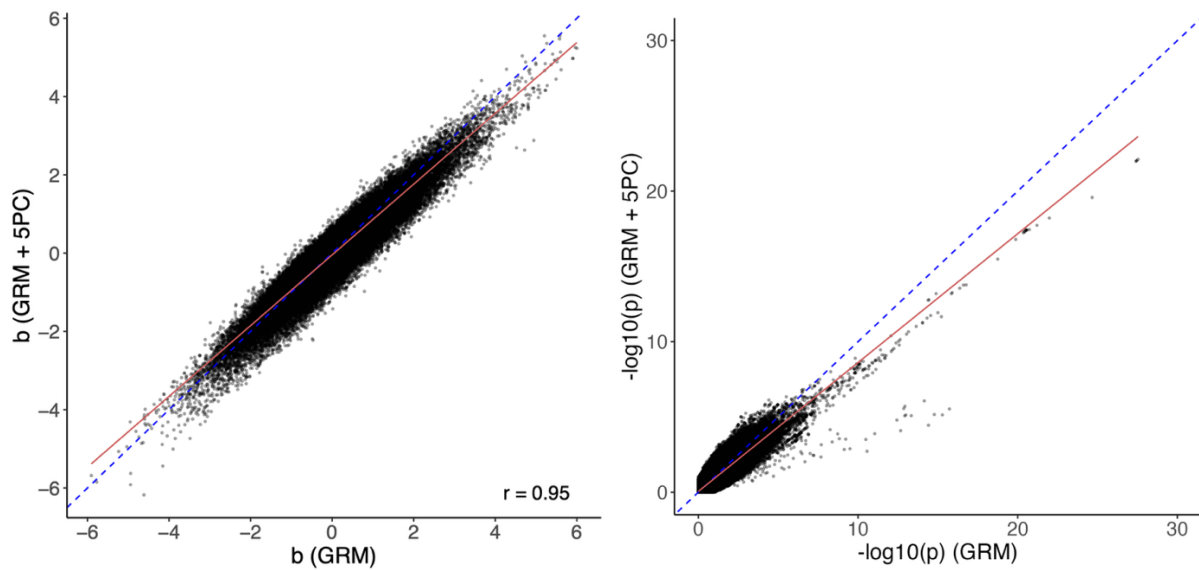


Supplemental Figure 4.15) Gene expression data from GTEx for 53 specific tissue types analyzed using FUMA's MAGMA gene-property analysis (see FUMA tutorial for details).

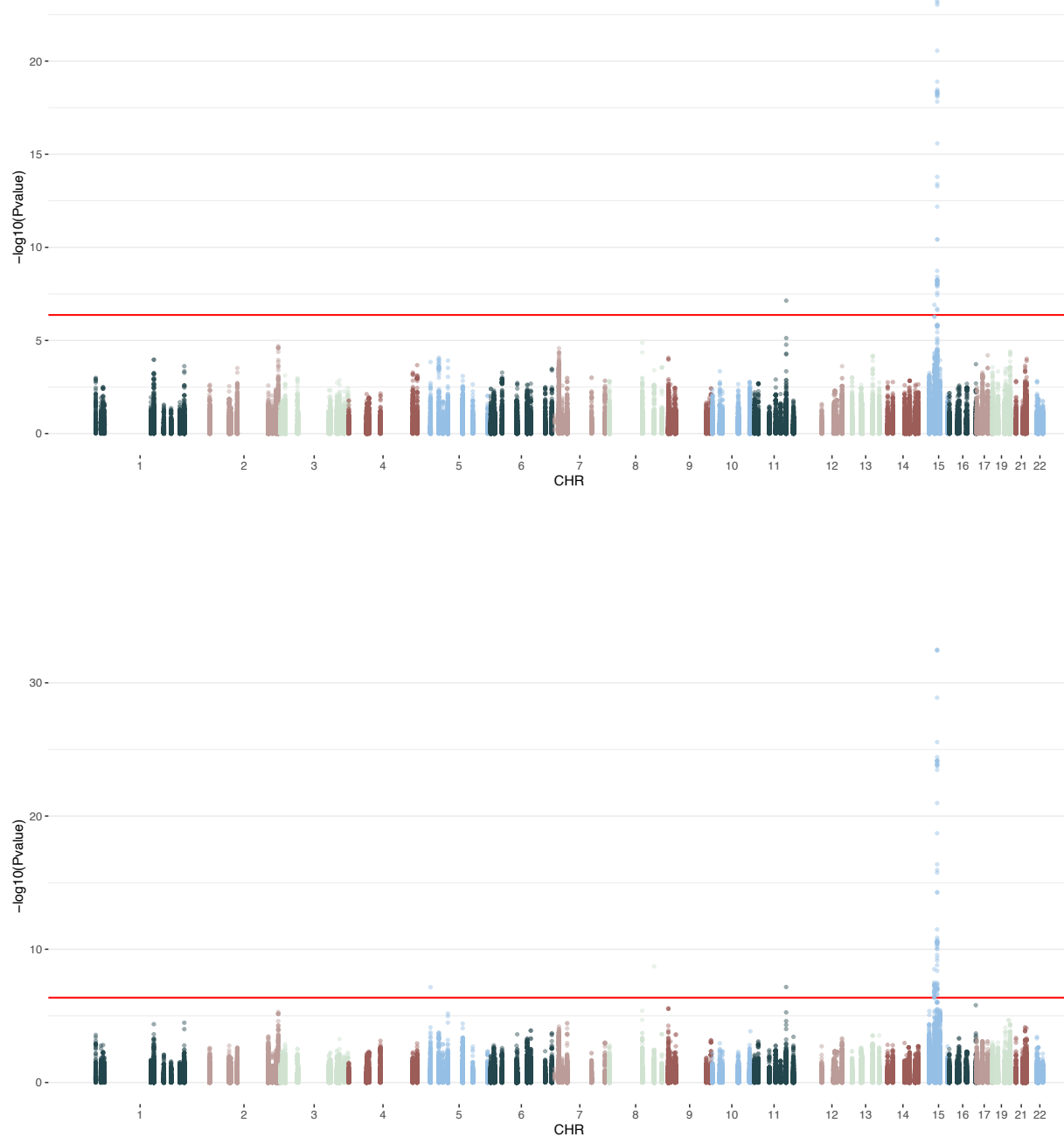
Constraints with including PCs in GWA



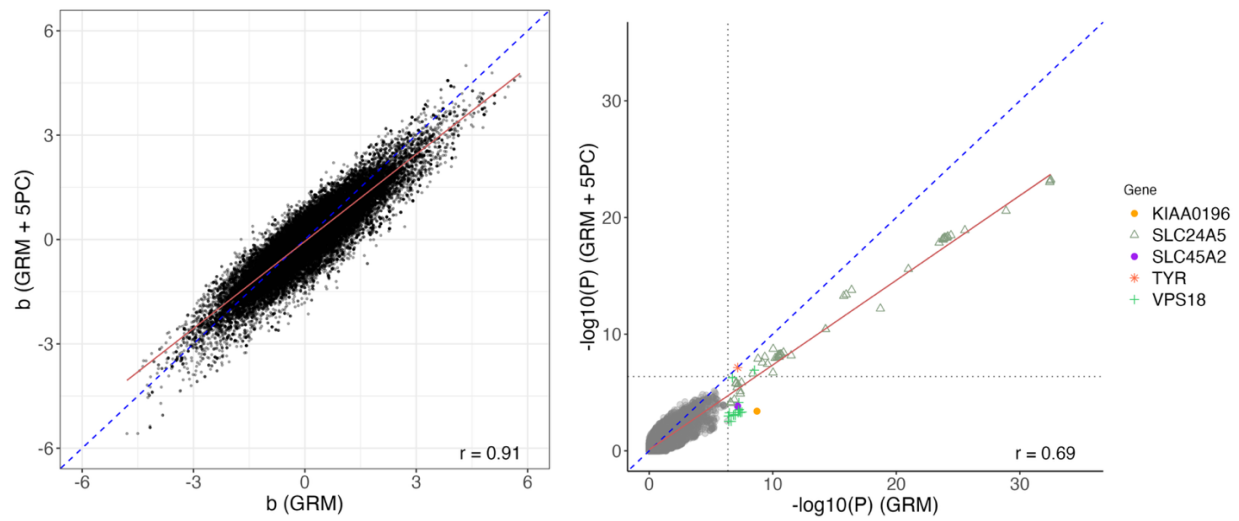
Supplemental Figure 4.16) Manhattan plot of GRM-only GWAS model. The red line indicates the Bonferroni-corrected significance threshold, while the pink, orange, and green lines represent false discovery rate (FDR) corrections at 0.05, 0.10, and 0.20, respectively.



Supplemental Figure 4.17) Beta and P-value comparison of GWAS models. Two models were evaluated, one including GRM and top 5 PCs as covariates (y-axes), and the only including a GRM (x-axes). Top figure visualizes correlation of beta values between models, and the bottom figure visualizes correlation of p-values. Red line indicates correlation; blue line is $x=y$. P-value correlation is 0.77.



Supplemental Figure 4.18) Manhattan plots of partitioned GWAS models. Data was attenuated to the 169 pigmentation loci identified in Bajpai et al. (2023). Two models were evaluated, one including GRM and top 5 PCs (top), and the only including a GRM (bottom).

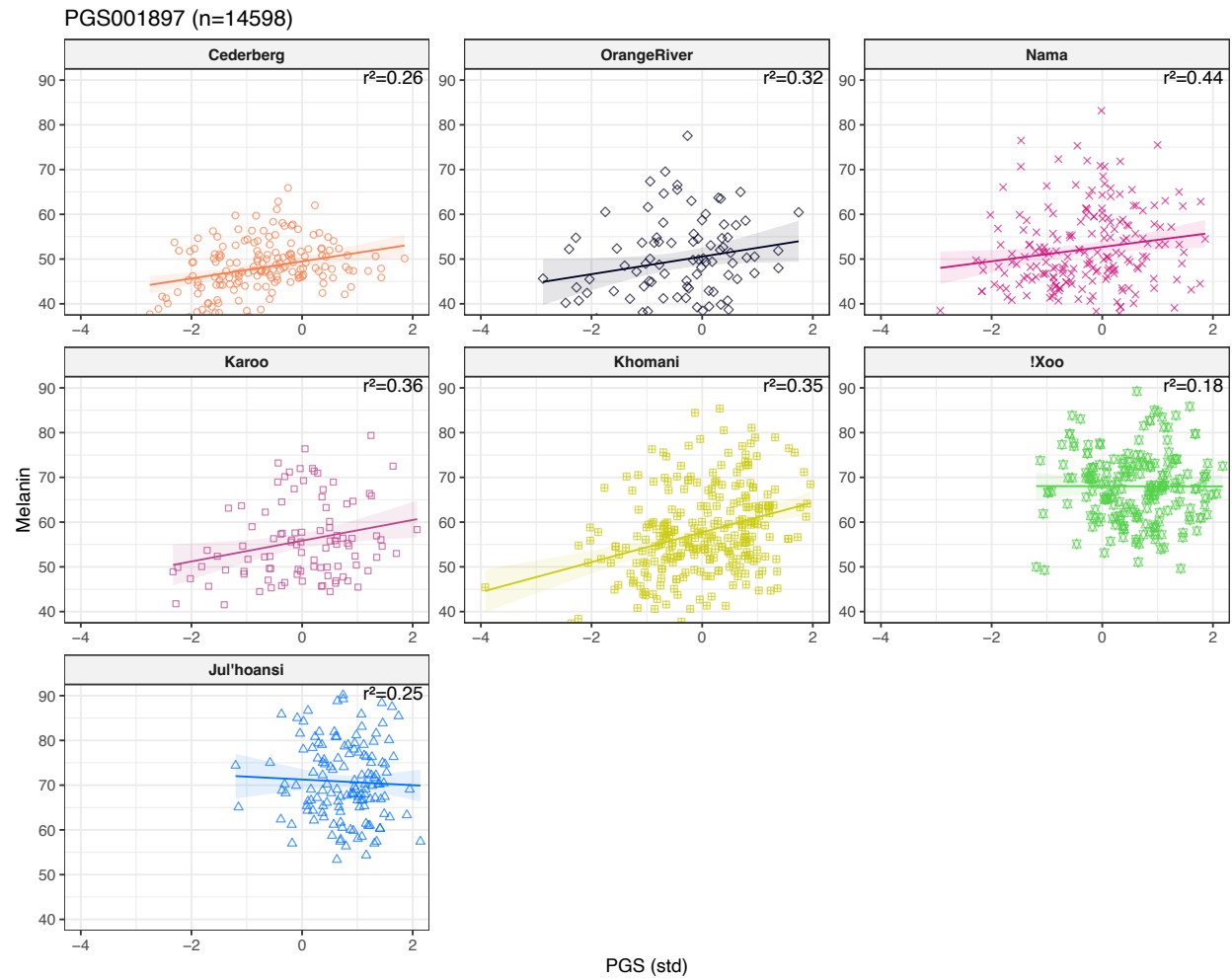


Supplemental Figure 4.19) Beta and P-value comparison of partitioned GWAS models. Data was attenuated to the 169 pigmentation loci identified in Bajpai et al. (2023). Two models were evaluated, one including GRM and top 5 PCs as covariates (y-axes), and the only including a GRM (x-axes). Top figure visualizes correlation of beta values between models, and the bottom figure visualizes correlation of p-values. Red line indicates correlation; blue line is $x=y$. Correlation values are at the bottom right of each figure.

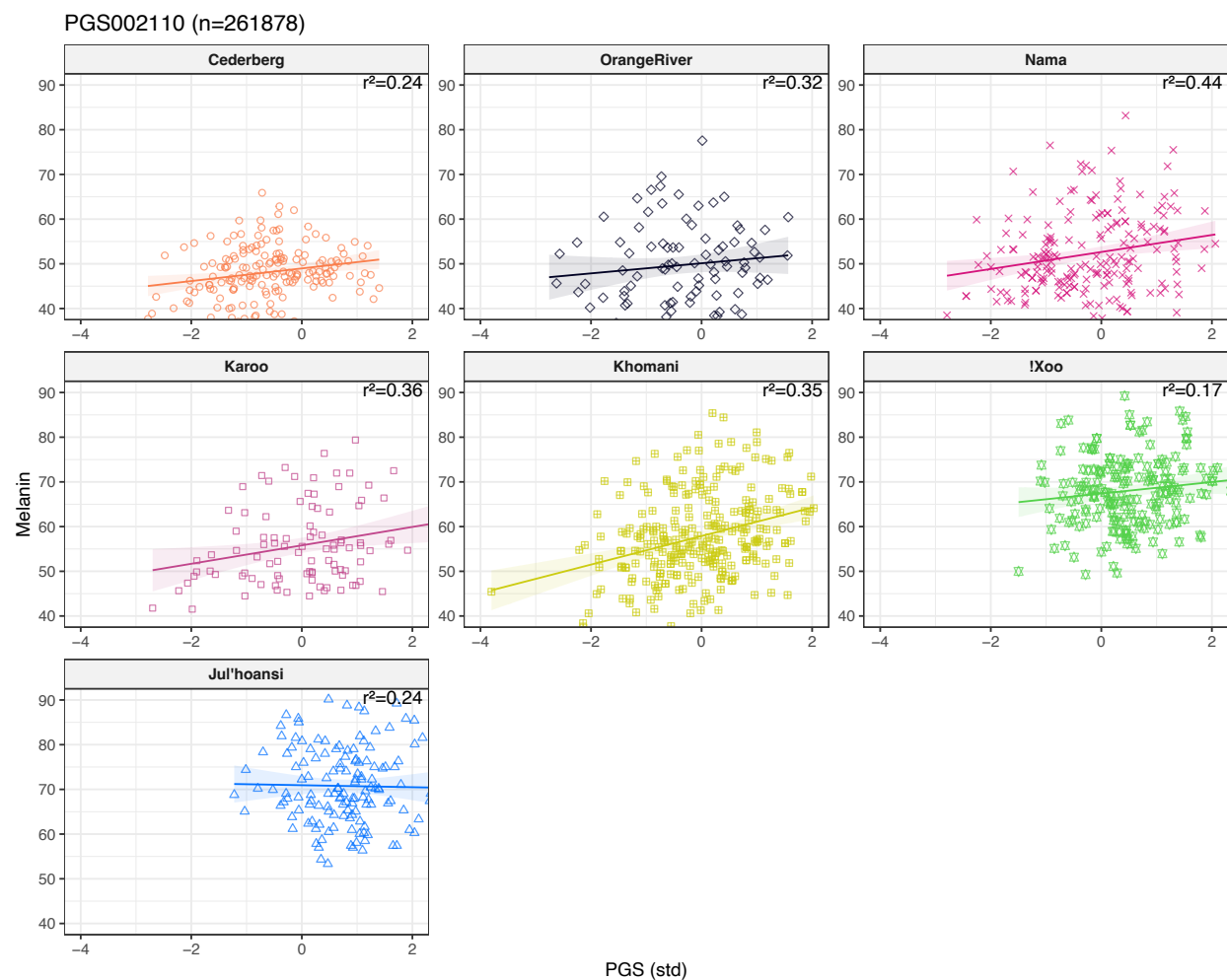
Poor Transferability of European-Derived Polygenic Scores

Panel	N	Model	R ²
PGS001897	14598	Base (PHENO ~ Score)	0.206
PGS001897	14598	PHENO ~ Score + 16PCs	0.578
PGS002110	261878	Base (PHENO ~ Score)	0.173
PGS002110	261878	PHENO ~ Score + 16PCs	0.578

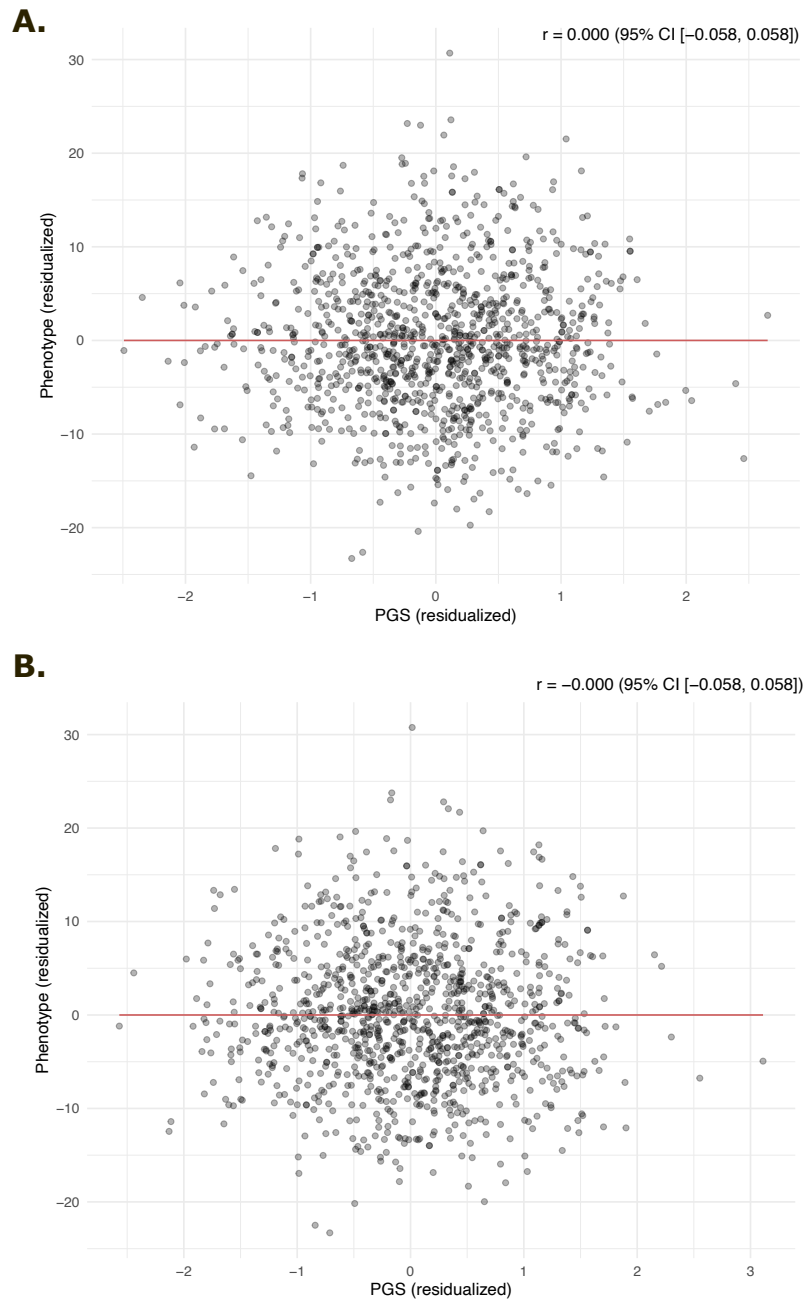
Supplemental Table 4.6) Correlation of PGS models and melanin.



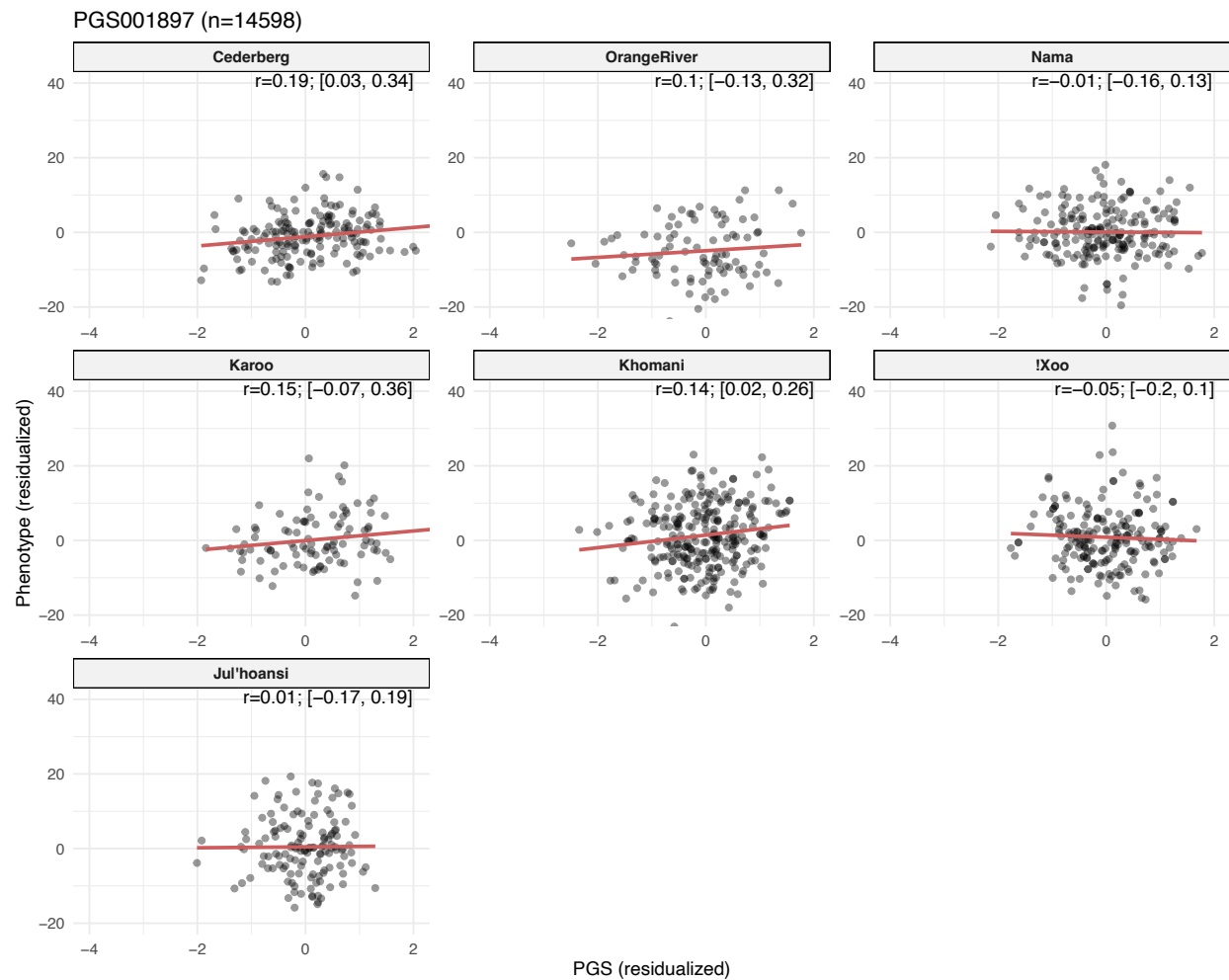
Supplemental Figure 4.20) Melanin against PGS001897 PGS loci. Included in the model are the top 16PCs.



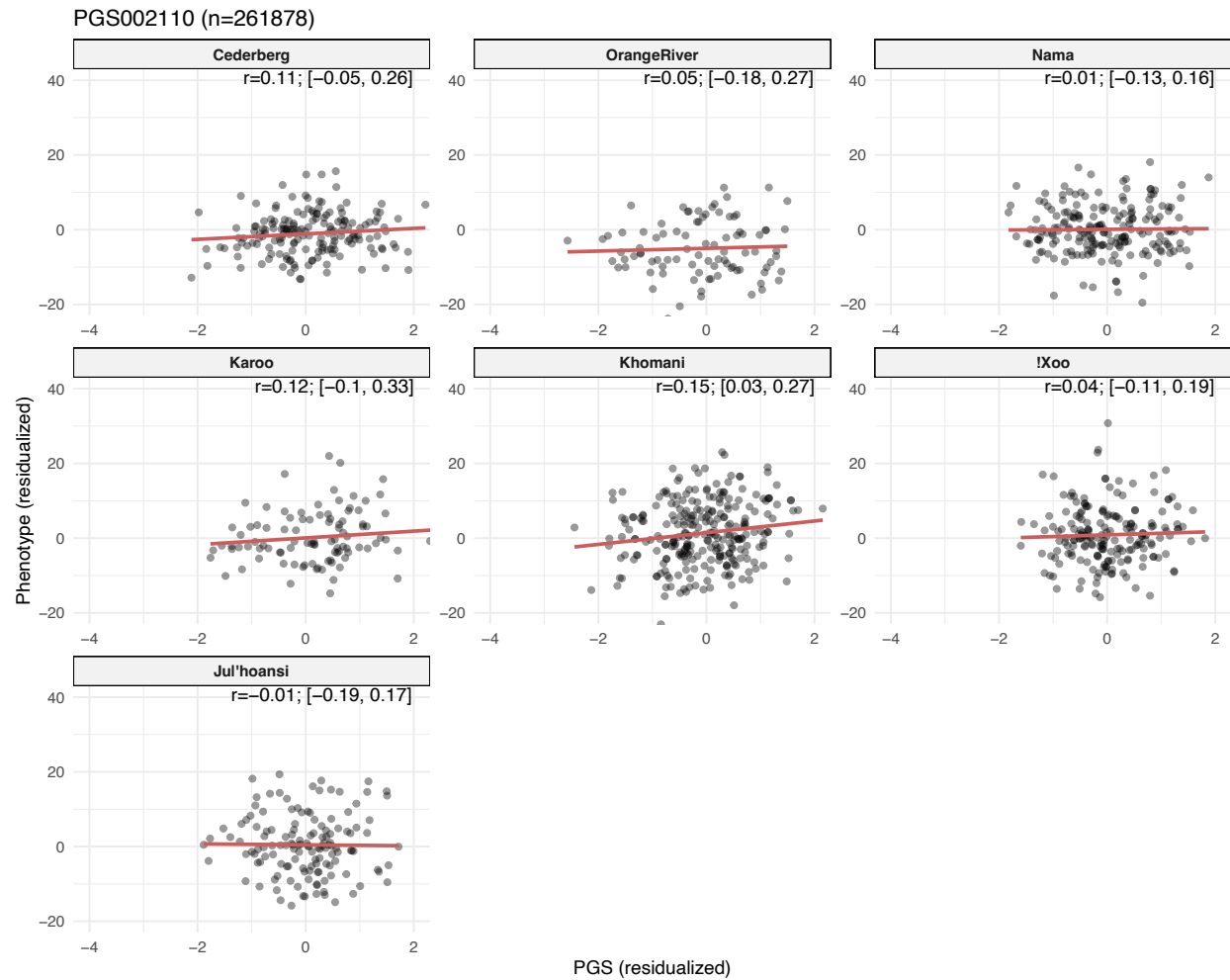
Supplemental Figure 4.21) Melanin against PGS002110 PGS loci. Included in the model are the top 16PCs.



Supplemental Figure 4.22) Partial correlation of PGS and melanin. Included in the model are the top 16PCs. Highlighted in brackets are lower and upper bounds of 95% confidence interval, respectively. A) PGS001897 B) PGS002110.



Supplemental Figure 4.22) Partial correlation of PGS001897 and melanin. Included in the model are the top 16PCs. Highlighted in brackets are lower and upper bounds of 95% confidence interval, respectively. Phenotype is melanin index values revisualized.



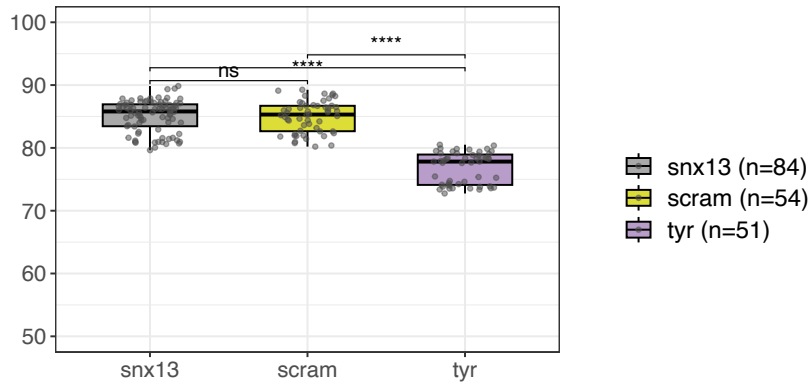
Supplemental Figure 4.23) Partial correlation of PGS002110 and melanin. Included in the model are the top 16PCs. Highlighted in brackets are lower and upper bounds of 95% confidence interval, respectively. Phenotype is melanin index values revisualized.

Functionality

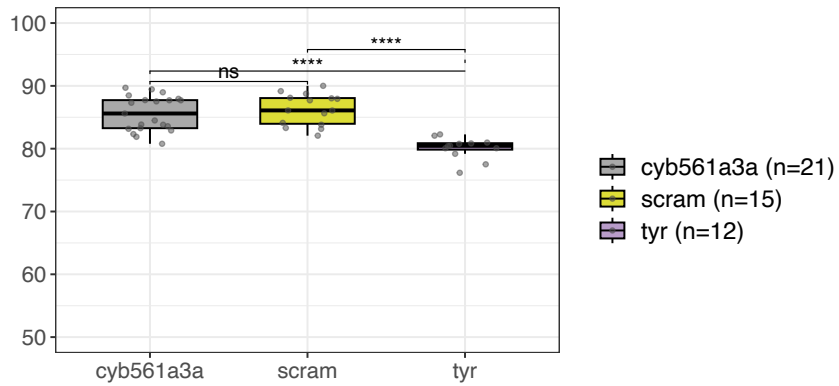
Chr	Gene	Function	Study	Orthologue(s)	Zebrafish Orthologue Co-expression(r)
12	P2RX7	receptor for extracellular ATP	this study	p2rx7	tyr(0.085), slc24a5(0.076), oca2(0.075), tyrp1a(0.074)
19	BLVRB	links bilirubin biosynthesis to yellowing of human skin	Fang et al. (2022); this study	blvrbl	tyrp1a(0.165), tyrp1b(0.159), slc24a5(0.146), slc45a2(0.141), oca2(0.155)
8	ARMC1	mediator of cation binding and mitochondrial dispersal	this study	armc1	slc22a7a(-0.056)
9	TMEFF1	receptor ligand inhibitor activity	this study	tmeff1a / tmeff1b	slc22a7a(-0.17)
10	PTPRE	regulates cell growth and differentiation	this study	ptprea / ptpreb	slc45a2(-0.057)
7	SNX13	regulates lysosomal degradation and G-protein signaling	Martin et al. (2017); Bajpai et al. (2023)	snx13	slc22a7a(-0.076)
11	CYB561A3	regulates SLC24A5 and TYR	Feng et al. (2024)	cyb561a3a / cyb561a3b	slc22a7a(0.380), tyrp1a(-0.180), tyrp1b(-0.193), slc24a5(-0.144), slc45a2(-0.220), oca2(-0.159), dct(-0.195)
15	FRMD5	regulates cell motility, adhesion, and migration	this study	frmd5	slc22a7a(-0.031), tyrp1a(-0.031), tyrp1b(-0.036),
15	ADAL	Implicated as a regulator of TYR in zebrafish	Kim et al. (2014); this study	adal	slc22a7a(-0.023)

Supplemental Table 4.7) Candidate genes investigated for functional validation. Columns are in the following order, left to right: Chromosome (Chr), Gene, function of gene as reported by NCBI, reported studies, zebrafish orthologue(s) with paralogues with expressions in chromatophores in bold, and zebrafish orthologue co-expression with canonical pigmentation genes from Sur et al. (2023). We made the exception to knock out CYB561A3 because of its previously reported strong signal in MNT-1 cell lines, but do not prioritize other genes with paralogues in zebrafish.

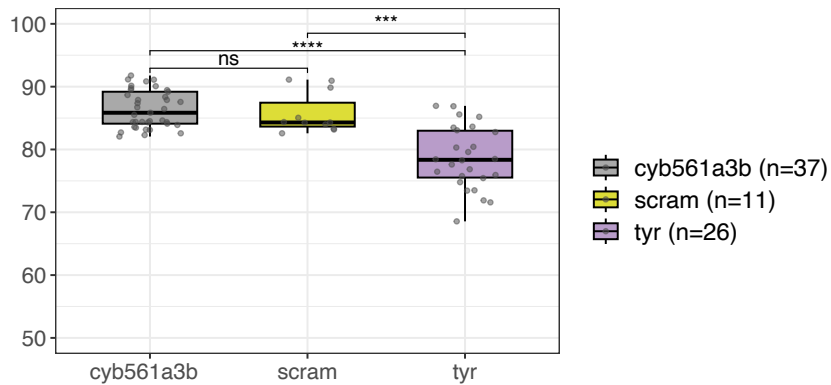
A.



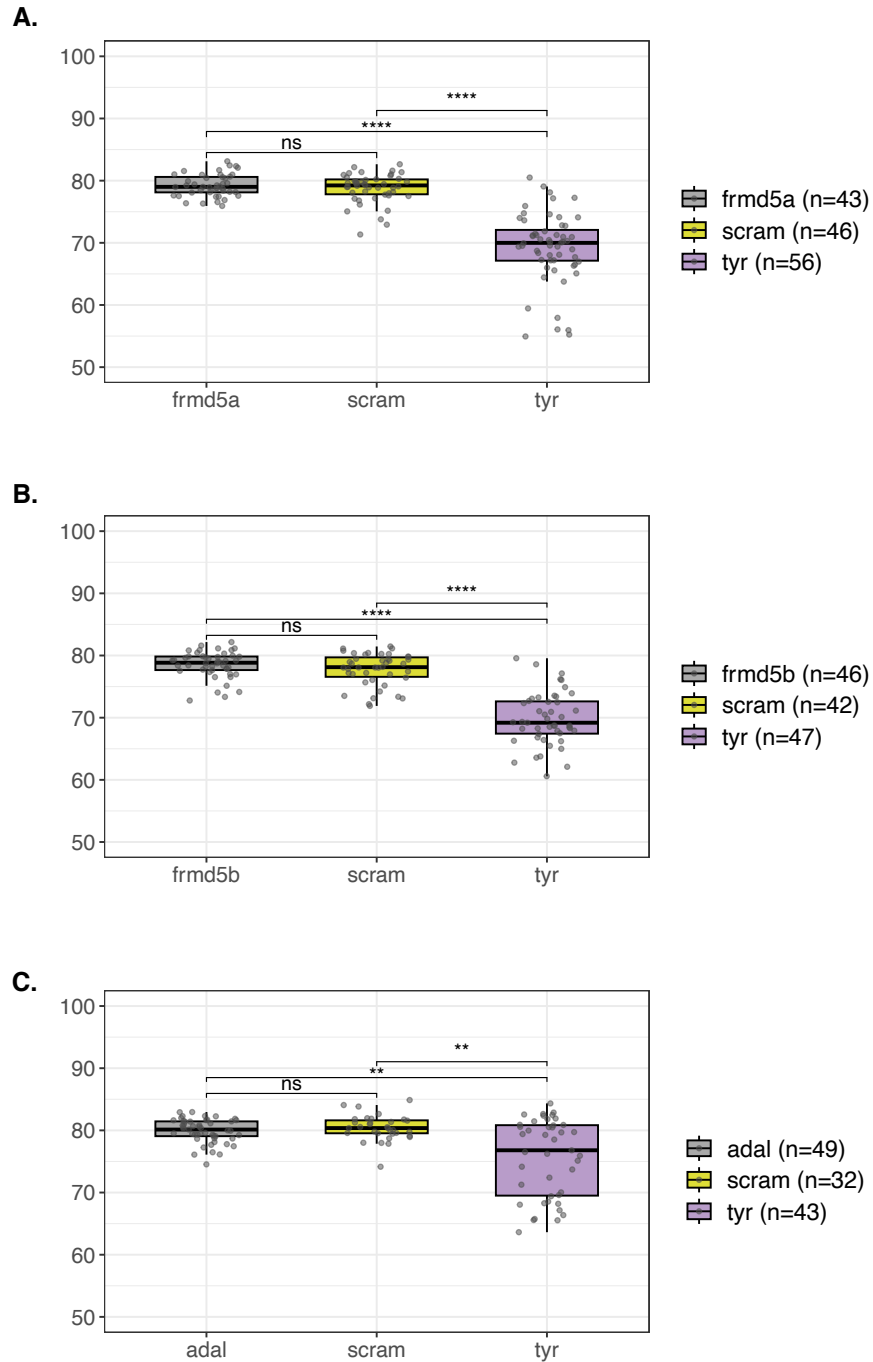
B.



C.

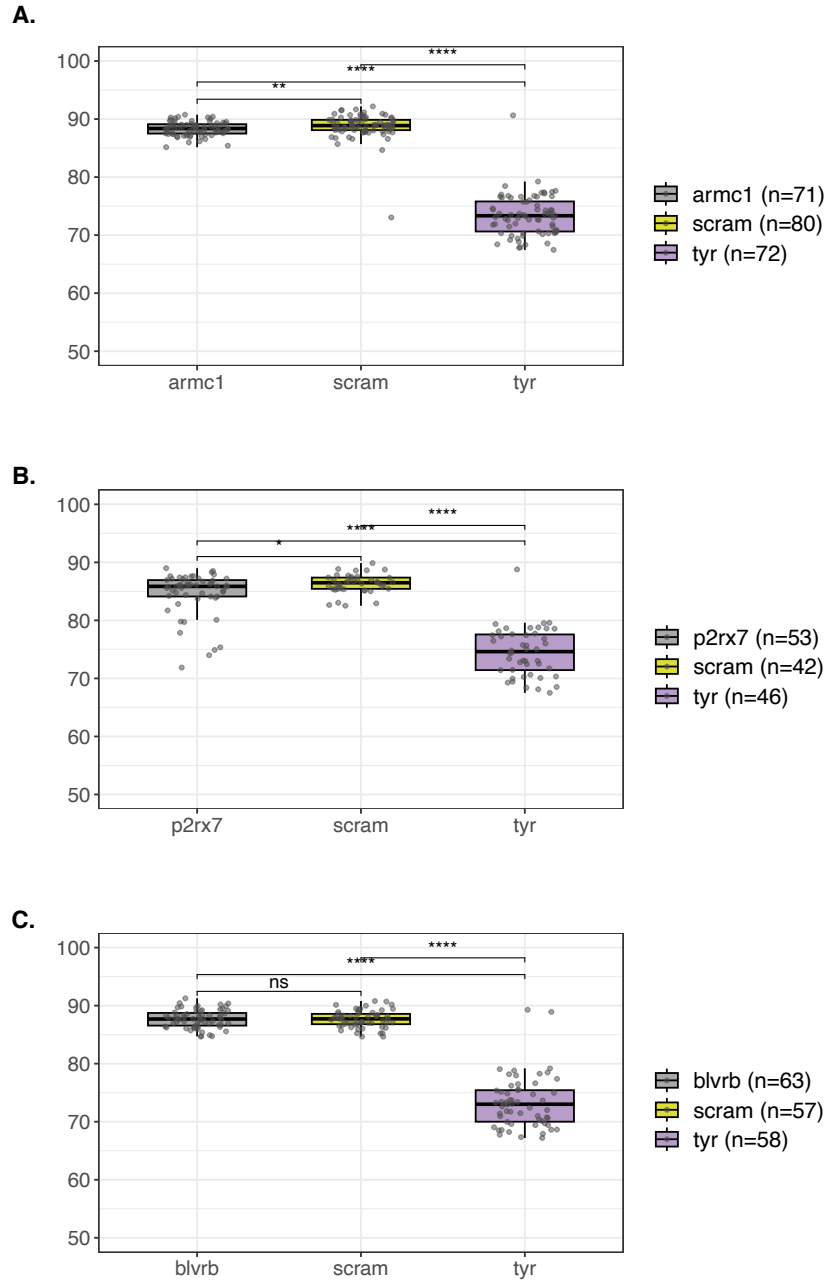


Supplemental Figure 4.24) SNX13 and CYB561A3 experiments, where ‘ns’, ‘**’, ‘***’ and ‘****’ indicate non-significant differences, $p < 0.01$, $p < 0.001$, and $p < 0.0001$, respectively. A) *snx13* B) *cyb561a3a* and C) *cyb561a3b* knockout experiments results. Y-axis is pigmentation score.

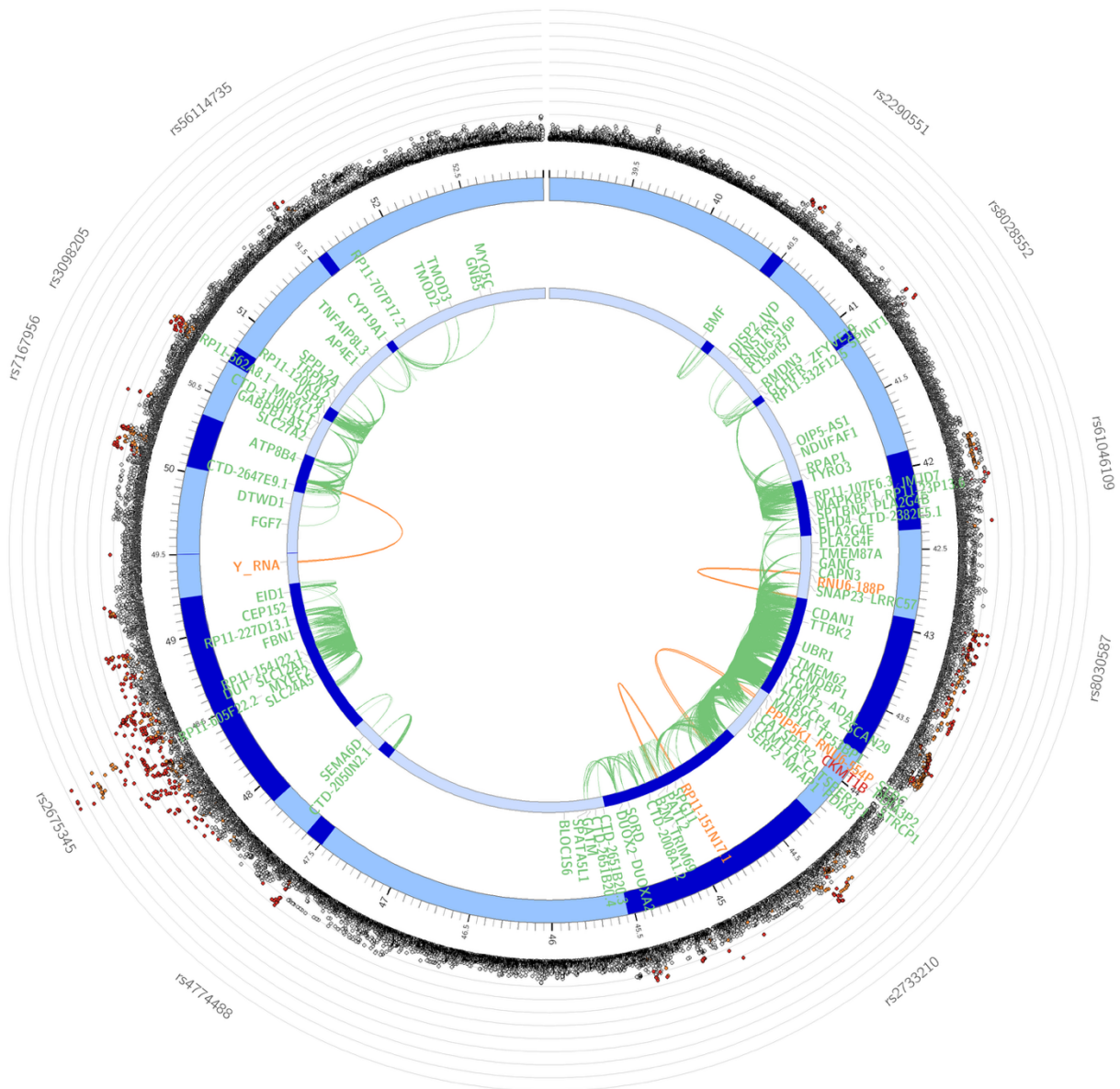


Supplemental Figure 4.25) FRMD5A and ADAL knockouts experiments, where 'ns', '**', '****' and '*****' indicate non-significant differences, $p < 0.01$, $p < 0.001$, and $p < 0.0001$, respectively.

A) *frmd5a* B) *frmd5b* and C) *adal* knockout experiment results. Y-axis is pigmentation score.



Supplemental Figure 4.26) ARMC1, P2RX7 and BLVRB knockouts experiments, where ‘ns’, ‘*’, ‘**’, ‘***’ and ‘****’ indicate non-significant differences, $p < 0.05$, $p < 0.01$, $p < 0.001$, and $p < 0.0001$, respectively. A) *armc1* B) *p2rx7* and C) *blvrbl* knockout experiment results. Y-axis is pigmentation score.



Supplemental Figure 4.27) Chromosome 15 circular Manhattan plot displays SNPs with $P < 0.05$, color-coded by linkage disequilibrium with independent significant SNPs: red ($r^2 > 0.8$), orange ($r^2 > 0.6$), green ($r^2 > 0.4$), blue ($r^2 > 0.2$), and grey ($r^2 \leq 0.2$). Top SNP rsIDs are labeled in the outer ring, with the y-axis scaled to the maximum $-\log_{10}(P)$. The second chromosome ring

marks genomic risk loci in blue, while mapped genes are shown by chromatin interaction (orange), eQTL (green), or both (red). The third ring aligns gene positions without coordinates. Inner links represent chromatin interactions (orange) and eQTL associations (green) used for mapping under user-defined parameters. Plot is limited to 50,000 SNPs and 10,000, displaying only the top-ranked results if thresholds are exceeded. Plot was generated by FUMA and was not further configured.

References

1. Fan, S., Kelly, D.E., Beltrame, M.H., Hansen, M.E.B., Mallick, S., Ranciaro, A., Hirbo, J., Thompson, S., Beggs, W., Nyambo, T., et al. (2019). African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biol* 20, 82. <https://doi.org/10.1186/s13059-019-1679-2>.
2. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538, 201–206. <https://doi.org/10.1038/nature18964>.
3. Bergström, A., Stringer, C., Hajdinjak, M., Scerri, E.M.L., and Skoglund, P. (2021). Origins of modern human ancestry. *Nature* 590, 229–237. <https://doi.org/10.1038/s41586-021-03244-5>.
4. Lorente-Galdos, B., Lao, O., Serra-Vidal, G., Santpere, G., Kuderna, L.F.K., Arauna, L.R., Fadhlou-Zid, K., Pimenoff, V.N., Soodyall, H., Zalloua, P., et al. (2019). Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-Saharan populations. *Genome Biol* 20, 77. <https://doi.org/10.1186/s13059-019-1684-5>.
5. Schlebusch, C.M., Skoglund, P., Sjödin, P., Gattepaille, L.M., Hernandez, D., Jay, F., Li, S., Jongh, M.D., Singleton, A., Blum, M.G.B., et al. (2012). Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History. *Science* 338, 374–379. <https://doi.org/10.1126/science.1227721>.
6. Campbell, M.C., and Tishkoff, S.A. (2008). African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. *Annu Rev Genom Hum G* 9, 403–433. <https://doi.org/10.1146/annurev.genom.9.081307.164258>.
7. Henn, B.M., Steele, T.E., and Weaver, T.D. (2018). Clarifying distinct models of modern human origins in Africa. *Curr Opin Genet Dev* 53, 148–156. <https://doi.org/10.1016/j.gde.2018.10.003>.
8. Skoglund, P., Thompson, J.C., Prendergast, M.E., Mitnik, A., Sirak, K., Hajdinjak, M., Salie, T., Rohland, N., Mallick, S., Peltzer, A., et al. (2017). Reconstructing Prehistoric African Population Structure. *Cell* 171, 59–71.e21. <https://doi.org/10.1016/j.cell.2017.08.049>.
9. Schlebusch, C.M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., Munters, A.R., Vicente, M., Steyn, M., Soodyall, H., et al. (2017). Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science* 358, 652–655. <https://doi.org/10.1126/science.aao6266>.

10. Scerri, E.M.L., Thomas, M.G., Manica, A., Gunz, P., Stock, J.T., Stringer, C., Grove, M., Groucutt, H.S., Timmermann, A., Rightmire, G.P., et al. (2018). Did Our Species Evolve in Subdivided Populations across Africa, and Why Does It Matter? *Trends Ecol Evol* 33, 582–594. <https://doi.org/10.1016/j.tree.2018.05.005>.
11. Chan, E.K.F., Timmermann, A., Baldi, B.F., Moore, A.E., Lyons, R.J., Lee, S.-S., Kalsbeek, A.M.F., Petersen, D.C., Rautenbach, H., Förtsch, H.E.A., et al. (2019). Human origins in a southern African palaeo-wetland and first migrations. *Nature* 575, 185–189. <https://doi.org/10.1038/s41586-019-1714-1>.
12. Henn, B.M., Gignoux, C.R., Jobin, M., Granka, J.M., Macpherson, J.M., Kidd, J.M., Rodríguez-Botigué, L., Ramachandran, S., Hon, L., Brisbin, A., et al. (2011). Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc National Acad Sci* 108, 5154–5162. <https://doi.org/10.1073/pnas.1017511108>.
13. Atkinson, Q.D. (2011). Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa. *Science* 332, 346–349. <https://doi.org/10.1126/science.1199295>.
14. Cavalli-Sforza, L.L., Minch, E., and Mountain, J.L. (1992). Coevolution of genes and languages revisited. *Proc National Acad Sci* 89, 5620–5624. <https://doi.org/10.1073/pnas.89.12.5620>.
15. Greenberg, J.H. (1963). *The Languages of Africa* (Indiana University Press).
16. Schlebusch, C. (2010). Issues raised by use of ethnic-group names in genome study. *Nature* 464, 487–487. <https://doi.org/10.1038/464487a>.
17. Güldemann, T., and Fehn, A.-M. (2014). Beyond ‘Khoisan.’ <https://doi.org/10.1075/cilt.330>.
18. Güldemann, T. (2020). The Language of Hunter-Gatherers. 114–146. <https://doi.org/10.1017/9781139026208.007>.
19. Güldemann, T., and Stoneking, M. (2008). A Historical Appraisal of Clicks: A Linguistic and Genetic Population Perspective. *Annu Rev Anthropol* 37, 93–109. <https://doi.org/10.1146/annurev.anthro.37.081407.085109>.
20. Pickrell, J.K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., Güldemann, T., Kure, B., Mpoloka, S.W., Nakagawa, H., Naumann, C., et al. (2012). The genetic prehistory of southern Africa. *Nat Commun* 3, 1143. <https://doi.org/10.1038/ncomms2140>.
21. Uren, C., Kim, M., Martin, A.R., Bobo, D., Gignoux, C.R., Helden, P.D. van, Möller, M., Hoal, E.G., and Henn, B.M. (2016). Fine-Scale Human Population Structure in Southern Africa Reflects Ecogeographic Boundaries. *Genetics* 204, 303–314. <https://doi.org/10.1534/genetics.116.187369>.

22. Montinaro, F., Busby, G.B.J., Gonzalez-Santos, M., Oosthuizen, O., Oosthuizen, E., Anagnostou, P., Destro-Bisol, G., Pascali, V.L., and Capelli, C. (2017). Complex Ancient Genetic Structure and Cultural Transitions in Southern African Populations. *Genetics* 205, 303–316. <https://doi.org/10.1534/genetics.116.189209>.
23. Vicente, M., Jakobsson, M., Ebbesen, P., and Schlebusch, C.M. (2019). Genetic affinities among southern Africa hunter-gatherers and the impact of admixing farmer and herder populations. *Mol Biol Evol* 36, 1849–1861. <https://doi.org/10.1093/molbev/msz089>.
24. Schlebusch, C.M., Jongh, M. de, and Soodyall, H. (2011). Different contributions of ancient mitochondrial and Y-chromosomal lineages in ‘Karretjie people’ of the Great Karoo in South Africa. *J Hum Genet* 56, 623–630. <https://doi.org/10.1038/jhg.2011.71>.
25. Wood, E.T., Stover, D.A., Ehret, C., Destro-Bisol, G., Spedini, G., McLeod, H., Louie, L., Bamshad, M., Strassmann, B.I., Soodyall, H., et al. (2005). Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. *Eur J Hum Genet* 13, 867–876. <https://doi.org/10.1038/sj.ejhg.5201408>.
26. Barbieri, C., Hübner, A., Macholdt, E., Ni, S., Lippold, S., Schröder, R., Mpoloka, S.W., Purps, J., Roewer, L., Stoneking, M., et al. (2016). Refining the Y chromosome phylogeny with southern African sequences. *Hum Genet* 135, 541–553. <https://doi.org/10.1007/s00439-016-1651-0>.
27. Barbieri, C., Vicente, M., Rocha, J., Mpoloka, S.W., Stoneking, M., and Pakendorf, B. (2013). Ancient substructure in early mtDNA lineages of southern Africa. *Am J Hum Genet* 92, 285–292. <https://doi.org/10.1016/j.ajhg.2012.12.010>.
28. Barbieri, C., Güldemann, T., Naumann, C., Gerlach, L., Berthold, F., Nakagawa, H., Mpoloka, S.W., Stoneking, M., and Pakendorf, B. (2013). Unraveling the complex maternal history of Southern African Khoisan populations. *Am J Phys Anthropol* 153, 435–448. <https://doi.org/10.1002/ajpa.22441>.
29. Scozzari, R., Massaia, A., D’Atanasio, E., Myres, N.M., Perego, U.A., Trombetta, B., and Cruciani, F. (2012). Molecular Dissection of the Basal Clades in the Human Y Chromosome Phylogenetic Tree. *Plos One* 7, e49170. <https://doi.org/10.1371/journal.pone.0049170>.
30. Poznik, G.D., Henn, B.M., Yee, M.-C., Sliwerska, E., Euskirchen, G.M., Lin, A.A., Snyder, M., Quintana-Murci, L., Kidd, J.M., Underhill, P.A., et al. (2013). Sequencing Y Chromosomes Resolves Discrepancy in Time to Common Ancestor of Males Versus Females. *Science* 341, 562–565. <https://doi.org/10.1126/science.1237619>.
31. Gonder, M.K., Mortensen, H.M., Reed, F.A., Sousa, A. de, and Tishkoff, S.A. (2006). Whole-mtDNA Genome Sequence Analysis of Ancient African Lineages. *Mol Biol Evol* 24, 757–768. <https://doi.org/10.1093/molbev/msl209>.

32. Wang, K., Goldstein, S., Bleasdale, M., Clist, B., Bostoen, K., Bakwa-Lufu, P., Buck, L.T., Crowther, A., Dème, A., McIntosh, R.J., et al. (2020). Ancient genomes reveal complex patterns of population movement, interaction, and replacement in sub-Saharan Africa. *Sci Adv* 6, eaaz0183. <https://doi.org/10.1126/sciadv.aaz0183>.
33. Consortium, I.H. 3, Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., et al. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* 467, 52–58. <https://doi.org/10.1038/nature09298>.
34. Behar, D.M., Villemes, R., Soodyall, H., Blue-Smith, J., Pereira, L., Metspalu, E., Scozzari, R., Makkan, H., Tzur, S., Comas, D., et al. (2008). The dawn of human matrilineal diversity. *Am J Hum Genet* 82, 1130–1140. <https://doi.org/10.1016/j.ajhg.2008.04.002>.
35. Robbins, L.H., Brook, G.A., Murphy, M.L., Ivester, A.H., and Campbell, A.C. (2016). Vertebrate Paleobiology and Paleoanthropology. 175–193. https://doi.org/10.1007/978-94-017-7520-5_10.
36. Mackay, A., Stewart, B.A., and Chase, B.M. (2014). Coalescence and fragmentation in the late Pleistocene archaeology of southernmost Africa. *J Hum Evol* 72, 26–51. <https://doi.org/10.1016/j.jhevol.2014.03.003>.
37. Grine, F.E., Bailey, R.M., Harvati, K., Nathan, R.P., Morris, A.G., Henderson, G.M., Ribot, I., and Pike, A.W.G. (2007). Late Pleistocene Human Skull from Hofmeyr, South Africa, and Modern Human Origins. *Science* 315, 226–229. <https://doi.org/10.1126/science.1136294>.
38. Crevecoeur, I., Rougier, H., Grine, F., and Froment, A. (2009). Modern human cranial diversity in the Late Pleistocene of Africa and Eurasia: Evidence from Nazlet Khater, Peștera cu Oase, and Hofmeyr. *Am J Phys Anthropol* 140, 347–358. <https://doi.org/10.1002/ajpa.21080>.
39. Mcbrearty, S., and Brooks, A.S. (2000). The revolution that wasn't: a new interpretation of the origin of modern human behavior. *J Hum Evol* 39, 453–563. <https://doi.org/10.1006/jhev.2000.0435>.
40. Grine, F.E. (2016). Africa from MIS 6-2, Population Dynamics and Paleoenvironments. *Vertebr Paleobiol Pa*, 323–381. https://doi.org/10.1007/978-94-017-7520-5_17.
41. Marean, C.W. (2010). Pinnacle Point Cave 13B (Western Cape Province, South Africa) in context: The Cape Floral kingdom, shellfish, and modern human origins. *J Hum Evol* 59, 425–443. <https://doi.org/10.1016/j.jhevol.2010.07.011>.
42. Berniell-Lee, G., Calafell, F., Bosch, E., Heyer, E., Sica, L., Mouguiama-Daouda, P., Veen, L. van der, Hombert, J.-M., Quintana-Murci, L., and Comas, D. (2009). Genetic and Demographic Implications of the Bantu Expansion: Insights from Human Paternal Lineages. *Mol Biol Evol* 26, 1581–1589. <https://doi.org/10.1093/molbev/msp069>.

43. Henn, B.M., Gignoux, C., Lin, A.A., Oefner, P.J., Shen, P., Scozzari, R., Cruciani, F., Tishkoff, S.A., Mountain, J.L., and Underhill, P.A. (2008). Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. *Proc. Natl. Acad. Sci.* *105*, 10693–10698. <https://doi.org/10.1073/pnas.0801184105>.
44. Soodyall, H., Makkan, H., Haycock, P., and Naidoo, T. The genetic prehistory of the Khoe and San. *Southern African Humanities* *20*, 37–48.
45. Cruciani, F., Fratta, R.L., Santolamazza, P., Sellitto, D., Pascone, R., Moral, P., Watson, E., Guida, V., Colomb, E.B., Zaharova, B., et al. (2004). Phylogeographic Analysis of Haplogroup E3b (E-M215) Y Chromosomes Reveals Multiple Migratory Events Within and Out Of Africa. *Am J Hum Genetics* *74*, 1014–1022. <https://doi.org/10.1086/386294>.
46. Tishkoff, S.A., Gonder, M.K., Henn, B.M., Mortensen, H., Knight, A., Gignoux, C., Fernandopulle, N., Lema, G., Nyambo, T.B., Ramakrishnan, U., et al. (2007). History of Click-Speaking Populations of Africa Inferred from mtDNA and Y Chromosome Genetic Variation. *Mol Biol Evol* *24*, 2180–2195. <https://doi.org/10.1093/molbev/msm155>.
47. Pargeter, J., MacKay, A., Mitchell, P., Shea, J., and Stewart, B.A. (2016). Primordialism and the ‘Pleistocene San’ of southern Africa. *Antiquity* *90*, 1072–1079. <https://doi.org/10.15184/aqy.2016.100>.
48. Stewart, B.A., Zhao, Y., Mitchell, P.J., Dewar, G., Gleason, J.D., and Blum, J.D. (2020). Ostrich eggshell bead strontium isotopes reveal persistent macroscale social networking across late Quaternary southern Africa. *Proc National Acad Sci* *117*, 6453–6462. <https://doi.org/10.1073/pnas.1921037117>.
49. Klein, R.G., and Steele, T.E. (2013). Archaeological shellfish size and later human evolution in Africa. *Proc National Acad Sci* *110*, 10910–10915. <https://doi.org/10.1073/pnas.1304750110>.
50. Breton, G., Schlebusch, C.M., Lombard, M., Sjödin, P., Soodyall, H., and Jakobsson, M. (2014). Lactase Persistence Alleles Reveal Partial East African Ancestry of Southern African Khoe Pastoralists. *Curr. Biol.* *24*, 852–858. <https://doi.org/10.1016/j.cub.2014.02.041>.
51. Macholdt, E., Lede, V., Barbieri, C., Mpoloka, S.W., Chen, H., Slatkin, M., Pakendorf, B., and Stoneking, M. (2014). Tracing Pastoralist Migrations to Southern Africa with Lactase Persistence Alleles. *Curr Biol* *24*, 875–879. <https://doi.org/10.1016/j.cub.2014.03.027>.
52. H., G., and R., E. (2014). *The Shaping of South African Society, 1652–1840*.
53. Worden, N. (2017). Slavery at the Cape. *Oxford Research Encyclopedias, African History*, 1–20. <https://doi.org/10.1093/acrefore/9780190277734.013.76>.
54. Rohde, R.F., and Hoffman, M.T. One Hundred Years of Separation: The Historical Ecology of a South African “Coloured Reserve.”

55. Penn, N. (2005). *The Forgotten Frontier: Colonist and Khoisan on the Cape's Northern Frontier in the 18th Century* (Ohio University Press).
56. Adhikari, M. (2014). *Genocide on Settler Frontiers: When Hunter-Gatherers and Commercial Stock Farmers Clash* M. Adhikari, ed. (University of Cape Town Press)
<https://doi.org/10.3167/9781782387381>.
57. Africa, S.S.A., Republic of South (2023). *Census 2022 Population Count Results 10 October 2023*.
https://census.statssa.gov.za/assets/documents/2022/P03014_Census_2022_Statistical_Release.pdf.
58. Lankheet, I., Hammarén, R., Caballero, L.X.A., Larena, M., Malmström, H., Jolly, C., Soodyall, H., Jongh, M. de, and Schlebusch, C. (2024). Wide-scale Geographical Analysis of Genetic Ancestry in the South African Coloured Population. *bioRxiv*, 2024.06.13.598620.
<https://doi.org/10.1101/2024.06.13.598620>.
59. Daya, M., Merwe, L. van der, Galal, U., Möller, M., Salie, M., Chimusa, E.R., Galanter, J.M., Helden, P.D. van, Henn, B.M., Gignoux, C.R., et al. (2013). A Panel of Ancestry Informative Markers for the Complex Five-Way Admixed South African Coloured Population. *PLoS ONE* 8, e82224. <https://doi.org/10.1371/journal.pone.0082224>.
60. Wit, E. de, Delport, W., Rugamika, C.E., Meintjes, A., Möller, M., Helden, P.D. van, Seoighe, C., and Hoal, E.G. (2010). Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. *Hum Genet* 128, 145–153.
<https://doi.org/10.1007/s00439-010-0836-1>.
61. Quintana-Murci, L., Harmant, C., Quach, H., Balanovsky, O., Zaporozhchenko, V., Bormans, C., Helden, P.D. van, Hoal, E.G., and Behar, D.M. (2010). Strong Maternal Khoisan Contribution to the South African Coloured Population: A Case of Gender-Biased Admixture. *Am J Hum Genetics* 86, 611–620. <https://doi.org/10.1016/j.ajhg.2010.02.014>.
62. Petersen, D.C., Libiger, O., Tindall, E.A., Hardie, R.-A., Hannick, L.I., Glashoff, R.H., Mukerji, M., Consortium, I.G.V., Fernandez, P., Haacke, W., et al. (2013). Complex Patterns of Genomic Admixture within Southern Africa. *Plos Genet* 9, e1003309.
<https://doi.org/10.1371/journal.pgen.1003309>.
63. Chen, Y.-S., Olckers, A., Schurr, T.G., Kogelnik, A.M., Huoponen, K., and Wallace, D.C. (2000). mtDNA Variation in the South African Kung and Khwe—and Their Genetic Relationships to Other African Populations. *Am J Hum Genetics* 66, 1362–1383.
<https://doi.org/10.1086/302848>.
64. Underhill, P.A., Passarino, G., Lin, A.A., Shen, P., Lahr, M.M., Foley, R.A., Oefner, P.J., and Cavalli-Sforza, L.L. (2001). The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet* 65, 43–62.
<https://doi.org/10.1046/j.1469-1809.2001.6510043.x>.

65. Barnard, A. (2012). *Hunters and Herders of Southern Africa* (Cambridge University Press) <https://doi.org/10.1017/cbo9781139166508>.
66. Patterson, N., Petersen, D.C., Ross, R.E. van der, Sudoyo, H., Glashoff, R.H., Marzuki, S., Reich, D., and Hayes, V.M. (2010). Genetic structure of a unique admixed population: implications for medical research. *Hum. Mol. Genet.* *19*, 411–419. <https://doi.org/10.1093/hmg/ddp505>.
67. Fortes-Lima, C.A., Burgarella, C., Hammarén, R., Eriksson, A., Vicente, M., Jolly, C., Semo, A., Gunnink, H., Pacchiarotti, S., Mundeke, L., et al. (2024). The genetic legacy of the expansion of Bantu-speaking peoples in Africa. *Nature* *625*, 540–547. <https://doi.org/10.1038/s41586-023-06770-6>.
68. Sengupta, D., Choudhury, A., Fortes-Lima, C., Aron, S., Whitelaw, G., Bostoen, K., Gunnink, H., Chousou-Polydouri, N., Delius, P., Tollman, S., et al. (2021). Genetic substructure and complex demographic history of South African Bantu speakers. *Nat. Commun.* *12*, 2080. <https://doi.org/10.1038/s41467-021-22207-y>.
69. Patterson, N., Price, A.L., and Reich, D. (2006). Population Structure and Eigenanalysis. *PLoS Genet.* *2*, e190. <https://doi.org/10.1371/journal.pgen.0020190>.
70. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* *19*, 1655–1664. <https://doi.org/10.1101/gr.094052.109>.
71. Swinford, N.A., Prall, S.P., Gopalan, S., Williams, C.M., Sheehama, J., Scelza, B.A., and Henn, B.M. (2023). Increased homozygosity due to endogamy results in fitness consequences in a human population. *Proc. Natl. Acad. Sci.* *120*, e2309552120. <https://doi.org/10.1073/pnas.2309552120>.
72. Robbins, L.H., Campbell, A.C., Murphy, M.L., Brook, G.A., Liang, F., Skaggs, S.A., Srivastava, P., Mabuse, A.A., and Badenhorst, S. (2008). Recent Archaeological Research at Toteng, Botswana: Early Domesticated Livestock in the Kalahari. *J. Afr. Archaeol.* *6*, 131–149. <https://doi.org/10.3213/1612-1651-10106>.
73. Vicente, M., Lankheet, I., Russell, T., Hollfelder, N., Coetzee, V., Soodyall, H., Jongh, M.D., and Schlebusch, C.M. (2021). Male-biased migration from East Africa introduced pastoralism into southern Africa. *BMC Biol.* *19*, 259. <https://doi.org/10.1186/s12915-021-01193-z>.
74. Ioannidis, A.G., Blanco-Portillo, J., Sandoval, K., Hagelberg, E., Miquel-Poblete, J.F., Moreno-Mayar, J.V., Rodríguez-Rodríguez, J.E., Quinto-Cortés, C.D., Auckland, K., Parks, T., et al. (2020). Native American gene flow into Polynesia predating Easter Island settlement. *Nature* *583*, 572–577. <https://doi.org/10.1038/s41586-020-2487-2>.

75. Hilmarsson, H., Kumar, A.S., Rastogi, R., Bustamante, C.D., Montserrat, D.M., and Ioannidis, A.G. (2021). High Resolution Ancestry Deconvolution for Next Generation Genomic Data. *bioRxiv*, 2021.09.19.460980. <https://doi.org/10.1101/2021.09.19.460980>.
76. African Languages (2000). (Cambridge University Press).
77. Blench, R. (2006). *Archaeology, Language, and the African Past* (AltaMira Press).
78. Schlebusch, C.M., and Jakobsson, M. (2018). Tales of Human Migration, Admixture, and Selection in Africa. *Annu Rev Genom Hum G* 19, 1–24. <https://doi.org/10.1146/annurev-genom-083117-021759>.
79. Beck, R.B. (2014). *The History of South Africa* (ABC-CLIO).
80. Hudjashov, G., Karafet, T.M., Lawson, D.J., Downey, S., Savina, O., Sudoyo, H., Lansing, J.S., Hammer, M.F., and Cox, M.P. (2017). Complex Patterns of Admixture across the Indonesian Archipelago. *Mol. Biol. Evol.* 34, 2439–2452. <https://doi.org/10.1093/molbev/msx196>.
81. Mbeki, L., and Rossum, M. van (2016). Private slave trade in the Dutch Indian Ocean world: a study into the networks and backgrounds of th. *Slavery & Abolition*. <https://doi.org/10.1080/0144039x.2016.1159004>.
82. Division, L. of Congress.F.R. (1997). *South Africa : A Country Study* R. M. Byrnes, ed. (The Division: For sale by the Supt. of Docs., U.S. G.P.O.).
83. Al-Asadi, H., Petkova, D., Stephens, M., and Novembre, J. (2019). Estimating recent migration and population-size surfaces. *PLoS Genet.* 15, e1007908. <https://doi.org/10.1371/journal.pgen.1007908>.
84. Thompson, E.A. (2013). Identity by Descent: Variation in Meiosis, Across Genomes, and in Populations. *Genetics* 194, 301–326. <https://doi.org/10.1534/genetics.112.148825>.
85. Hollfelder, N., Breton, G., Sjödin, P., and Jakobsson, M. (2021). The deep population history in Africa. *Hum Mol Genet* 30, ddab005. <https://doi.org/10.1093/hmg/ddab005>.
86. Choudhury, A., Aron, S., Sengupta, D., Hazelhurst, S., and Ramsay, M. (2018). African genetic diversity provides novel insights into evolutionary history and local adaptations. *Hum. Mol. Genet.* 27, R209–R218. <https://doi.org/10.1093/hmg/ddy161>.
87. Choudhury, A., Ramsay, M., Hazelhurst, S., Aron, S., Bardien, S., Botha, G., Chimusa, E.R., Christoffels, A., Gamielien, J., Sefid-Dashti, M.J., et al. (2017). Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nat. Commun.* 8, 2062. <https://doi.org/10.1038/s41467-017-00663-9>.

88. Chimusa, E.R., Daya, M., Möller, M., Ramesar, R., Henn, B.M., Helden, P.D. van, Mulder, N.J., and Hoal, E.G. (2013). Determining Ancestry Proportions in Complex Admixture Scenarios in South Africa Using a Novel Proxy Ancestry Selection Method. *Plos One* 8, e73971. <https://doi.org/10.1371/journal.pone.0073971>.
89. Hollfelder, N., Erasmus, J.C., Hammaren, R., Vicente, M., Jakobsson, M., Greeff, J.M., and Schlebusch, C.M. (2020). Patterns of African and Asian admixture in the Afrikaner population of South Africa. *Bmc Biol* 18, 16. <https://doi.org/10.1186/s12915-020-0746-1>.
90. Philip, D. (2003). The Cape Huguenots and Their Legacy in Apartheid South Africa. In *Memory and Identity: The Huguenots in France and the Atlantic Diaspora*, B. V. Ruymbeke and R. J. Sparks, eds. (Columbia, University of South Carolina Press), pp. 285–309.
91. Rees, W. (1970). The Oxford History of South Africa. Volume I, South Africa to 1870. Edited by Monica Wilson and Leonard Thompson. (New York: Oxford University Press. 1969. Pp. xviii, 502. \$8.00.). *Am. Hist. Rev.* 75, 899–900. <https://doi.org/10.1086/ahr/75.3.899>.
92. Etherington, N. (2002). The Great Treks: The Transformation of Southern Africa, 1815-1854. *J. South. Afr. Stud.* 28, 651–660. <https://doi.org/10.1080/0305707022000006567>.
93. Kootker, L.M., Mbeki, L., Morris, A.G., Kars, H., and Davies, G.R. (2016). Dynamics of Indian Ocean Slavery Revealed through Isotopic Data from the Colonial Era Cobern Street Burial Site, Cape Town, South Africa (1750-1827). *PLoS ONE* 11, e0157750. <https://doi.org/10.1371/journal.pone.0157750>.
94. Vink, M. (2003). “The World’s Oldest Trade”: Dutch Slavery and Slave Trade in the Indian Ocean in the Seventeenth Century. *J. World Hist.* 14, 131–177. <https://doi.org/10.1353/jwh.2003.0026>.
95. Worden, N. (2005). Indian Ocean slavery and its demise in the Cape Colony. In: *Abolition and its aftermath in the Indian Ocean Africa and Asia*. In, G. Campbell, ed., pp. 29–49.
96. Plessis, I.D. du (1947). *The Cape Malays* (Maskew Miller).
97. GUELKE, L. (1976). Frontier settlement in early Dutch South Africa. *Ann. Assoc. Am. Geogr.* 66, 25–42. <https://doi.org/10.1111/j.1467-8306.1976.tb01070.x>.
98. Mitchell, L.J. (2002). Traces in the Landscape: Hunters, Herders and Farmers on the Cedarberg Frontier, South Africa, 1725-95. *The Journal of African History* 43, 431–450.
99. Reynolds, A.W., Grote, M.N., Myrick, J.W., Al-Hindi, D.R., Siford, R.L., Mastoras, M., Möller, M., and Henn, B.M. (2023). Persistence of Matrilocal Postmarital Residence Across Multiple Generations in Southern Africa. *Hum. Nat.* 34, 295–323. <https://doi.org/10.1007/s12110-023-09452-4>.

100. Desai, A., and Vahed, G. (2022). The Oxford Handbook of South African History. <https://doi.org/10.1093/oxfordhb/9780190921767.013.28>.
101. Schlebusch, C.M., and Soodyall, H. (2012). Extensive Population Structure in San, Khoe, and Mixed Ancestry Populations from Southern Africa Revealed by 44 Short 5-SNP Haplotypes. *Hum Biol* 84, 695–724. <https://doi.org/10.3378/027.084.0603>.
102. Gopalan, S., Berl, R.E.W., Myrick, J.W., Garfield, Z.H., Reynolds, A.W., Bafens, B.K., Belbin, G., Mastoras, M., Williams, C., Daya, M., et al. (2022). Hunter-gatherer genomes reveal diverse demographic trajectories during the rise of farming in Eastern Africa. *Curr. Biol.* 32, 1852-1860.e5. <https://doi.org/10.1016/j.cub.2022.02.050>.
103. Fenner, J.N. (2005). Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthr.* 128, 415–423. <https://doi.org/10.1002/ajpa.20188>.
104. Barry, M. (2017). Now Another Thing Must Happen: Richtersveld and the Dilemmas of Land Reform in Post-Apartheid South Africa. *S Afr J Hum Rights* 20, 355–382. <https://doi.org/10.1080/19962126.2004.11864826>.
105. Ross, R. (1983). *Cape of Torments: Slavery and Resistance in South Africa* (Routledge & Kegan Paul).
106. Dooling, W. (2005). The Origins and Aftermath of the Cape Colony’s “Hottentot Code” of 1809. *Kronos*, 50–61.
107. Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M.O., Choudhury, A., et al. (2015). The African Genome Variation Project shapes medical genetics in Africa. *Nature* 517, 327–332. <https://doi.org/10.1038/nature13997>.
108. Cann, H.M., Toma, C. de, Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A Human Genome Diversity Cell Line Panel. *Science* 296, 261–262. <https://doi.org/10.1126/science.296.5566.261b>.
109. Vespasiani, D.M., Jacobs, G.S., Cook, L.E., Brucato, N., Leavesley, M., Kinipi, C., Ricaut, F.-X., Cox, M.P., and Romero, I.G. (2022). Denisovan introgression has shaped the immune system of present-day Papuans. *PLOS Genet.* 18, e1010470. <https://doi.org/10.1371/journal.pgen.1010470>.
110. Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. <https://doi.org/10.1038/nature15393>.
111. Crawford, N.G., Kelly, D.E., Hansen, M.E.B., Beltrame, M.H., Fan, S., Bowman, S.L., Jewett, E., Ranciaro, A., Thompson, S., Lo, Y., et al. (2017). Loci associated with skin

pigmentation identified in African populations. *Sci New York N Y* 358, ean8433.
<https://doi.org/10.1126/science.aan8433>.

112. Genovese, G., Rockweiler, N.B., Gorman, B.R., Bigdeli, T.B., Pato, M.T., Pato, C.N., Ichihara, K., and McCarroll, S.A. (2024). BCFtools/liftover: an accurate and comprehensive tool to convert genetic variants across genome assemblies. *Bioinformatics* 40, btac038.
<https://doi.org/10.1093/bioinformatics/btac038>.

113. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience* 10, giab008. <https://doi.org/10.1093/gigascience/giab008>.

114. Delaneau, O., Zagury, J.-F., Robinson, M.R., Marchini, J.L., and Dermitzakis, E.T. (2019). Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* 10, 5436.
<https://doi.org/10.1038/s41467-019-13225-y>.

115. Williams, C.M., Scelza, B.A., Daya, M., Lange, E.M., Gignoux, C.R., and Henn, B.M. (2020). A rapid, accurate approach to inferring pedigrees in endogamous populations. *Biorxiv*, 2020.02.25.965376. <https://doi.org/10.1101/2020.02.25.965376>.

116. Caballero, M., Seidman, D.N., Qiao, Y., Sannerud, J., Dyer, T.D., Lehman, D.M., Curran, J.E., Duggirala, R., Blangero, J., Carmi, S., et al. (2019). Crossover interference and sex-specific genetic maps shape identical by descent sharing in close relatives. *PLoS Genet.* 15, e1007979.
<https://doi.org/10.1371/journal.pgen.1007979>.

117. Campbell, C.L., Furlotte, N.A., Eriksson, N., Hinds, D., and Auton, A. (2015). Escape from crossover interference increases with maternal age. *Nat. Commun.* 6, 6260.
<https://doi.org/10.1038/ncomms7260>.

118. Bhérier, C., Campbell, C.L., and Auton, A. (2017). Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nat. Commun.* 8, 14994.
<https://doi.org/10.1038/ncomms14994>.

119. Freyman, W.A., McManus, K.F., Shringarpure, S.S., Jewett, E.M., Bryc, K., Team, T. 23 and M.R., and Auton, A. (2020). Fast and Robust Identity-by-Descent Inference with the Templated Positional Burrows–Wheeler Transform. *Mol. Biol. Evol.* 38, 2131–2151.
<https://doi.org/10.1093/molbev/msaa328>.

120. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics* 26, 2867–2873. <https://doi.org/10.1093/bioinformatics/btq559>.

121. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., Bakker, P.I.W. de, Daly, M.J., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genetics* 81, 559–575.
<https://doi.org/10.1086/519795>.

122. Behr, A.A., Liu, K.Z., Liu-Fang, G., Nakka, P., and Ramachandran, S. (2016). pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics* 32, 2817–2823. <https://doi.org/10.1093/bioinformatics/btw327>.
123. Henn, B.M., Botigué, L.R., Gravel, S., Wang, W., Brisbin, A., Byrnes, J.K., Fadhlou-Zid, K., Zalloua, P.A., Moreno-Estrada, A., Bertranpetit, J., et al. (2012). Genomic Ancestry of North Africans Supports Back-to-Africa Migrations. *PLoS Genet.* 8, e1002397. <https://doi.org/10.1371/journal.pgen.1002397>.
124. Wickham, H. (2011). ggplot2. *Wiley Interdiscip. Rev.: Comput. Stat.* 3, 180–185. <https://doi.org/10.1002/wics.147>.
125. Pless, E., Eckburg, A.M., and Henn, B.M. (2023). Predicting Environmental and Ecological Drivers of Human Population Structure. *Mol. Biol. Evol.* 40, msad094. <https://doi.org/10.1093/molbev/msad094>.
126. Zhou, Y., Browning, S.R., and Browning, B.L. (2020). A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data. *Am. J. Hum. Genet.* 106, 426–437. <https://doi.org/10.1016/j.ajhg.2020.02.010>.
127. Karger, D.N., Conrad, O., Böhrner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., Zimmermann, N.E., Linder, H.P., and Kessler, M. (2017). Climatologies at high resolution for the earth's land surface areas. *Sci. Data* 4, 170122. <https://doi.org/10.1038/sdata.2017.122>.
128. Amatulli, G., McInerney, D., Sethi, T., Strobl, P., and Domisch, S. (2020). Geomorpho90m, empirical evaluation and accuracy assessment of global high-resolution geomorphometric layers. *Sci. Data* 7, 162. <https://doi.org/10.1038/s41597-020-0479-6>.
129. Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J.C., Sampson, C.C., Kanae, S., and Bates, P.D. (2017). A high-accuracy map of global terrain elevations. *Geophys. Res. Lett.* 44, 5844–5853. <https://doi.org/10.1002/2017gl072874>.
130. Saada, J.N., Kalantzis, G., Shyr, D., Cooper, F., Robinson, M., Gusev, A., and Palamara, P.F. (2020). Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. *Nat. Commun.* 11, 6130. <https://doi.org/10.1038/s41467-020-19588-x>.
131. Andersen, J.D., Meyer, O.S., Simão, F., Jannuzzi, J., Carvalho, E., Andersen, M.M., Pereira, V., Børsting, C., Morling, N., and Gusmão, L. (2020). Skin pigmentation and genetic variants in an admixed Brazilian population of primarily European ancestry. *Int J Legal Med* 134, 1569–1579. <https://doi.org/10.1007/s00414-020-02307-y>.
132. Mishra, A., Nizammuddin, S., Mallick, C.B., Singh, S., Prakash, S., Siddiqui, N.A., Rai, N., Carlus, S.J., Sudhakar, D.V.S., Tripathi, V.P., et al. (2017). Genotype-Phenotype Study of the Middle Gangetic Plain in India Shows Association of rs2470102 with Skin Pigmentation. *J Invest Dermatol* 137, 670–677. <https://doi.org/10.1016/j.jid.2016.10.043>.

133. Beleza, S., Johnson, N.A., Candille, S.I., Absher, D.M., Coram, M.A., Lopes, J., Campos, J., Araújo, I.I., Anderson, T.M., Vilhjálmsson, B.J., et al. (2013). Genetic Architecture of Skin and Eye Color in an African-European Admixed Population. *Plos Genet* 9, e1003372. <https://doi.org/10.1371/journal.pgen.1003372>.
134. Larsson, M., and Pedersen, N.L. (2004). Genetic correlations among texture characteristics in the human iris. *Mol. Vis.* 10, 821–831.
135. Dorgaleleh, S., Naghipoor, K., Barahouie, A., Dastaviz, F., and Oladnabi, M. (2020). Molecular and biochemical mechanisms of human iris color: A comprehensive review. *J. Cell. Physiol.* 235, 8972–8982. <https://doi.org/10.1002/jcp.29824>.
136. Sulem, P., Gudbjartsson, D.F., Stacey, S.N., Helgason, A., Rafnar, T., Magnusson, K.P., Manolescu, A., Karason, A., Palsson, A., Thorleifsson, G., et al. (2007). Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat Genet* 39, 1443–1452. <https://doi.org/10.1038/ng.2007.13>.
137. Keyser, C., Gibert, M., Soulier, A., Balaesque, P., and King, T.E. (2016). Chapter Nine Human Phenotypic Diversity An Evolutionary Perspective. *Curr Top Dev Biol* 119, 349–390. <https://doi.org/10.1016/bs.ctdb.2016.02.001>.
138. Sturm, R.A., and Duffy, D.L. (2012). Human pigmentation genes under environmental selection. *Genome Biol* 13, 248. <https://doi.org/10.1186/gb-2012-13-9-248>.
139. Duffy, D.L., and Sturm, R.A. (2007). A Three–Single-Nucleotide Polymorphism Haplotype in Intron 1 of OCA2 Explains Most Human Eye-Color Variation.
140. Sturm, R.A., and Frudakis, T.N. (2004). Eye colour: portals into pigmentation genes and ancestry. *Trends Genet* 20, 327–332. <https://doi.org/10.1016/j.tig.2004.06.010>.
141. Sulem, P., Gudbjartsson, D.F., Stacey, S.N., Helgason, A., Rafnar, T., Jakobsdottir, M., Steinberg, S., Gudjonsson, S.A., Palsson, A., Thorleifsson, G., et al. (2008). Two newly identified genetic determinants of pigmentation in Europeans. *Nat Genet* 40, 835–837. <https://doi.org/10.1038/ng.160>.
142. Donnelly, M.P., Paschou, P., Grigorenko, E., Gurwitz, D., Barta, C., Lu, R.-B., Zhukova, O.V., Kim, J.-J., Siniscalco, M., New, M., et al. (2012). A global view of the OCA2-HERC2 region and pigmentation. *Hum Genet* 131, 683–696. <https://doi.org/10.1007/s00439-011-1110-x>.
143. Wilde, S., Timpson, A., Kirsanow, K., Kaiser, E., Kayser, M., Unterländer, M., Hollfelder, N., Potekhina, I.D., Schier, W., Thomas, M.G., et al. (2014). Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proc National Acad Sci* 111, 4832–4837. <https://doi.org/10.1073/pnas.1316513111>.
144. Frost, P. (2006). European hair and eye color A case of frequency-dependent sexual selection? *Evol. Hum. Behav.* 27, 85–103. <https://doi.org/10.1016/j.evolhumbehav.2005.07.002>.

145. Cain, F.E., and Yamaguchi, K. (2024). Effect of iris pigmentation of blue and brown eyed individuals with European ancestry on ability to see in low light conditions after a short-term dark adaption period. *bioRxiv*, 2024.01.17.576074. <https://doi.org/10.1101/2024.01.17.576074>.
146. Goel, N., Terman, M., and Terman, J.S. (2002). Depressive symptomatology differentiates subgroups of patients with seasonal affective disorder. *Dépress. Anxiety* 15, 34–41. <https://doi.org/10.1002/da.1083>.
147. Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 528, 499–503. <https://doi.org/10.1038/nature16152>.
148. Liu, F., Wen, B., and Kayser, M. (2013). Colorful DNA polymorphisms in humans. *Semin. Cell Dev. Biol.* 24, 562–575. <https://doi.org/10.1016/j.semcdb.2013.03.013>.
149. Cook, A.L., Chen, W., Thurber, A.E., Smit, D.J., Smith, A.G., Bladen, T.G., Brown, D.L., Duffy, D.L., Pastorino, L., Bianchi-Scarra, G., et al. (2009). Analysis of Cultured Human Melanocytes Based on Polymorphisms within the SLC45A2/MATP, SLC24A5/NCKX5, and OCA2/P Loci. *J Invest Dermatol* 129, 392–405. <https://doi.org/10.1038/jid.2008.211>.
150. Edwards, M., Cha, D., Krithika, S., Johnson, M., Cook, G., and Parra, E.J. (2016). Iris pigmentation as a quantitative trait: variation in populations of European, East Asian and South Asian ancestry and association with candidate gene polymorphisms. *Pigment Cell Melanoma Res.* 29, 141–162. <https://doi.org/10.1111/pcmr.12435>.
151. Norton, H.L., Edwards, M., Krithika, S., Johnson, M., Werren, E.A., and Parra, E.J. (2015). Quantitative assessment of skin, hair, and iris variation in a diverse sample of individuals and associated genetic variation: QUANTITATIVE ASSESSMENT OF PIGMENTARY PHENOTYPE. *Am J Phys Anthropol* 160, 570–581. <https://doi.org/10.1002/ajpa.22861>.
152. Lamason, R.L., Mohideen, M.-A.P.K., Mest, J.R., Wong, A.C., Norton, H.L., Aros, M.C., Jurynec, M.J., Mao, X., Humphreville, V.R., Humbert, J.E., et al. (2005). SLC24A5, a Putative Cation Exchanger, Affects Pigmentation in Zebrafish and Humans. *Science* 310, 1782–1786. <https://doi.org/10.1126/science.1116238>.
153. Norton, H.L., Kittles, R.A., Parra, E., McKeigue, P., Mao, X., Cheng, K., Canfield, V.A., Bradley, D.G., McEvoy, B., and Shriver, M.D. (2007). Genetic Evidence for the Convergent Evolution of Light Skin in Europeans and East Asians. *Mol Biol Evol* 24, 710–722. <https://doi.org/10.1093/molbev/msl203>.
154. Norton, H.L., Edwards, M., Krithika, S., Johnson, M., Werren, E.A., and Parra, E.J. (2016). Quantitative assessment of skin, hair, and iris variation in a diverse sample of individuals and associated genetic variation. *Am J Phys Anthropol* 160, 570–581. <https://doi.org/10.1002/ajpa.22861>.

155. Beleza, S., Santos, A.M., McEvoy, B., Alves, I., Martinho, C., Cameron, E., Shriver, M.D., Parra, E.J., and Rocha, J. (2012). The Timing of Pigmentation Lightening in Europeans. *Mol Biol Evol* 30, 24–35. <https://doi.org/10.1093/molbev/mss207>.
156. Beleza, S., Santos, A.M., McEvoy, B., Alves, I., Martinho, C., Cameron, E., Shriver, M.D., Parra, E.J., and Rocha, J. (2013). The Timing of Pigmentation Lightening in Europeans. *Mol Biol Evol* 30, 24–35. <https://doi.org/10.1093/molbev/mss207>.
157. Meredith, P., and Sarna, T. (2006). The physical and chemical properties of eumelanin. *Pigment Cell Res.* 19, 572–594. <https://doi.org/10.1111/j.1600-0749.2006.00345.x>.
158. Byard, P.J. (1981). Quantitative genetics of human skin color. *Am. J. Phys. Anthr.* 24, 123–137. <https://doi.org/10.1002/ajpa.1330240506>.
159. Relethford, J.H. (2000). Human skin color diversity is highest in sub-Saharan African populations. *Hum Biol* 72, 773–780.
160. Relethford, J.H. (2002). Apportionment of global human genetic diversity based on craniometrics and skin color. *Am J Phys Anthropol* 118, 393–398. <https://doi.org/10.1002/ajpa.10079>.
161. Jablonski, N.G., and Chaplin, G. (2000). The evolution of human skin coloration. *J Hum Evol* 39, 57–106. <https://doi.org/10.1006/jhev.2000.0403>.
162. Chaplin, G. (2004). Geographic distribution of environmental factors influencing human skin coloration. *Am J Phys Anthropol* 125, 292–302. <https://doi.org/10.1002/ajpa.10263>.
163. Brenner, M., and Hearing, V.J. (2008). The Protective Role of Melanin Against UV Damage in Human Skin†. *Photochem Photobiol* 84, 539–549. <https://doi.org/10.1111/j.1751-1097.2007.00226.x>.
164. Jablonski, N.G., and Chaplin, G. (2000). The evolution of human skin coloration. *J Hum Evol* 39, 57–106. <https://doi.org/10.1006/jhev.2000.0403>.
165. Jablonski, N.G., and Chaplin, G. (2010). Human skin pigmentation as an adaptation to UV radiation. *Proc National Acad Sci* 107, 8962–8968. <https://doi.org/10.1073/pnas.0914628107>.
166. Martin, A.R., Lin, M., Granka, J.M., Myrick, J.W., Liu, X., Sockell, A., Atkinson, E.G., Werely, C.J., Möller, M., Sandhu, M.S., et al. (2017). An Unexpectedly Complex Architecture for Skin Pigmentation in Africans. *Biorxiv*, 200139. <https://doi.org/10.1101/200139>.
167. Crawford, N.G., Kelly, D.E., Hansen, M.E.B., Beltrame, M.H., Fan, S., Bowman, S.L., Jewett, E., Ranciaro, A., Thompson, S., Lo, Y., et al. (2017). Loci associated with skin pigmentation identified in African populations. *Science* 358, eaan8433. <https://doi.org/10.1126/science.aan8433>.

168. Lin, M., Siford, R.L., Martin, A.R., Nakagome, S., Möller, M., Hoal, E.G., Bustamante, C.D., Gignoux, C.R., and Henn, B.M. (2018). Rapid evolution of a skin-lightening allele in southern African Khoesans. *Proc National Acad Sci* 115, 201801948. <https://doi.org/10.1073/pnas.1801948115>.
169. Martin, A.R., Lin, M., Granka, J.M., Myrick, J.W., Liu, X., Sockell, A., Atkinson, E.G., Werely, C.J., Möller, M., Sandhu, M.S., et al. (2017). An Unexpectedly Complex Architecture for Skin Pigmentation in Africans. *Cell* 171, 1340-1353.e14. <https://doi.org/10.1016/j.cell.2017.11.015>.
170. Feng, Y., Xie, N., Inoue, F., Fan, S., Saskin, J., Zhang, C., Zhang, F., Hansen, M.E.B., Nyambo, T., Mpoloka, S.W., et al. (2024). Integrative functional genomic analyses identify genetic variants influencing skin pigmentation in Africans. *Nat. Genet.*, 1–15. <https://doi.org/10.1038/s41588-023-01626-1>.
171. HARRISON, G.A., and OWEN, J.J.T. (1964). Studies on the inheritance of human skin colour. *Ann. Hum. Genet.* 28, 27–37. <https://doi.org/10.1111/j.1469-1809.1964.tb00457.x>.
172. Conery, M., and Grant, S.F.A. (2023). Human height: a model common complex trait. *Ann. Hum. Biol.* 50, 258–266. <https://doi.org/10.1080/03014460.2023.2215546>.
173. Mallick, C.B., Iliescu, F.M., Möls, M., Hill, S., Tamang, R., Chaubey, G., Goto, R., Ho, S.Y.W., Romero, I.G., Crivellaro, F., et al. (2013). The Light Skin Allele of SLC24A5 in South Asians and Europeans Shares Identity by Descent. *Plos Genet* 9, e1003912. <https://doi.org/10.1371/journal.pgen.1003912>.
174. Candille, S.I., Absher, D.M., Beleza, S., Bauchet, M., McEvoy, B., Garrison, N.A., Li, J.Z., Myers, R.M., Barsh, G.S., Tang, H., et al. (2012). Genome-Wide Association Studies of Quantitatively Measured Skin, Hair, and Eye Pigmentation in Four European Populations. *Plos One* 7, e48294. <https://doi.org/10.1371/journal.pone.0048294>.
175. Zaitlen, N., Pasaniuc, B., Sankararaman, S., Bhatia, G., Zhang, J., Gusev, A., Young, T., Tandon, A., Pollack, S., Vilhjálmsdóttir, B.J., et al. (2014). Leveraging population admixture to characterize the heritability of complex traits. *Nat. Genet.* 46, 1356–1362. <https://doi.org/10.1038/ng.3139>.
176. Soejima, M., and Koda, Y. (2007). Population differences of two coding SNPs in pigmentation-related genes SLC24A5 and SLC45A2. *Int. J. Leg. Med.* 121, 36–39. <https://doi.org/10.1007/s00414-006-0112-z>.
177. Graf, J., Voisey, J., Hughes, I., and Daal, A. van (2007). Promoter polymorphisms in the MATP (SLC45A2) gene are associated with normal human skin color variation. *Hum. Mutat.* 28, 710–717. <https://doi.org/10.1002/humu.20504>.
178. Sturm, R.A. (2009). Molecular genetics of human pigmentation diversity. *Hum Mol Genet* 18, R9–R17. <https://doi.org/10.1093/hmg/ddp003>.

179. Jacobs, L.C., Hamer, M.A., Gunn, D.A., Deelen, J., Lall, J.S., Heemst, D. van, Uh, H.-W., Hofman, A., Uitterlinden, A.G., Griffiths, C.E.M., et al. (2015). A Genome-Wide Association Study Identifies the Skin Color Genes *IRF4*, *MC1R*, *ASIP*, and *BNC2* Influencing Facial Pigmented Spots. *J. Investig. Dermatol.* 135, 1735–1742. <https://doi.org/10.1038/jid.2015.62>.
180. Wei, A.-H., Zang, D.-J., Zhang, Z., Liu, X.-Z., He, X., Yang, L., Wang, Y., Zhou, Z.-Y., Zhang, M.-R., Dai, L.-L., et al. (2013). Exome Sequencing Identifies *SLC24A5* as a Candidate Gene for Nonsyndromic Oculocutaneous Albinism. *J. Investig. Dermatol.* 133, 1834–1840. <https://doi.org/10.1038/jid.2013.49>.
181. Kim, M.Y., Lee, H.-E., Im, M., Lee, Y., Kim, C.-D., Lee, J.-H., and Seo, Y.-J. (2013). Effect of Adenosine on Melanogenesis in B16 Cells and Zebrafish. *Ann. Dermatol.* 26, 209–213. <https://doi.org/10.5021/ad.2014.26.2.209>.
182. Price, A.L., Weale, M.E., Patterson, N., Myers, S.R., Need, A.C., Shianna, K.V., Ge, D., Rotter, J.I., Torres, E., Taylor, K.D., et al. (2008). Long-Range LD Can Confound Genome Scans in Admixed Populations. *Am. J. Hum. Genet.* 83, 132–135. <https://doi.org/10.1016/j.ajhg.2008.06.005>.
183. Park, L. (2019). Population-specific long-range linkage disequilibrium in the human genome and its influence on identifying common disease variants. *Sci. Rep.* 9, 11380. <https://doi.org/10.1038/s41598-019-47832-y>.
184. Watanabe, K., Taskesen, E., Bochoven, A. van, and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* 8, 1826. <https://doi.org/10.1038/s41467-017-01261-5>.
185. Lin, M., Park, D.S., Zaitlen, N.A., Henn, B.M., and Gignoux, C.R. (2021). Admixed Populations Improve Power for Variant Discovery and Portability in Genome-Wide Association Studies. *Front. Genet.* 12, 673167. <https://doi.org/10.3389/fgene.2021.673167>.
186. Mester, R., Hou, K., Ding, Y., Meeks, G., Burch, K.S., Bhattacharya, A., Henn, B.M., and Pasaniuc, B. (2023). Impact of cross-ancestry genetic architecture on GWASs in admixed populations. *Am. J. Hum. Genet.* 110, 927–939. <https://doi.org/10.1016/j.ajhg.2023.05.001>.
187. Yao, Y., and Ochoa, A. (2023). Limitations of principal components in quantitative genetic association models for human studies. *eLife* 12, e79238. <https://doi.org/10.7554/elife.79238>.
188. Bajpai, V.K., Swigut, T., Mohammed, J., Naqvi, S., Arreola, M., Tycko, J., Kim, T.C., Pritchard, J.K., Bassik, M.C., and Wysocka, J. (2023). A genome-wide genetic screen uncovers determinants of human pigmentation. *Science* 381, eade6289. <https://doi.org/10.1126/science.ade6289>.
189. Liu, G.-S., Peshavariya, H., Higuchi, M., Brewer, A.C., Chang, C.W.T., Chan, E.C., and Dusting, G.J. (2012). Microphthalmia-associated transcription factor modulates expression of

- NADPH oxidase type 4: A negative regulator of melanogenesis. *Free Radic. Biol. Med.* 52, 1835–1843. <https://doi.org/10.1016/j.freeradbiomed.2012.02.040>.
190. Privé, F., Aschard, H., Carmi, S., Folkersen, L., Hoggart, C., O'Reilly, P.F., and Vilhjálmsson, B.J. (2022). Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *Am. J. Hum. Genet.* 109, 12–23. <https://doi.org/10.1016/j.ajhg.2021.11.008>.
191. Holtzman, N.G., Iovine, M.K., Liang, J.O., and Morris, J. (2016). Learning to Fish with Genetics: A Primer on the Vertebrate Model *Danio rerio*. *Genetics* 203, 1069–1089. <https://doi.org/10.1534/genetics.116.190843>.
192. Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L., et al. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature* 496, 498–503. <https://doi.org/10.1038/nature12111>.
193. Dooley, K., and Zon, L.I. (2000). Zebrafish: a model system for the study of human disease. *Curr Opin Genet Dev* 10, 252–256. [https://doi.org/10.1016/s0959-437x\(00\)00074-5](https://doi.org/10.1016/s0959-437x(00)00074-5).
194. Postlethwait, J.H., Woods, I.G., Ngo-Hazelett, P., Yan, Y.-L., Kelly, P.D., Chu, F., Huang, H., Hill-Force, A., and Talbot, W.S. (2000). Zebrafish Comparative Genomics and the Origins of Vertebrate Chromosomes. *Genome Res* 10, 1890–1902. <https://doi.org/10.1101/gr.164800>.
195. Jao, L.-E., Wente, S.R., and Chen, W. (2013). Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. *P Natl Acad Sci Usa* 110, 13904–13909. <https://doi.org/10.1073/pnas.1308335110>.
196. Quigley, I.K., and Parichy, D.M. (2002). Pigment pattern formation in zebrafish: A model for developmental genetics and the evolution of form. *Microsc. Res. Tech.* 58, 442–455. <https://doi.org/10.1002/jemt.10162>.
197. Gur, D., Bain, E.J., Johnson, K.R., Aman, A.J., Pasolli, H.A., Flynn, J.D., Allen, M.C., Deheyn, D.D., Lee, J.C., Lippincott-Schwartz, J., et al. (2020). In situ differentiation of iridophore crystallotypes underlies zebrafish stripe patterning. *Nat. Commun.* 11, 6391. <https://doi.org/10.1038/s41467-020-20088-1>.
198. Qu, J., Yan, M., Fang, Y., Zhao, J., Xu, T., Liu, F., Zhang, K., He, L., Jin, L., and Sun, D. (2023). Zebrafish in dermatology: a comprehensive review of their role in investigating abnormal skin pigmentation mechanisms. *Front. Physiol.* 14, 1296046. <https://doi.org/10.3389/fphys.2023.1296046>.
199. Saunders, L.M., Mishra, A.K., Aman, A.J., Lewis, V.M., Toomey, M.B., Packer, J.S., Qiu, X., McFaline-Figueroa, J.L., Corbo, J.C., Trapnell, C., et al. (2019). Thyroid hormone regulates distinct paths to maturation in pigment cell lineages. *eLife* 8, e45181. <https://doi.org/10.7554/elife.45181>.

200. Praetorius, C., Grill, C., Stacey, S.N., Metcalf, A.M., Gorkin, D.U., Robinson, K.C., Van Otterloo, E., Kim, R.S.Q., Bergsteinsdottir, K., Ogmundsdottir, M.H., et al. (2013). A Polymorphism in IRF4 Affects Human Pigmentation through a Tyrosinase-Dependent MITF/TFAP2A Pathway. *Cell* 155, 1022–1033. <https://doi.org/10.1016/j.cell.2013.10.022>.
201. King, R.A., Pietsch, J., Fryer, J.P., Savage, S., Brott, M.J., Russell-Eggitt, I., Summers, C.G., and Oetting, W.S. (2003). Tyrosinase gene mutations in oculocutaneous albinism 1 (OCA1): definition of the phenotype. *Hum Genet* 113, 502–513. <https://doi.org/10.1007/s00439-003-0998-1>.
202. Sur, A., Wang, Y., Capar, P., Margolin, G., Prochaska, M.K., and Farrell, J.A. (2023). Single-cell analysis of shared signatures and transcriptional diversity during zebrafish development. *Dev. Cell* 58, 3028–3047.e12. <https://doi.org/10.1016/j.devcel.2023.11.001>.
203. Liu, F., Visser, M., Duffy, D.L., Hysi, P.G., Jacobs, L.C., Lao, O., Zhong, K., Walsh, S., Chaitanya, L., Wollstein, A., et al. (2015). Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up. *Hum Genet* 134, 823–835. <https://doi.org/10.1007/s00439-015-1559-0>.
204. Ju, D., and Mathieson, I. (2021). The evolution of skin pigmentation-associated variation in West Eurasia. *Proc National Acad Sci* 118, e2009227118. <https://doi.org/10.1073/pnas.2009227118>.
205. Pośpiech, E., Wojas-Pelc, A., Walsh, S., Liu, F., Maeda, H., Ishikawa, T., Skowron, M., Kayser, M., and Branicki, W. (2014). The common occurrence of epistasis in the determination of human pigmentation and its impact on DNA-based pigmentation phenotype prediction. *Forensic Sci Int Genetics* 11, 64–72. <https://doi.org/10.1016/j.fsigen.2014.01.012>.
206. Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci.* 109, 1193–1198. <https://doi.org/10.1073/pnas.1119675109>.
207. Yang, G., Mishra, M., and Perera, M.A. (2023). Multi-Omics Studies in Historically Excluded Populations: The Road to Equity. *Clin. Pharmacol. Ther.* 113, 541–556. <https://doi.org/10.1002/cpt.2818>.
208. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am. J. Hum. Genet.* 88, 76–82. <https://doi.org/10.1016/j.ajhg.2010.11.011>.
209. Kroll, F., Powell, G.T., Ghosh, M., Antinucci, P., Hearn, T.J., Tunbak, H., Lim, S., Dennis, H.W., Fernandez, J.M., Whitmore, D., et al. (2020). A simple and effective F0 knockout method for rapid screening of behaviour and other complex phenotypes. *Biorxiv*, 2020.06.04.133462. <https://doi.org/10.1101/2020.06.04.133462>.

210. Cui, Y., Xu, J., Cheng, M., Liao, X., and Peng, S. (2018). Review of CRISPR/Cas9 sgRNA Design Tools. *Interdiscip Sci Comput Life Sci* 10, 455–465. <https://doi.org/10.1007/s12539-018-0298-z>.
211. Heigwer, F., Kerr, G., and Boutros, M. (2014). E-CRISP: fast CRISPR target site identification. *Nat Methods* 11, 122–123. <https://doi.org/10.1038/nmeth.2812>.
212. Fan, S., Spence, J.P., Feng, Y., Hansen, M.E.B., Terhorst, J., Beltrame, M.H., Ranciaro, A., Hirbo, J., Beggs, W., Thomas, N., et al. (2023). Whole-genome sequencing reveals a complex African population demographic history and signatures of local adaptation. *Cell* 186, 923–939.e14. <https://doi.org/10.1016/j.cell.2023.01.042>.
213. Scrucca, L., Fop, M., Murphy, T.B., and Raftery, A.E. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J.* 8, 289–317.
214. McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* 3, 861. <https://doi.org/10.21105/joss.00861>.
215. Malzer, C., and Baum, M. (2020). A Hybrid Approach To Hierarchical Density-based Cluster Selection. 2020 IEEE Int. Conf. Multisens. Fusion Integr. Intell. Syst. (MFI) 00, 223–228. <https://doi.org/10.1109/mfi49285.2020.9235263>.
216. Diaz-Papkovich, A., Zabad, S., Ben-Eghan, C., Anderson-Trocmé, L., Femerling, G., Nathan, V., Patel, J., and Gravel, S. (2023). Topological stratification of continuous genetic variation in large biobanks. *bioRxiv*, 2023.07.06.548007. <https://doi.org/10.1101/2023.07.06.548007>.
217. Reich, D., Thangaraj, K., Patterson, N., Price, A.L., and Singh, L. (2009). Reconstructing Indian population history. *Nature* 461, 489–494. <https://doi.org/10.1038/nature08365>.
218. Metspalu, M., Mondal, M., and Chaubey, G. (2018). The genetic makings of South Asia. *Curr. Opin. Genet. Dev.* 53, 128–133. <https://doi.org/10.1016/j.gde.2018.09.003>.
219. Kerdoncuff, E., Skov, L., Patterson, N., Zhao, W., Lueng, Y.Y., Schellenberg, G.D., Smith, J.A., Dey, S., Ganna, A., Dey, A., et al. (2024). 50,000 years of Evolutionary History of India: Insights from ~2,700 Whole Genome Sequences. *bioRxiv*, 2024.02.15.580575. <https://doi.org/10.1101/2024.02.15.580575>.
220. Moorjani, P., Thangaraj, K., Patterson, N., Lipson, M., Loh, P.-R., Govindaraj, P., Berger, B., Reich, D., and Singh, L. (2013). Genetic Evidence for Recent Population Mixture in India. *Am. J. Hum. Genet.* 93, 422–438. <https://doi.org/10.1016/j.ajhg.2013.07.006>.