The study of present-day human variation to explore past evolutionary events

Alexandra Blant

A thesis submitted to the Faculty of Graduate Studies
in partial fulfillment of the requirements of the degree:

Doctor of Philosophy

Department of Biochemistry and Medical Genetics
University of Manitoba
Winnipeg

Table of Contents

List of Tables

List of Figures

List of Equations (Chapter 3)

List of Appendix Tables

Acknowledgements

Many people have supported me throughout my PhD journey, and I would like to use this space to express my thanks.

To begin with, I would like to acknowledge my original supervisor Dr. Trevor Pemberton who was among the first professors stationed in Winnipeg with a strong biology foundation specializing in computational genetics. As your PhD student, I learned a great deal and I remain thankful for the opportunity to transition into a computational field which complemented my prior genetics background.

A heartfelt thank you to Dr. Michelle Liu and Dr. Colin Garroway, both of whom kindly stepped in to accept a supervisory role in 2019, spent time familiarizing themselves with my thesis projects and provided support to facilitate the completion of my PhD. Michelle and Colin: thank you for your advice and your willingness to help me in all respects. I am additionally grateful for my committee members Dr. Linda Larcombe and Dr. Barb Triggs-Raine; to Linda for ensuring my thesis appropriately reflects anthropological thinking and to Barb for your always practical advice and support.
Additionally, a word of gratitude to Drs. Gilbert Arthur, Sara Good, Lisa Lix, Mark Nachtigal, Donna Turner and Jeff Wigle for discussion or help at various moments of my PhD journey.

Next, I'd like to acknowledge those I had the pleasure of meeting and working with during the 2016 Cameroon field-mission. First and foremost, I am indebted to Drs. Paul Verdu, Alain Froment and Fernando Ramirez-Rozzi for guiding me in the practical aspects of collecting biological data in a remote setting. Thank you in particular to Dr. Paul Verdu for the great conversations and for going above and beyond to support my academic progress. Additionally, thank you to Mr. Alain Fezzu (our driver during the fieldwork) for your positive energy, your help in the various aspects of our field-work and for letting me practice my French with you.

My sincere gratitude to the Bezan, Tikar, Baka and Nzime people of Cameroon who participated in our study. The work would not have been possible without you. In particular, thank you to Lily and Sylvie from Le Bosquet for making me feel welcome in your village and your home, respectively.

I am additionally grateful to the funding sources that supported me, and the research conducted during my PhD: NSERC's Postgraduate Scholarship and Discovery Grant and the University of Manitoba's Mindel Rady Olenick Fellowship in Human Genetics, Phyllis J McAlpine Graduate Fellowship and the President's Graduate Scholarship in Human Genetics.

Finally, thank you to my family for your unwavering support.

Abstract

This work examines present-day human variation from two distinct perspectives and methodologies: **Chapter 2** catalogs morphological characteristics of 337 Central African Pygmies and non-Pygmies and provides a novel perspective on the body proportions of Pygmy groups as compared to their non-Pygmy neighbours. The latter sets the stage for downstream genetic studies to inform on the molecular mechanisms which underpin biological scaling and the evolutionary histories of populations inhabiting the Congo basin. **Chapter 3** describes a bioinformatics method (wLOD) to detect genomic signals of relatedness (autozygosity) in worldwide populations. The performance of wLOD is assessed in both simulated and next-generation sequencing data of various marker density and is shown to have comparable or improved performance to several other publicly available tools. More broadly, the work presented here explores the distribution and properties of autozygous regions at different ancestral depths and demonstrates the ability of wLOD (and other autozygosity detection tools) to shed light on the role of recessive variation in human diversity and on the overarching effect of limited gene pools of past generations on the genomic architecture of present-day populations.

List of Abbreviations

Abbreviations are presented for Chapters 2 and 3 in alphabetical order:

AIC	Aikake information criterion
AOH	Absence of heterozygosity
aDNA	Ancient DNA
BC	Before the Common Era
BIC	Bayesian Information Criterion
BMI	Body mass index
CC	Central Cameroon
CEPH	Centre d'Etude du Polymorphisme Humain
CI	Confidence interval
DBP	Diastolic blood pressure
DRC	Democratic Republic of Congo
GARLIC	Genomic Autozygosity Regions Likelihood-based Inference and Classification
GHBP	Growth hormone binding protein
GIANT	The Genetic Investigation of ANthropometric Traits Consortium
GWAS	Genome wide association study
HDL	High-density lipoprotein
HGDP	Human genome diversity project
HLA	Human Leukocyte antigen
HWE	Hardy Weinberg Equilibrium
IBD	Identity by descent
IBS	Identical by state
IGF1-GH	Insulin growth factor 1/growth hormone axis
Kya	Thousand years ago
LCSH	long contiguous stretch of homozygosity
LCT	Lactase gene
LD	Linkage disequilibrium
LDL	Low-density lipoprotein
LMM	Linear mixed effect model

LOD	Logarithm of odds score
LOH	Loss of heterozygosity
MAF	Minor allele frequency
MDS	Multidimensional scaling
mtDNA	Mitochondrial DNA
MRCA	Most recent common ancestor
Mya	Million years ago
NCD	Non-communicable disease (risk factor collaboration)
NGS	Next-generation sequencing
OMIM	Online Mendelian Inheritance in Man
OmniExp	Illumina OmniExpress-24 BeadChip
Omni2.5	Illumina HumanOmni2.5-8 BeadChip
SBP	Systolic blood pressure
SEC	Southeast Cameroon
SF	Skin-fold
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
ROA	Run of autozygosity
ROH	Run of homozygosity
TGP	The 1000 Genomes Project Phase 3, release v5a
UCSC	University of California – Santa Cruz
WES	Roche Nimblegen SeqCap EZ Human Exome Library v3.0 system
WGS	Whole genome sequencing
WHtR	Waist to height ratio
wLOD	Weighted likelihood logarithm of odds score
WWI	World War I
ya	years ago

Chapter 1 Introductory chapter

At the present time, the underlying reason for human variation is known to be the result of past evolutionary events. This understanding, however, is the product of at least three thousand years of human inquiry into medicine, biology, and the earth sciences. In fact, the notion that human variation was intrinsically connected to the study of evolutionary processes on Earth is a relatively new concept, even though both fields existed in parallel for many centuries before intersecting. The realization that these two concepts are inexorably linked arose ~100 years ago and culminated into concerted efforts by numerous scientists to unite the fields of anthropology, paleontology, genetics and evolution into one "Modern Synthesis" [1].

1.1 The study of human variation through history

For the majority of recorded history, the study of human variation, perhaps unsurprisingly is synonymous with the classification of superficial, easily observable characteristics like skin color and facial features. In the Western traditions, information pertaining to human differences in ancient literature (Classical Antiquity, Middle Ages) is scattered and mostly in the form of incidental observations about foreign peoples [2]. A shift in perspective begins with the Renaissance and the exploration of the America's in the 1500's, eventually transforming into the systematic study of human groups familiar to us today [3].

To say the least, the historical progression of the study of human variation is an expansive topic, given that human variation is a central theme in nearly any inquiry relating to the biology or history of the human species. The closer one gets to present time, the more difficult it becomes to discuss human variation studies without discussing its intersections with other related fields, such as medicine, paleontology (anatomical studies of skeletal remains), earth science, psychology and philosophy.

This thesis explores patterns of human variation. To provide context to the research presented in Chapters 2 and 3, the aim of this introductory chapter is to briefly orient the reader to the progression of human variation studies throughout history.

1.1.1 Classical Antiquity

Though true beginnings are impossible to define concretely, it is likely that the first people to document human variability were artists. One of the first and clearest examples of human classifications is depicted by an ancient Egyptian mural found in the Royal Tombs of the 19th Dynasty which dates to ~12th century BC. Four racial groups from the point of view of the Egyptian artist are classified in different paint colors: the Egyptians in red, Africans in black, Asians in yellow, and Europeans in white [2]. Other art from this period instead represents commercial or political relationships between the Egyptians and other groups; the latter made distinct not only by physical traits but also via attention to the costumes and material culture (e.g. tools and weapons). Such classifications during this time appear fluid and vary depending on the artist who drew it. Written sources additionally included references to foreign lands and the types of people who inhabit them [2][4]. Though ancient Egyptian art and written records make clear distinctions between human groups using phenotypic (facial characteristics, skin and hair pigmentation) cultural or geographic criteria, there are no records to suggest a systematic or concerted study of these differences. Instead, these records appear to represent the groups of people Egyptian citizens saw either as part of their daily life or came into contact with via specific events such as trade or war.

Available texts from the ancient Greek and Roman period predominantly discuss foreign lands and its people from a geographic or political perspective [2]. That being said, several works written during this period nonetheless make relevant contributions to the field of human variability. During the 7th century BCE it had become customary for Greek travellers to make written records of their travels. Over time these accumulated into guidebooks intended to ease the uncertainties of travel to foreign lands by merchants or sailors. These guidebooks contained predominantly practical information, but often included geographic distributions and physical and cultural characteristics of the human settlements they encountered. Though the original purpose of these texts was not scientific in nature, they nonetheless imply a thorough investigation of foreign lands and its people. Thus, from a historical perspective, these texts are often considered foundational to the development of ethnographic and anthropological thinking that becomes formalized over 2,000 years later. Second, and somewhat in contrast to the cultural landscape of this period, two notable authors wrote extensively on the subject of foreign human groups. Herodotus lived in Greece

during the 4th century BC and is the first known person to have left a structured and detailed survey of the ancient world and its people and to point out the association between physical characteristics and climates or geographical locations. Pliny the Elder lived in Rome approximately 500 years after Herodotus and is best known for authoring a seven-volume compendium of "Natural History" collated from numerous sources and authors to provide an encyclopedia on every aspect of the natural world known at that time [2]. One of the seven volumes was devoted to descriptions of people, in which Pliny discussed human variation from either a Greek/Roman perspective or with a focus on the unusual: people that are very tall or short, or those displaying impressive strength, speed or even fecundity. A number of "monstrous races" are also mentioned, and these take a decisive slant towards the fantastical: headless, nose-less, mouth-less, huge eared or one-legged races. Though impossible to take at face-value, modern interpretations suggest that many of these descriptions were either exaggerations or misinterpretations of real phenomena [5]. One relevant example from Pliny's work is his description of the Pygmy people, whom he cited to have a standing height of 40 to 70 cm tall (Refer to Chapter 2 for a literature review of the Central African Pygmy people).

1.1.2 The Renaissance

The scholars and artists of the Renaissance displayed great interest in Classical Antiquity and among the many renewed interests was the human body and its proportions (See Chapter 2 for an overview of anthropometry). The study of human variation and research into human body proportions facilitated the establishment of anthropometry as a science which in subsequent centuries takes center stage in the study of human variation. The Renaissance period is described as a turning point for comparative studies. As pointed out by the ethnographer John Howland Rowe, the interest displayed by Renaissance scholars towards ancient Roman and Greek civilizations catalyzed the development of objective methods by which to study human groups and the necessary 'distance perspective' required to study the unfamiliar. In other words, the interest in Classical Antiquity lead to a major perspective shift in the way that other human groups were perceived and studied and served as precedent to the curiosity toward and subsequent study of modern-day "non-Western" groups [3].

The Americas were explored in 1492, with the first descriptions published in 1511 by Pietro Martire d'Anghiera, an Italian scholar and follower of the Renaissance movement. Martire served as a Royal chronicler in the Spanish court which allowed him to oversee what information was to be collected by the explorers as well as garner public interest in the study of human populations [3].

1.1.3 The Enlightenment (1600-1800s)

The intellectual Enlightenment movement (scientific method and emphasis on rational thought) began in the late 1600's and lasted until the early 1800s and represents a key period in the history of evolutionary thought and the phenotypic study of humans. Open questions debated by naturalists during this time which related to the study of human variation included: **1)** Is there a single or multiple origin for all humans? **2)** Is human variability fixed (discrete) or fluid? **3)** What underlying factors determine human skin pigmentation?

During this time-period, environmental factors were considered critically important in the determination of health and outward physical features. Though the connection between skin color and climate has been recorded by naturalists since Classical Antiquity, Samuel Stanhope Smith (1751–1819), an American minister conducted the first formal correlation between skin-color gradients to latitude and sun intensity [6].

It should be noted that "fixity of species" (an evolutionary theory dating back to ancient Greece, also called *Scala naturae* or "The Great Chain of Being") was a popularly held belief during this time. The latter theory represents the natural world as a linear hierarchy, beginning with the simplest life forms at the bottom (plant) and progressively increasing in complexity towards the top (human) [7]. As such, when it came to discussions of the phenotypic variation observed among humans, many naturalists during this time favored 'immutability of races' or the 'unchanging racial hierarchies' theory, including the Swedish botanist Carl Linnaeus (1707-1778). Linnaeus was the first to create a uniform a system of taxonomical nomenclature for living organisms. In regard to human classification, he divided humans into 4 *ordered* groups based on skin color and geography (*Americanus rubescens* -American reddish, *Europaeus albus* -European white, *Asiaticus fuscus* -Asian tawny, *Africanus niger* -African black) which he later supplemented with

details on physical traits, habits, mode of dress and importantly, mental characteristics [6][8]. This peculiar association of mental characteristics with skin-color is thought to have origins to the Humoral theory. To provide some context to Linnaeus' perspective, the Humoral theory (with origins beyond that of Classical Antiquity) is a holistic system of medicine in which health of the mind and body is attained with the proper balance of 4 bodily fluids, or "humors" (yellow bile, black bile, phlegm, blood) [9]. Each of the humors were interconnected to numerous other phenomena, which depending on the timeframe included the elements (fire, water, air, earth), climactic conditions (hot, cold, moist, dry), personality traits, and physical body-types. Though Linnaeus' human group classifications clearly emphasize the connection between the human phenotype and geography via the four known continents [8][6], it remains altogether unclear why skin-color became aligned with mental features typically associated with an over-abundance of one of the humors.

Contrasting the immutability of races was another theory called 'degeneration', as popularized by the French naturalist George-Louis Leclerc, Comte de Buffon (1707–1788). Though Buffon correctly theorized that all humans have a common origin and variation in skin-color reflects the environmental circumstances under which people lived, the degenerative theory stated that humans living in equitable climates (40-50° latitude) possessed both the original and ideal skin-color; while excessive heat or cold produced inferior deviations from the original "white skin and brown hair" phenotype [6].

During the 18th century, skull shape became the second phenotype to be extensively used in the study of human variation and was pioneered by the German physician Johann Friedrich Blumenbach (1752-1840) [10]. Craniometric scientists from this era can be broadly grouped into two opposing schools of thought: the size and shape of the skull represented either 1) immutable racial characteristics informing on intelligence, talent, or moral character *or* 2) broad racial groupings that were highly variable and lacking in mental hierarchies [11]. Blumenbach belonged to the latter; by studying a collection of diverse skulls he distinguished 5 broad human races: Caucasian, Ethiopian, American, Mongolian and Malay [12]. Like Buffon, Blumenbach was a degenerist and believed human races to be fluid entities. Though he refuted the existence of any

measurement to reliably distinguish races, his skull series was nonetheless arranged in order of personal aesthetic preference, and compared to a Caucasian ideal [11].

Another relevant figure, whose views contrasted the degenerist perspective, was the American physician Samuel George Morton (1799-1851) who is best known today for his racial hierarchies using cranial capacities. Morton's work showed that cranial capacity was largest in Caucasians and smallest in Africans which he directly correlated to intellectual ability. Though this work was contested by his peers at the time of publication, Morton's racial hierarchies of intelligence exacerbated racist ideologies [11] [6].

The overt Eurocentric bias and the custom of attributing moral characteristics to human phenotypes during the Enlightenment period undeniably contributed to the rise of scientific racism in the 20th century. Nonetheless, it remains important to make a distinction between the phenomenon of racial discrimination and the process of cataloguing physical differences across human groups. Phenotypic differences (such as skin color or cranial shape) across human groups do not hold a presupposed rank. Thus, assignment of hierarchy to a phenotypic trait is an imposition on an inherently neutral phenomenon of variation that contaminates a concept quite central to the fields of anthropology and biology.

1.1.4 Classical biomarkers

On a parallel front, the turn of the 20th century represents a significant shift in biological perspective and marks the beginning of the classical genetics era. Improved biological understanding and related advancements in methodology allowed for the quantification of physiometric variations (or, biomarkers) found within humans. The first of these biomarkers were antigens, proteins and enzymes of the blood [13][14]. Though the underlying component of heredity was still unknown at this time, the attention shift towards cellular/molecular components in the blood was a significant step towards the discovery of the DNA blueprint.

In 1901, the first red blood cell antigens (now called the ABO blood groups) were discovered by Karl Landsteiner [15]. In 1919, Hirschfeld and Hirschfeld published the first population genetics study using the principles uncovered by Landsteiner. The latter work used a diverse study

population (soldiers stationed on the Macedonian front in WWI) to report distinct proportions of the A and B alleles across several worldwide populations [16]. The characterization of ABO blood groups, in retrospect, marks the discovery of the first genetic locus in humans and was seminal to the fields of population genetics and transfusion medicine. The latter also catalyzed a concerted search effort for additional markers, which over the next 70 years lead to the discovery of over 30 blood cell and soluble markers (antigens and proteins) [14]. In addition to the accumulating repository of blood-related biomarkers, numerous other biological features (PTC tasting, ear wax consistency, color blindness, lactose tolerance, etc.) were identified by the latter half of the 20th century and subsequently applied to human population studies [17]–[22].

It is interesting to note that many of the markers that became most popular in anthropological and evolutionary studies were blood markers that subsequently become associated with disease and natural selection. Two such examples are glucose-6-phosphate dehydrogenase (G6PD) and human leucocyte antigen (HLA). G6PD is an oxidative stress response enzyme in red blood cells associated with the presence of the malaria parasite while HLAs are cell-surface proteins responsible for the regulation of the immune response. Both are notably polymorphic and have a distinct worldwide distribution: numerous G6PD deficiency variants exist at higher proportions in tropic and sub-tropic regions while diversity levels of HLA variants also follow broad geographic patterns and are implicated in numerous human disease processes [22]–[29].

By the mid-1950's, accumulating data on biomarker frequencies began to be compiled for numerous worldwide human polymorphisms [13][21][30]–[32]. Among these was the seminal publication by Cavalli-Sforza in 1963, marking the first construction of a human evolutionary tree via 5 classical biomarker frequencies (ABO, MNS, Rh, Diego and Duffy) across 15 human populations and the first application of Principal Component Analysis (PCA) to worldwide marker frequencies [33][13][30].

1.1.5 Genomics

Enabled by further improvements in biological understanding and technology, the late 1970s and 1980's represent an important transition period in the field of human population genetics. Arguably among the most notable accomplishments is the polymerase chain reaction (PCR) first described

in 1987 [34], though the first approach for studying genetic polymorphisms preceded PCR by almost a decade.

Genetic polymorphism refers to a location in the genome that differs across individuals or populations. In the broadest sense, the genetic unit of change refers to one or more base-pair changes within a stretch of DNAs either on a "nucleotide" or "structural" level [35]. Structural changes usually refer to processes that restructure chromosomal segments, for example, insertions, deletions, copy number variations (CNVs) or rearrangements of DNA segments. The latter is catalyzed by a variety of biological processes like mutations, segregation errors, transposons (short and long interspersed nuclear elements/SINES and LINES) or disease. Though structural-level changes have been reported across human populations both from the standpoint of normal phenotypic variation and disease susceptibility [36]–[39], they are emerging topics and will not be discussed in the present review. Nucleotide-level changes are usually discussed from the perspective of single nucleotide polymorphism (SNP); defined as a genomic region comprised of a single nucleotide whose identity varies across different individuals or populations. In the present time, SNPs are the most common nucleotide-level polymorphisms used to study human variation, however it was not until the early 2000's that large-scale mapping of candidate SNPs became technologically feasible [40][22].

Prior to the wide-spread use of SNPs, two other types of DNA variation were used: restriction fragment length polymorphisms (RFLPs) and microsatellites. RFLPs (first reported in 1978) harness the DNA-cleaving ability of restriction enzymes (endonucleases) at specific nucleotide sequences. For a given strand of DNA, an RFLP may be detected if the product of the endonuclease cleavage differs across individuals [13]. In 1987, RFLP analysis was used to create the first human evolutionary tree using mitochondrial DNA (mtDNA) variation patterns across five worldwide populations [41]. Notably, the latter publication by *Cann et al* provided the first piece of genetic evidence to support the "out of Africa hypothesis" (since confirmed using both classical markers and genomic analyses). The advent of PCR in the late 1980's increased the range of possibilities, including the systematic mapping of the human genome. This enabled the discovery of polymorphisms at the nucleotide level, the first of which were microsatellites or short tandem repeats (STRs) identified in 1989 [42]. Microsatellites hold an important place in history due to

their use in the Human Genome Project and the first population genetics studies. Though SNPs are presently the most frequently used polymorphism, certain applications continue to benefit from the higher mutation rate and polymorphic status of the microsatellite [43][13].

The study of genome-wide features became possible at the turn of the century. The latter was facilitated both by the accumulation of high-density maps of genetic polymorphisms and the development of key conceptual and mathematical principles during the first half of the 20th century (the Hardy Weinberg equilibrium, linkage disequilibrium, the inbreeding coefficient, and the logarithm of odds score) [44]–[50]. In 1999, stretches of homozygosity were first reported in human genomes using microsatellite markers [51]. In contrast to the polymorphisms described above, a run of homozygosity is a measure of similarity (rather than difference) and refers to the DNA segments on homologous chromosomes (as opposed to a single DNA sequence). See Chapter 3 for a review of this topic and the associated bioinformatics software developed as part of this PhD thesis.

The first large-scale study of human genetic variation was published in 2002, using 377 microsatellites from >1,000 individuals (from 52 populations) spanning 5 major worldwide regions [52]. It is important to highlight that large-scale studies like those of *Rosenberg et* were made possible via two international data collection initiatives: the CEPH (as the Centre d'Etude du Polymorphisme Humain) and HGDP (The Human Diversity Genome Project), established in 1984 and 2002, respectively. Collectively, the HGDP-CEPH Human Genome Cell Line Panel (The Diversity Panel) contained DNA samples (LCL cells) from 1,050 individuals from 51 worldwide populations [52]–[54].

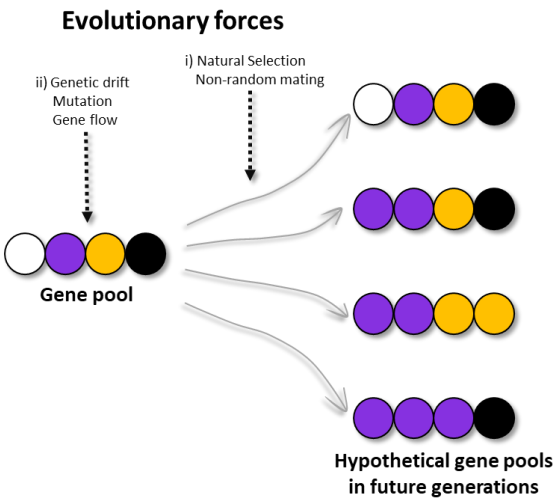
Two final initiatives worth mentioning are the International HapMap project (2002-2008) and The 1000 Genomes Project (2008-2015), both of which to some extent overlap with individuals sampled for the CEPH-HGDP. Though HapMap and 1000 Genomes differed in sequencing technology they shared the goal of characterizing common genetic variation across diverse human populations. At the time of its initialization, the HapMap project characterized common genetic variation across human genomes and served as a key precursor to genome wide association studies (GWAS); the statistical association between genome-wide variation patterns and a phenotypic

outcome (presence or absence of a complex disease or trait) [55]–[57]. In other words, the HapMap project resulted in a genomic database from several worldwide populations which served as the infrastructure for subsequent studies interested in uncovering correlations between phenotypic outcomes and genomic patterns [56][58]–[60]. The 1000 Genomes Project currently represents the largest public catalogue of common human genetic variation collected to date (>2,000 individuals from 26 populations) [61] and accordingly is the dataset used in Chapter 3 of this thesis.

1.2 Evolution

As described in the previous section, the progression of evolution and human diversity studies have historically existed as separate entities until the mid 1900's. Though the historical progression of these connected topics is distinct, there are unsurprisingly numerous points of intersection (for example, the fixity of species concept). The 17th and 18th centuries are perhaps best-known for numerous inquiries into the diversity of life on Earth (human and otherwise), with formal theories of evolution rising to popularity in 1858 with Darwin and Wallace's publication on the principles of natural selection [62]. However, much like ideas of human variation, ideas about evolution have existed long before the 19th century, with traces of evolutionary thinking attributed to numerous Greek, Chinese, and Roman philosophers from the 6th century BCE [63].

In contemporary biology, evolution refers to the gradual process of change in the genetic pool of living organisms over many generations. Variation in a given trait is thus an accumulation of heritable change. This process and the resulting biological diversity are predominantly mediated by evolutionary mechanisms that influence either the size or composition of a gene pool (additionally, recombination during sexual reproduction can also impart heritable change to the gene pool of subsequent generations). There are a total of 5 mechanisms by which evolution takes place: natural selection, non-random mating, mutation, gene flow and genetic drift [64]. The latter mechanisms can be split into two broad categories: *i) selective processes*, or those that change the transmission probability of existing gene pool variants and *ii) stochastic processes* that influence the extent of variation present in the gene pool (see diagram).



Natural selection and non-random mating were the first two evolutionary mechanisms defined by Darwin and Wallace in 1858 [62]. Both forces are selective, in that individuals experience differential rates of reproductive success depending on their underlying genetics. Under *natural selection*, an individual's probability of survival (and reproduction) is considered in the context of their surrounding environment. The better-suited an individual is to their local habitat, the more likely it is they will reach adulthood and produce offspring. On the other hand, *non-random mating* (also called sexual selection) occurs when a given phenotype is considered more attractive by the opposite-sex which influences an individual's probability of reproductive success. Non-random mating can follow assortative or non-assortative patterns; the pairing of individuals with either similar or dissimilar characteristics, respectively. In humans, these are complex and multifaceted processes given that selected-for traits may be both phenotypic (height, weight, pigmentation, major histocompatibility complex, etc.) or socio-cultural (geography, ancestry, age, level of education, religion, behaviour, etc.) [65]–[68].

There are three stochastic processes under category ii). *Genetic drift* is the random fluctuations in allele frequency, which over time has the potential to significantly change the frequency of a given allele. In other words, genetic drift occurs because of statistical sampling error: the genetic variants passed down to the next generation are not representative of the current gene pool. Genetic drift reduces genetic diversity: the smaller the population on which genetic drift is acting, the more pronounced the loss of genetic variation. There are two mechanisms by which genetic drift can occur: Bottleneck or Founder effect. A population bottleneck is a sharp reduction in population size due to events external to an individual's underlying genetics (such as war or natural disaster) while founder effect occurs when a random sample of individuals is separated from the original larger population (geography). *Genetic mutation*, in an evolutionary context refers to a random change in germ-line DNA that is passed onto the offspring. In the specific context of the environment, this mutation has the potential to impart neutral, positive, or negative consequences

on the reproductive success of that individual. Thus, genetic mutations have the potential to affect both size and composition of the subsequent gene pool. *Gene flow* or gene migration is the movement of genetic material into or out of a population. Depending on the direction of gene flow and the genetic make-up of the populations involved, genetic diversity of either group can increase, decrease or remain unchanged [64] [69].

Both Chapters 2 and 3 are an exploration of human variation which is by definition the downstream product of numerous evolutionary processes. Chapter 2 catalogs a series of complex anthropometric traits in Central African Pygmy populations thought to have evolved, at least in part, in response to a tropical rainforest environment (natural selection). Chapter 3 describes a method to infer ancestral 'relatedness' signals in human populations which arise as a consequence of limited gene pools (genetic drift) and non-random mating practices (endogamy, consanguinity). The study of genomic variation patterns (or lack thereof as in Chapter 3) informs on the molecular and genetic mechanisms responsible for human traits, assists in the reconstruction of ancestral demographics and improves the collective understanding of how past evolutionary processes shape human populations.

Chapter 2 Phenotypic profile of Central African Pygmies and non-Pygmies in Cameroon

2.1 Literature review

2.1.1 Brief history of anthropometry

Anthropometry, the measurement of size, shape and proportion of the human body is the oldest method of studying human phenotypic variation. Etymologically, it is derived from the Greek *anthropos* and *matron*, meaning human (or more broadly humanity) and to measure. Despite being one of the earliest branches of physical anthropology, anthropometric methods have not changed significantly since the 19th century. Interpretations however, regarding how and why variation in the human form exists has changed dramatically [1].

Prior to the introduction of the imperial and metric systems, anatomical measurements of the human figure were used as rough standards of measurement. The cubit, which is the length of the forearm from elbow to tip of finger, is seen across numerous ancient populations, though the exact size definition varies. The royal cubit is thought to be the earliest standard measure, originating in Egypt around ~3,000 BC [2]. Other anatomical measurements were also heavily used, some of which are still around today in a more standardized form: foot length, the width of the hand, the fathom (arm-span), or thumb breadth –predecessor for todays 'inch' [2][3].

Among the first systematic studies of the human body were of the head. The study of the skull, or craniometry is commonly attributed to the Ancient Greek physician Hippocrates (4th century BC) who described different skull forms and their variations [4]. Renaissance artists were highly interested in proportions of the human body. Of particular note are Leonardo da Vinci and Albrecht Durer who systematically measured and recorded both head and facial components [4]. In 1529, Durer published a manuscript on crania measurements, which represents the first known attempt to apply anthropometry to aesthetics. Da Vinci's drawing of the *Vitruvian man* depicting a male figure enclosed within a circle and a square is a well-known image in today's popular culture, but in the 15th century, it marked the re-discovery of a seminal Roman manuscript, *De Architectura*, written 1,500 years prior to the Renaissance period. In 21 BC, Vitruvius wrote a compendium of architectural knowledge throughout which he drew on the parallels between architecture and the

human body; positing that architectural design should follow the natural and “archetypal” source of proportion: the human body [5].

The first known study approaching human body proportion from a quantitative perspective dates back to a manuscript titled *Anthropometrica* written by Elsholtz in 1654, where the author discusses the relationship between disease and the human body [3][6].

Measurements relating to the head have been of particular interest throughout human history and cranial measurement continued to be an active topic of study for several centuries after the Renaissance [4]. Physiognomy, the practice of correlating facial features to certain personality traits originated in ancient Greece and became popular in the 16-18th centuries despite having no scientific basis. The field of craniometry continued to advance and in the 18th-19th centuries was used to compare humans to other animals and later to ancient hominid remains. Unlike physiognomy, craniometry remains an important sub-category of morphometrics. Broadly, the study of the human body and in particular the face and head have a disreputable record through much of human history; commonly misused to propagate racist, classist and sexist ideologies [3], [7]–[10] (refer to the introductory chapter for a more in depth discussion of this topic). A recent example occurred during the first half of the 20th century surrounding World War II, where anthropometric studies conducted by German Nazis were used to classify humans into Aryan/non-Aryan categories [5].

In contrast, by the 18th century, standing height was seen as an important determinant of health and strength, and was routinely used and recorded by the military [6][11][12]. Taller individuals were thought to be of better stock, however, the determinants of height itself were not yet understood. By the start of the 19th century, there was great interest in measuring human body shape and size in the context of describing of both modern and past populations. Studies relating to variation and inheritance of human standing height began in the late 19th century and laid the groundwork for the subsequent birth of the quantitative genetics field. Of note, some of the principal researchers involved were F. Galton, K. Pearson and R.A Fisher, all of whom also made fundamental contributions to the field of statistics [13]–[16].

The 19th century was also broadly important for the fields of physical anthropology, paleoanthropology and evolutionary biology. In particular, the fossil discovery of the Neanderthal man in 1859 set in motion a long and heated series of debates surrounding the common ancestry of humans and apes, the evolutionary time-line of humans, and prompted many additional comparative studies between apes, humans and ancient remains [17]–[19].

Correlations between human stature and environmental conditions were first observed in the early 20th century. Comparative studies of United States immigrants and their American born children found distinct differences in standing height, which suggested an environmental role and plasticity of the human body [20] (see section 1.1.3 – Secular changes in standing height). From the 1960s and onwards the scale of anthropometric studies increased dramatically. Numerous anthropometric longitudinal studies were conducted globally in an attempt to elucidate the processes contributing to human growth patterns and worldwide phenotypic variation, often involving extension of tenants already known from studies of plants and animals [21]–[24].
Modern anthropometry is used in public health surveillance, assessment of disease risk, nutrition and paleonanthropology, both in analysis of ancient hominin remains as well as reconstruction of population histories using demographic and economic records. Uses for anthropometry also extend to disciplines like exercise science, economics, forensic investigation, design applications and ergonomics [25].

2.1.2 Selective pressures on human body size and shape

Body size and proportion is central to the biology of all living organisms as it determines ecological opportunity, relationship to the environment, locomotion, physiological requirements, behaviour, and fitness characteristics [26].

In the course of our evolutionary history, humans moved into different temporal regions of the Earth and experienced new environments, resulting in numerous gradual adaptations pertaining to changes in body size, form and proportion. This process is not unique to humans; it has been known for well over two centuries that body size and proportion of many animals, including humans, are systematically related to climate [27]–[33].

Patterns of body shape variation are generally described via two ecogeographic principles, first formally outlined by Bergmann in 1847 and Allan in 1877, even though these variations were observed much earlier than the 19th century. *Bergmann's rule* posits that individuals in colder climates tend to be larger as compared to those from warmer climates [34] while *Allen's rule* states that protruding body parts and extremities are relatively shorter in individuals in colder vs warmer climates [35]. The above rules summarize empirical observations between phenotypic variation and the environment and are best understood as generalizations. Though they are statistically valid, there are numerous exceptions to these rules [36]–[38].

Thus, variations in body size and shape across different temporal regions is thought to be a result of thermoregulatory adaptation. Both *Bergmann's* and *Allan's* rules describe adaptations to thermal stress and the phenotype that accompanies changes in the ratio of body surface area to body mass [39]–[41]**.** Heavier-set, wider, shorter-limbed bodies have a smaller body surface area to body mass, which decreases heat dissipation and thus allows better adaptation to colder climates. Likewise, narrower longer-limbed bodies have a larger surface area to body mass, which increases heat dissipation and thus is better suited to tropical climates [29][40] [42]–[48].

The above patterns are corroborated in worldwide populations. For example, circumpolar populations like the Inuit, Aleut and Sami tend to be heavier, have shorter limbs and broader trunks [48]. Similar trends are also seen in populations in higher altitudes [45]. In contrast, Sub-Saharans and Australian Aborigines populations have proportionally larger extremities [49][50]. Interestingly, ethnic-specific skeletal patterns often persist after several generations following immigration from a different country.

Body proportions and standing height in humans are highly variable both between and within populations. There is considerable inter- and intra-population variation in body shape and size. It has been observed however, that short statured people tend to be concentrated near the equator [45], but in general standing height is less associated with climate than are body and limb breadths. Worldwide, variation in standing height, body breadth and body mass across populations is ~10%, 25% and 50%, respectively [29]. Body breadth in particular has been shown to have a strong correlation with latitude, much larger than for height [39]. It has also been noted that independence

of stature, trunk length generally gets longer in colder climates [29]. Skull shape to a certain extent is also predictive of ethnicity [51], and examination of eye sockets, nasal apertures, nasal bridges, jaw shape as well as tooth size, shape and spacing can be used to estimate overall facial structure and predict racial affinity. A number of cranial and facial features also appear to follow a temporal distribution, though this topic is out of scope for the present review [45].

Limb length changes occurs predominantly as a result of changes in distal limb bones. In other words, arm and leg length tends to change due to changes in the length of the radius or the tibia, respectively. Intralimb bone proportions are commonly described in terms of two ratios: the brachial index represents a ratio of the distal and proximal bones of the arm, and the crural index represents the same ratio but for leg bones [52][53]. Limbs dissipate heat, with distal regions doing so at a faster rate [41]. Consistent with thermoregulatory principles, the distal bones of both upper and lower body limbs tend to be proportionally longer in warmer climates than in colder climates [47]. Thus people living in cold and/or high-altitude climates tend to have shorter distal limb lengths (presumably to lessen the surface area that loses heat most effectively), and the associated brachial and crural indices are smaller [48][54][36][55]. Interestingly distal bones of the hands and feet (to a lesser extent) also tend to be shorter in populations from cold climates [46].

As it currently stands, the underlying reasons for the patterns that Bergmann's and Allan's rules describe are not well defined. The relationship between genetics, environment and body size are by no means simple. Though traits reflecting thermoregulatory adaptations in ancestral human groups are undeniably heritable, trait plasticity and immediate environmental effects (during a single generation) also play a non-trivial role and can create similar phenotypes [56]–[58] [59].

2.1.3 Trait plasticity and the environment

Several mammalian skeletal features exhibit a high degree of phenotypic plasticity, that is, the potential of an organism to modify a phenotype in response to environmental stimuli within the confines of their genomic profile [60]. Numerous studies have demonstrated shortened limb lengths in mammals housed in *cold temperatures*. In particular, a 2008 study showed distinct differences between outbred mice raised at either at 7 °C or 27 °C. Though body mass between the two remained the same, the ears, tails and limbs of the mice in the warm climate were significantly

longer [57]. There is also significant evidence to suggest that *exercise* has the capacity to increase limb length in mammals [61][62]. Interestingly, a study in 2010 tested the effect of combining cold temperature and exercise; they showed that groups of mice housed in either cold or hot environments had similar distal limb lengths *if* they were also exposed to exercise. In contrast, mice housed in cool temperatures without exercise exposure had significantly shorter limb lengths than their exercising counterparts [61]–[63]. *Poor nutrition or exposure to disease* during key developmental periods can also negatively affect final limb length and standing height [39]. Specifically, tibia length is a commonly used metric for nutritional availability and exposure to disease or stress during childhood [64].

2.1.4 Secular changes in standing height

In the last hundred years, many countries worldwide have experienced un-even increases in standing height and relative leg length. This is broadly attributed to secular changes: industrialization, urbanization and the resulting improvements in nutrition and healthcare [65][16][66]. In 2016, the NCD Risk Factor Collaboration published a comprehensive summary of available standing heights from worldwide population born between 1896 and 1996. Though there has been a worldwide increase in height over the past century, the increase is not uniform across countries or between sexes within the same country, though the latter generally follows a similar trend. As of 1996, the tallest country-wide average is in the Netherlands for men and in Latvia for women; 182.5 cm (5 feet 11.9 inches) and 169.8 cm (5 feet 6.7 inches), respectively. In contrast, the shortest country averages are held by Timor-Leste (Southeast Asia) at 158 cm (~5 feet 2 inches) for men and Guatemala at 149.4 cm (~4 feet 11 inches) for women [67].

Over the past century, the largest gains were made by South Korean women and Iranian men at ~20 cm and ~17 cm, respectively. Both sexes in Japan and Greenland also experienced among the largest increases in average height. On the other hand, numerous African and South Asian countries made some of the smallest gains. For example, women in Nimibia, Niger, and Bangladesh experienced just ~2 cm increase in height, while men in South Africa, Pakistan and Marshall Islands (Oceania) increased even less (0.5 cm). Though the majority of countries have experienced relatively steady gains over the past century, this has not been the case for countries like Niger, Rwanda, Sierra Leona and Uganda, where standing height averages increased until the

early 1960's but have since either remained the same or decreased. For a full summary of standing height trends in the last century, see [67] and an interactive version of the studies figures/visuals [www.ncdrisc.org].

Body proportions and standing height are highly variable both between and within populations. As such, it is important to consider that country-wide averages are not necessarily representative of the regional or ethnicity-specific groups of a given country. For example, it is not apparent from the country-wide averages that both the largest and smallest average standing heights are found on the African continent. The Nuer and Dinka people from parts of Sudan and Ethiopia are known for exceptionally tall stature, where men's average heights were reported to be 185 cm (6 feet 0.8 inches) and 181 cm (5 feet, 11.3 inches), respectively, from a 1963 study [68]. This is significantly higher than either Sudan's or Ethiopia's current average of ~166 cm (5 feet 5 inches) for men [67]. Likewise, a study conducted in Bosnia and Herzegovina on young men (born in 2000) found that some cities in the southwestern regions of the country reach considerably higher values than the country-wide male average of 181.7cm (5 feet 11.5 inches), with some of the highest city averages exceeding 185 cm (6 feet 0.7 inch) [69].

The shortest people are the Efe Pygmies in the Democratic Republic of Congo, Efe men and women have average respective heights of just ~145 cm (4 feet, 9 inches) and ~136 cm (4 feet, 5.5 inches) – over 20 cm shorter than the country-wide average. In contrast the country-wide average height for men and women is 167.4 cm (5 feet 6 inches) and 157.6 cm (5 feet 2 inches), respectively [70][71].

2.1.5 Body size epidemiology

Both height and relative leg length have been linked to a variety of non-communicable disorders [16][72][49][73]. For example, taller stature is associated with numerous cancers [74][75] while shorter stature is instead associated with increased risk of type 2 diabetes, chronic heart disease, atherosclerosis and osteoarthritis, the latter possibly a result of ancient selective pressures for shorter bones in colder climates [76][77]. Relative leg length has also been found to be medically relevant. Individuals who have relatively shorter legs were found to have similar risks as those

with short stature, as well as increased risk for obesity and liver disfunction [77]–[80]. The underlying reasons behind these connections are not well understood.

2.1.6 Genetics of body size

Body proportion and size are highly complex traits, representing the cumulative effects of interactions between genetic, epigenetic and environmental factors. Standing height in particular is a composite phenotype resulting from the additive size of the head, torso and leg length, each of which in turn may have different levels of plasticity mediated by differential growth patterns [81]–[83]. Though temporal/behavioural/socioeconomic factors have the potential to impart visible changes to body size during the lifetime of one individual (and thus confound heritable patterns) it nevertheless remains clear that innate distinctions in body proportions across worldwide populations do exist. For example, ethnicity-specific differences in distal bone lengths are readily apparent during fetal development and in newborns [84][52][39]. Additionally, modern tropical populations tend to have relatively longer limbs relative to populations in cold climates, despite having lower socioeconomic status [39][83].

To date, studies have found that standing height follows an additive polygenic model with a roughly normal distribution [16][81][86][87]. Height is one of the more heritable human phenotypes known to date. Family and twin studies have determined that ~80% of standing height variation can be attributed to genetics [87][81].

The past decade of genome-wide association studies (GWAS) and, most recently, the GIANT consortium report of predominantly European and Asian populations has uncovered 697 independent genetic sequence variants associated with height, explaining ~20% of the variability in standing height observed across individuals [88]. In 2017, a follow-up study in European populations determined that up to 27.4% of heritability in height can be explained by rare, low-frequency and common height associated variants [89]. Other work reports much higher heritability estimates. A study in 2010 found that 45% of variance in standing height across ~3,000 individuals could be explained by considering ~4,000 common genome-wide SNPs [90]. Several years later, work by the same group showed that GWAS exome data imputed to the

1000 Genomes Project reference panel explained 56% of variance in standing height across ~40,000 individuals [91].

Currently, the majority of known sequence variants associated with height are outside of coding regions [81] [91]. However, many of the variants were found in proximity to genes coding for extracellular matrix components, chondrocyte biology, skeletal development as well as proteases, transcription factors and cell-cycle modulators [89][92]. Though the majority of height-associated variants discovered so far are common, a 2017 study uncovered a number of rare and low-frequency coding variants associated with adult height [93]. The variants discovered so far tend to have small individual effects. It is highly likely that there are many more additional variants "lost" in the genome [16][92][82].

Despite the effort put into understanding the genetics of human body size, there is relatively little information available. Genetic variants are typically of small effect size, tend to be for standing height only (as opposed to other body proportions) and predominantly based on European populations [16][49].

2.1.7 Central African Pygmies

Central African hunter-gatherer populations, traditionally referred to as "Pygmies," exist at the lower extreme of human body size variation.

The original meaning of the word "Pygmy" emphasizes size and etymologically derives from the Greek *pugon*, meaning "length of a forearm". The first use of "Pygmy" is seen during the 8th century in Homer's *Iliad* where the Pygmies are a mythical tribe of people in a war with cranes. The term was popularized in the late 1800s by Western explorers when they used this term to describe the short-statured rainforest people they encountered on their travels [94]. Thus the term Pygmy is deeply rooted in Western society, and exclusively refers to Central African populations; though other short-statured populations exist in South America and Southeast Asia. Pygmies are the largest and most diverse populations of hunter-gatherers remaining in the world. Total numbers are estimated to be between 250,000-350,000 individuals. Today, the term "Pygmy" functions as an umbrella term referring to 20 ethnolinguistically distinct groups residing in the Congo Basin.

These groups are on average short-statured and linguistically, socially and economically distinct from their neighbouring populations. Pygmy groups themselves are distinct genetically, socially, and geographically; individuals from a given group are unlikely to know about other Pygmy populations unrelated to their own [94].

2.1.7.1 A note on naming convention

There has been a long-standing argument as to whether the term "Pygmy" is an appropriate term to use in reference to these populations. A variety of arguments exists against the use of the word. First and foremost, it artificially creates the illusion of Pygmies belonging to one ethnic group, where in reality there are over 20 Central African Pygmy groups that are diverse in both language and culture. Second, it puts emphasis on physical features which is not shared by all Pygmy groups, and thus is not adequately representative. Though most Pygmy groups do have a significantly lower average stature than their neighbours, there is no discontinuity in height between Pygmy and non-Pygmy groups. Third, some researchers argue that the use of "Pygmy" is inappropriate because it is a colonial and Western term. Lastly, depending on the location or context it is used in, "Pygmy" can be potentially derogatory [95].

Alternatives such as "hunter-gatherers", "foragers", "forest-foragers", "forest-farmers", or "autochones" have been proposed, however, many of these are similarly fraught with issues of misrepresentation, bias and non-specificity. Not all Pygmy groups follow an exclusively hunter-gatherer lifestyle or live in forests. Many Pygmy peoples now are sedentary villagers and are no longer full-time hunter-gatherers. Even if they do hunt and forage, many Pygmy groups also practice agriculture for at least some of the year. Some researchers criticize the representation of a people by their mode of subsistence, arguing that this once again introduces a Western bias by using an economic focal point to define a people. Terms like 'autochones', or 'indigenous' have also been proposed [94]. However since the majority of black Africans are indigenous to the African continent these terms are somewhat meaningless. Some advocate for the use of local terms for Pygmy groups, though these can also have negative connotations within non-Pygmy communities. There is also no consensus on appropriate nomenclature by the people themselves; numerous Pygmy populations refer to themselves as "Pygmies", while others do not.

As it stands, a universally agreeable term does not exist; “Pygmy” and non “Pygmy” terminology persists in literature with the understanding that these are both umbrella terms representing numerous distinct ethnic groups.

2.1.7.2 Existing hypotheses on Pygmy short stature and body size

Small body size in humans has evolved independently in at least three regions: Central Africa, Southeast Asia and South America. Thus, the assumption is that small body size likely conferred a direct or indirect fitness benefit to ancestral populations. With few exceptions, small body size is generally associated with modern populations that have traditionally subsisted on hunting and gathering activities in tropical rainforests [96].

There are 4 classical and non-mutually exclusive hypotheses for the evolution of the Central African pygmy phenotype, related to adaptive processes in responses to various aspect of hunter-gatherer lifestyles in tropical rainforests: 1) *Reduced caloric requirements*: rainforests are sometimes referred to as “ecological islands” where food supply is restrictive depending on season. Thus, small body size could allow for a selective advantage by reducing caloric requirements of smaller individuals [97][98]. 2*) Improved thermoregulation*: a smaller body size may confer an advantage by producing less internal heat during exercise [99]–[101]. 3) *Mobility advantage*: The longer the limbs, the worse the mobility in dense rainforests, so a smaller body size could confer mobility-related advantages [98] and 4) *Early onset reproduction/Pathogen exposure*: tropical rainforests are high pathogen environments with high mortality rates; in order to maximize fitness under these conditions, early cessation of growth, or maturation at the expense of overall size would in theory facilitate earlier age at first reproduction [96][102]. No one theory sufficiently explains the observed phenotype. Thus combined effects are highly likely [94].

2.1.7.3 A brief history

2.1.7.3.1 Ancestral Pygmies and non-Pygmies

The relationship between Central African Pygmies and non-Pygmies is thought to go back to the Bantu expansion; a series of migrations from east Nigeria and west Cameroon to the Central African belt estimated to have occurred ~5,000-3,000 years ago (ya) eventually spreading to the majority of the continent [103][104]. The Bantu expansion is considered a key timepoint in the

demographic and cultural history of Africa, facilitating the transmission of culture, technology, language and likely, genes [99][106][107]. Of particular interest, the Bantu expansion coincides with the spread of agriculture and metallurgy [107][105]. The Bantu people have also left a remarkably large impact on language; today, the Bantu language family predominates across Central, East and Southern Africa [108][104][109]. As ancestral Bantu farmers migrated into new territories, they encountered the ancestors of modern-day Pygmy [110]. The Bantu migration has had a fundamental effect on the language, culture and demography of ancestral Pygmy populations, though the exact timeline and demography of the origin of their relationship are unknown.

Archeological, linguistic and genetic evidence suggests that ancestral Pygmy groups and Bantu farmers have been in contact for at least several thousand years [110]–[112]. Archeological evidence from Gabon suggests a coexistence of hunter-gatherers and sedentary populations by 4,500 ya; however it is unclear whether these populations were ancestral to modern Pygmy and non-Pygmy populations [113][100]. The split among Western Pygmy groups is estimated at ~3,000 ya, which coincides with the expansion of Bantu migrants into the Congo Basin (2-5,000 ya). Thus, the Bantu expansion is hypothesized to have fragmented an initially homogenous Pygmy population by introducing new constraints on mobility that increased isolation and subsequent genetic differentiation among Pygmy groups [113][114]. Additionally, Pygmy groups have undergone a major language shift sometime in the last 3,000 years. Though an original Pygmy language is thought to have existed, the Pygmy language is impossible to reconstruct due to the high degree of linguistic integration with neighbouring groups [112]. Today, all languages spoken by Pygmy groups are related to Bantu non-Pygmy populations, however not necessarily to the one they currently live beside [112][107][115]. Such a major linguistic shift strongly suggests extensive contact between Pygmy and non-Pygmy groups [116]. Interestingly, it has been noted that Mbuti Pygmy groups tend to collectively speak with similar intonations, regardless of spoken language or geographical location [70].

It is striking to note that despite occupying the same space, speaking the same language and most recently, having similar modes of subsistence (as more Pygmies take up farming), boundaries between Pygmy and non-Pygmy groups are maintained. *Neither language nor mode of*

subsistence, the standard for the majority of known worldwide populations, *marks an ethnic or cultural boundary for the Pygmies* [94].

2.1.7.3.2 Two broad Pygmy groups

Central African Pygmy groups can be broadly sub-divided into geographically East and West groups. East Central African Pygmy groups include the Aka, Sua, Efe, Batwa (and others) and reside across Rwanda, Uganda and Eastern Democratic Republic of the Congo (DRC) number at ~30,000 individuals. The West Central African Pygmy groups such as the Bezan, Baka, Bongo, Kola, Koya and Twa reside in Cameroon, Central African Republic, Congo and the Western DRC, numbering at ~55,000 individuals [117][118]. East and West Central African Pygmies are estimated to have diverged approximately ~25,000 ya. Many phenotypic similarities across East and West Pygmy groups exist, however East Central African Pygmies tend to be smaller. Recent genetic studies suggest that the phenotypic similarity between East and West Pygmies is at least in part due to convergent evolution; in other words the exposure to similar environmental constraints rather than inherited from a common ancestor [118]–[120]. See Genetics section 2.1.7.6.

2.1.7.3.3 Complex sociocultural factors between Pygmies and non-Pygmies

All Pygmy and non-Pygmy groups share a long history of socioeconomical and ecological inter-dependence to various degrees.

Traditionally, most Central African Pygmy groups were semi-nomadic hunters, gatherers and foragers, and engaged in the trade of goods and services with agriculturalist non-Pygmy groups. Pygmies and non-Pygmies are economically and socially co-dependent; this includes trade of goods, exchange of work, and participation in social and ritual events. For example, Pygmies are the primary providers of forest products like honey and meat, and are commonly hired as musicians, dancers, or hired labor. They are also sought out for medicinal services, ritual practices and craftsmanship. Pygmies in turn rely on the non-Pygmies for a variety of iron tools, cassava and work (which often leads to unbalanced power dynamics) [94][112].

Modern relationships between Pygmies and non-Pygmies are complex, often paradoxical and highly variable depending on the group. Strong and long-standing cultural barriers exist against intermarriage [110][121], though marriage is sometimes permitted between Pygmy women and non-Pygmy men [103]. The power dynamic between the two groups is generally unequal; Pygmies are often marginalized and discriminated against. On the other hand, non-Pygmies generally believe that Pygmies are "quasi-supernatural", exist at the border of the human world, are able to control the supernatural powers of the forest, and are often regarded with both fear and admiration. Similarly, Pygmies describe the farmers as powerful sorcerers [94][112].

Paradoxically, the participation of the 'outside' group is often required for key cultural processes. For example, the bride-price in Pygmy marriage ceremonies are iron tools which are procured by non-Pygmies, while meat hunted by the Pygmies is crucial for non-Pygmy ceremonies, and the presence of Pygmies is crucial during various non-Pygmy rituals [94][112].

2.1.7.4 Anthropometry

Phenotypic data on Central African Pygmies accrued over the last 50 years predominantly focus on differences in height, weight, BMI, and, to a lesser extent, cranial morphology and leg, torso and arm lengths, in comparison to their non-Pygmy neighbours.

2.1.7.4.1 Cranial features

Broadly, Pygmy cranial and facial characteristics on average involve a greater orbit depth, a lower nasal height, and a lower positioned face, larger nostrils, smaller lips and smaller ears as compared to other African populations [94]. A study from 1962 observed that though East Pygmy infants were born smaller than non-Pygmy infants, their heads were proportionally less reduced than their overall size [122]. Generally, facial and cranial differences between the two groups are not obvious on an individual level [94].

2.1.7.4.2 Stature and body proportions

Studies in the beginning of the 1900s defined Pygmy populations using specific stature cut offs, however considering there is no discontinuity between the heights of individual Pygmies and non-Pygmies, these stature cut-offs are now considered relatively arbitrary [94].

Body weight and BMI are most commonly reported as reduced in Pygmy populations. Though levels of nutrition between Pygmy and non-Pygmy groups are often different, they are not sufficient to explain the observed differences in [71][123]. Additionally, not all research reports a unanimously lower Pygmy BMI [124].

Other than reduction in stature and weight, several common generalities have been cited regarding Central African Pygmies, however when taken together are somewhat contradictory. A study from the 1970's reported sitting heights, limb lengths and shoulder breadth averages in East Ituri Pygmies equaled ~ 87% of averages reported for non-Pygmy Africans. Chest girth and hip breadth averages, on the other hand, were more similar (~95-98% of non-Pygmy Africans), however another study reported an unusually small pelvis size in Pygmies [100].

A study of the East Mbuti Pygmies reported shorter legs, longer arms, wider shoulders and larger heads in comparison to their non-Pygmy counterparts [70]. A meta-analysis in the late 70's compared four groups of geographically close Pygmy and non-Pygmy males, three from the East and one from West Central Africa. They uncovered several trends: East Mbuti Pygmies had shoulder breadths, forearm lengths, overall arm length and spina ilica height measurements 2 or more standard deviations smaller than the geographically close non-Pygmies. The East Cwa and Twa Pygmies appeared to be the most proportionally reduced, with all measurements approximately 1 standard deviation less than their closest non-Pygmy counterparts. Lastly, the West Babinga/Binga Pygmies were found to have sitting heights and shoulder breadth 2 and 1.5 standard deviations smaller than their non-Pygmy neighbours, respectively [125]. An anthropometric profile of West Baka Pygmies, similarly reported a larger sitting height. **Finally, a recently published text summarizing currently known information on Central African Pygmies and non-Pygmies lists the commonly cited differences to date: longer arms, a longer cubit, a reduction of leg length and larger sitting to standing height and forearm to trunk ratios** [94].

It is clear from the work outlined above that Pygmies do not present the same pattern of differentiation but do have in common a general reduction in size. The fact that different Pygmy

groups present with anthropometric distinctions is not altogether surprising, considering their high degree of genetic heterogeneity and isolation from each other [94][125].

2.1.7.4.3 Allometry

Allometry is a somewhat convoluted term due to its numerous meanings during the past century. In the modern context it most often refers to the study of the relative growth of body parts in a given population of organisms, or to associations between size and shape in a given organism or across organisms and the respective physiological functions [126]–[128].

One of the main questions in the context of Central African Pygmies is whether scaling relationships between a given body segment and the rest of the body differs in Pygmies in contrast to their non-Pygmy neighbours. In other words, do Pygmy and non-Pygmy individuals *of the same height* present with different body proportions? Functionally, elucidation of these patterns has the potential to uncover underlying changes in shared developmental and molecular mechanisms.

Another related consideration is the possibility that shorter individuals in any human population may have different proportions than that of their taller counterparts. Indeed, it has been previously reported that shorter individuals in numerous human populations have larger sitting heights, shorter legs, longer arms and wider shoulders in comparison to taller individuals [70].

Over the past century, much attention has been given to the relatively small body size of Central African Pygmies, but surprisingly little to the question of explicit allometric comparisons [124]. In the 1930's, Huxley hypothesized that Pygmy body size and limb proportions are likely allometric correlates of their overall size reduction in the past [129][70], however specific morphological distinctions between Pygmies and neighbouring non-Pygmy groups are not a common subject of allometric investigation [124]. A study in 1977 reported that Konda/Twa/Cwa Pygmies are proportionally reduced in contrast to their non-Pygmy neighbours [125].

In 1996, a preliminary study of the Efe, another group of Eastern Pygmies, reported that even though adult Pygmies and non-Pygmies differ in their final size, their allometric trajectories are concordant, with the possible exception of hip breadth presumably constrained by child-birth. In

other words, neither of the studies found any differences in the underlying morphological patterns in the two groups. Other reports state that East Konda Pygmies exhibit more proportional reductions than the Binga (also East) or Mbuti (West) Pygmy groups. Other than the latter, there appears to be almost a complete absence of allometric studies of West Central African Pygmies, and whether they represent isometric correlates of their neighbours is unknown [124].

2.1.7.4.4 Secular changes

East Pygmies from the Ituri forest have been reported to have especially thin arms, legs and calves, possibly suggestive of a decrease in relative muscle mass proportional to their decrease in weight [100]. Similarly arm circumference and tricep skin-folds have been reported to be smaller in West Pygmies from Cameroon vs non-Pygmies [71][94]. Since these measures represent nutritional proxies, these disparities may be reflective of lower socioeconomic status of the Pygmies. A 2011 paper reports a lack of any secular increases in standing height in West Pygmies from 1911 to 2006, which was in contrast to neighbouring non-Pygmy populations who experienced significant increases in standing height from 1943 to 2006 [130]**.**

2.1.7.4.5 Growth and tooth eruption

Growth and malnutrition studies in Pygmies are difficult to conduct because the age of children is typically unknown [94]. Several hypotheses prior to 2015 have suggested that the short stature of Central African Pygmies could be explained by a lack of pubertal growth spurt [131]. However, a high-resolution longitudinal study of Baka growth patterns published in 2015 proved this to be untrue [117]: Baka babies are born the same size at birth, and reach maturity at a similar age in comparison to non-Pygmy African populations. The difference appears to be in the growth experienced in the first 2 years of life. Baka infants grow at a slower rate in the first 2 years of life and are smaller than other non-Pygmy African infants from 3 years and beyond. This is in contrast to some Sue and Efe Eastern Pygmy infants, who are significantly smaller at birth.

On the other hand, tooth eruption, a common proxy for growth and development is markedly different in Baka individuals: tooth eruption of all tooth classes occurs earlier in Baka children than in any other known human population [132]. Differences in dental size are also apparent in

Pygmy and non-Pygmy groups. Southeast Cameroonian Baka Pygmy adults were found to have smaller anterior dentition but larger molars than their non-Pygmy counterparts [133][134].

2.1.7.5 Physiometry

There is limited physiological data on Central African Pygmy populations, though cardiometabolic and endocrinological factors have been studied to a limited degree.

2.1.7.5.1 Cardiometabolic factors and lipoproteins

Traditional foraging societies are commonly cited as having lower blood pressure that either doesn't increase with age or does so more slowly [125][137][138]. This trend is partially reflected in the literature for Central African Pygmies: i) either no change or a mild increase in blood pressure with increased age in Central African Pygmies and ii) lower incidence of cardiovascular disease as compared to their non-Pygmy counterparts or other worldwide populations [137]–[142]. Interestingly, a 2012 study conducted in Cameroonian groups compared atherosclerotic risk by means of pulse wave velocity between Pygmies practicing predominantly hunting and gathering versus those that have migrated to a semi-urban area. The study found that semi-urban Pygmies experienced similar increases in blood pressure with age as non-Pygmies while traditional Pygmies had the lowest pulse wave velocity values and thus the lowest risks for atherosclerosis [123]. Though Pygmies living a more traditional lifestyle appear to have more favorable cardiometabolic profiles and lower rates of hypertension, it is nonetheless reported in 5-15% of Pygmy subjects, possibly reflective of increasingly sedentary lifestyles and/or availability of salt, tobacco and alcohol in the given communities [100][94][142]. A recent study found that Pygmies that fell into the hypertensive range had much lower urinary sodium excretion levels (representing dietary intake of salt) in contrast to hypertensive non-Pygmies [140][143].

Numerous studies report very low serum cholesterol values in Pygmies [142]. Some studies have also reported lower LDL cholesterol and higher HDL cholesterol in traditionally subsisting Pygmies than in their non-Pygmy counterparts. Other studies in Pygmies have instead reported lower HDL and higher triglyceride values [125][137][140][146][147].

2.1.7.5.2 Endocrinological factors

Other than differential free fatty acid profiles, endocrinological studies since the early 1970s have reported differences in glucose homeostasis, insulin secretion and perturbations in the insulin-growth factor 1/growth hormone (IGF1-GH) axis in Pygmies as contrasted to non-Pygmies [146]–[152].Central African Pygmies have normal serum levels of GH, but lower levels of IGF-1 and GH binding protein (GHBP) as compared to non-Pygmies [149][154][155]. Though the genetic basis of African Pygmies short stature is unknown, the IGF1-GH axis is suspected to be at least partially responsible.

2.1.7.6 Genetics

Genetic studies have predominantly focussed on reconstructing the evolutionary history of Central African Pygmies and non-Pygmies, elucidating the genetic basis of short stature and/or genetic loci under selection and identifying genetic structural differences in either group.

2.1.7.6.1 Evolutionary history of Central West Pygmies and non-Pygmies

Though there is a record of human presence in the Congo Basin for at least 40,000 years, this geographical area in general has a shortage of demographic, historical and archeological information resulting from the high acidity of the rainforest soil [94][112].

Central African Pygmies and non-Pygmies are assumed to share a common ancestor, and numerous genetic studies conducted in the last decade provide broad estimates of the last common ancestor existing ~ 50-70,000 ya (50-70 Kya) based on microsatellite, targeted resequencing and mitochondrial data and ~30,000-150,000 ya (30-150 Kya) based on exome and whole genome sequence data [118][119][113][154].

These time frames roughly coincide with the major expansions and migrations of human groups within and beyond the African continent estimated to have occurred ~60-80 Kya. Though it is impossible to pinpoint the catalysts underlying these movements, evidence from paleo-climatological and archeological studies provides possible clues which suggest this time period also saw a) the emergence of major technological and cultural advances, and b) the onset of extremely dry conditions in many regions of Africa [119] either of which could have indirectly triggered major movements. To further put the ancestral split timescale into perspective, the “Out

of Africa Migrations" are broadly estimated to have occurred 50-120 Kya [155], thus major demographic changes such as the split between ancestral Pygmy and Non-Pygmy groups very likely overlapped with the dispersal of modern humans throughout the world.

It is important to note that the original environment(s) that eventually resulted in divergent selection processes acting on the ancestral Pygmy and non-Pygmy populations are unknown. In other words, whether ancestral Pygmy populations evolved to be shorter or ancestral non-Pygmy populations evolved to be taller remains a fundamental question in the evolutionary history of Central Africa [153]. Genetic, linguistic and geographical evidence suggests that East and West Central African Pygmies have an ancient origin, with the last common ancestor existing ~25,000 ya. This coincides with the Last Glacial Maximum (19-26.5 Kya), a period in Earth's climactic history involving a vast reduction of Central African tropical forests. The phenotypic similarities across East and West Pygmy groups could be either a result of their shared ancestry, and/or via convergent evolution as a result of small body size conferring a fitness advantage– the later hypothesis more favored by the current data available [118]–[120].

Finally, West Central African Pygmy groups diverged ~ 3,000 ya [156][103][118][113] possibly as a result of Bantu farmers migrating into Pygmy territory and fragmenting a then homogenous Pygmy population [114] who subsequently underwent a period of isolation and genetic differentiation.

2.1.7.6.2 General trends and admixture patterns

From a population genetics perspective, Central African Pygmies exist near the root of African diversity, and along with other African hunter-gatherer populations represent one of the earliest branches in African pre-history [118].

Pygmy groups have a high level of genetic differentiation in contrast to the more homogenous non-Pygmy populations, possibly resulting from their recent expansion into Central Africa (the Bantu expansion). Pygmy groups on the other hand are highly diverse; a given Pygmy group is likely to be more similar to its neighbouring non-Pygmy group rather than another Pygmy group [97][104][114][120]–[122][158][159].

One study reports the statistic for genetic distance (F_{ST}) across Central African Pygmy populations is comparable to the F_{ST} values observed on the continental scale, striking considering the small geographical range in which Central African Pygmies reside [158].

The noted similarity between Pygmy and non-Pygmy groups is a result of asymmetric gene flow across the two groups; a feature present to various degrees in the majority of neighbouring Pygmy and non-Pygmy settlements. For example, Eastern Pygmy groups are generally much less admixed with their non-Pygmy neighbours, with levels of gene flow estimated to be 3-7x times less than admixture in Western Pygmy and non-Pygmy groups; which in part explains why Eastern Pygmies are on average smaller. Thus Eastern Pygmies are arguably one of the most isolated populations in sub- Saharan Africa [118][119][159].

Strict taboos exist surrounding marriage between Pygmies and non-Pygmies in the majority of Central West Africa, however unions between non-Pygmy males and Pygmy females are permitted to varying degrees. That being said, some Western Pygmy groups do not have obvious patterns of male-biased gene flow, as such it is unclear when these social boundaries came into existence [160]. A number of studies report that Pygmy standing height is positively correlated with the proportion of non-Pygmy admixture, thus strongly implicating genetic factors [121][122][155][163]–[165]. A 2017 study reported that Western non-Pygmies show Pygmy ancestry signatures in their HLA region (known to mediate the immune response), which suggests that admixture with ancestral hunter-gatherer populations may have proved advantageous to the Bantu migrants from an evolutionary perspective [164].

Genomic studies have established that asymmetric gene exchange between Pygmy and non-Pygmy groups has been occurring for at least 1,000 years. This however is significantly later than the presumed “initial” meeting between the two groups during the influx of Bantu populations into close proximity of ancestral Pygmy populations 3-6 Kya [160][119][165]. Interestingly, evidence for ancient interaction between the two groups is supported by a study of *Helicobactor pylori* (*H.pylori*), a pathogen that has co-evolved with humans for at least the past 100 Kya [165]. The study provides strong evidence that Baka Pygmies acquired their current subtype of *H.pylori* through contact with non-Pygmy neighbours ~2.5-4 Kya.

All evidence suggests that admixture did not occur immediately after the two groups met, rather interactions progressed in two-stages: a) interactions which have subsequently resulted in the socio-economic interdependencies observed today as well as a complete language shift on the part of the Pygmies and b) intermarriages with strict social boundaries. This is markedly different from other documented interactions between Bantu migrants and African hunter-gatherer groups. For example, the meeting between Bantu migrants and the Khoi-san of the Kalahari Desert 1-2 Kya resulted in immediate cultural and genetic exchanges [160].

2.1.7.6.3 Genetic signatures of body size in Central African Pygmies

The genetic basis of African Pygmies short stature is currently unknown, but the IGF1-GH axis pathway is strongly suspected to be partially responsible [166]. In 2009, RNA from peripheral blood cells from 35 Baka Pygmies and 14 neighbouring non-Pygmy groups was used to quantify genetic expression of the IGF1-GH axis pathway, revealing markedly different gene expression profiles. Growth hormone receptor and growth hormone was 8x and 1.8x reduced respectively, in the Baka Pygmies as compared to the non-Pygmies. Despite this, basal levels of serum growth hormone were comparable (agreeing with previous endocrinological studies) and no differences in the base-pair sequence of the GHR gene was found [152]. Several subsequent attempts to locate genetic variants that correlate to reduced stature or the pygmy phenotype in selected regions of GHR and IGF1 genes have also yielded negative results [120][167]. GWAS during the last decade have searched for the genetic basis of standing height in Central African Pygmies; exposing possible signatures of differential selective pressures acting on GHR and IGF1 genes in Pygmies and non-Pygmies [120][153].

Numerous additional candidate regions have been reported to have differential signals of selective pressure and include loci related to immunity, metabolism, reproduction, wound healing, pituitary function, muscle development, bone synthesis and olfactory/taste sensors; the majority of which have been reported by multiple studies [154][161][163]. Currently, only a small fraction of height associated genomic regions identified in Europeans have been identified in Central African Pygmies and non-Pygmies [163][165][16][122][155].

Other than the loci related to the GHR and IGF1 axis, or those previously associated with height in other populations, there is yet no certainty as to which of the above associations are authentic, which are false positives and how these associations combine into a cohesive biological framework of the pygmy phenotype. Likely the implicated regions have wider functions than currently understood.

In an attempt to distil the genetic elements involved in the determination of standing height, a recent study used trunk and leg length in addition to standing height as a trait variable in their GWAS and admixture association studies. Though some genetic regions associated with all three traits to similar extents, some regions associated with standing height and sitting height only, which suggests those variants only contribute to trunk growth as opposed to long bone lengths. They also reported a high correlation between shorter sitting heights and greater levels of non-Pygmy admixture in males, with much lower associations observed in females, implicating sex-hormone mediated growth patterns [161].

Taken together, currently available data are strongly suggestive of a polygenic model of adaptation; in which reduced body size in Central African Pygmies has occurred as a result of small but concerted allele frequency shifts at many loci [168]. It is highly likely that there have been multiple selective events in the history of both Pygmies and non-Pygmies, encompassing different genetic loci at different time periods throughout the overall process of adaptation. Additionally, it is possible that small body size could have occurred indirectly as a result of selection on traits other than stature such as immunity or metabolism [120].

2.2 Purpose of study

The purpose of the present study is two-fold. First and foremost, this work aims to add to the compendium of available knowledge pertaining to Central West African Pygmies and non-Pygmies. To this end, a comprehensive characterization of anthropometric and physiometric trait differences between the two ethnic groups is provided and placed in context of previously published data. Secondly the effect that standing height imparts on anthropometric traits for both Central West African Pygmies and non-Pygmies is explored. The latter analyses is expected to provide insight on whether Central African Pygmies are proportional reductions of neighbouring non-Pygmy populations and whether anthropometric measures are uniquely affected by standing height variation within each ethnic group.

2.3 Methods

2.3.1 Study participants

Anthropometric, physiometric and family-history data were collected from 379 individuals in May-June of 2016 from four distinct ethnic groups from two broad regions in Cameroon: Central (CC) and Southeast (SEC). Each region contains two ethnically distinct groups, one defined as ‘Pygmy’ and the second neighbouring group as ‘non-Pygmy’. The two populations in CC are the Pygmy Bezan and the non-Pygmy Tikar, while in SEC the populations studied are the Pygmy Baka and the non-Pygmy Nzime (See Figure 2.1).

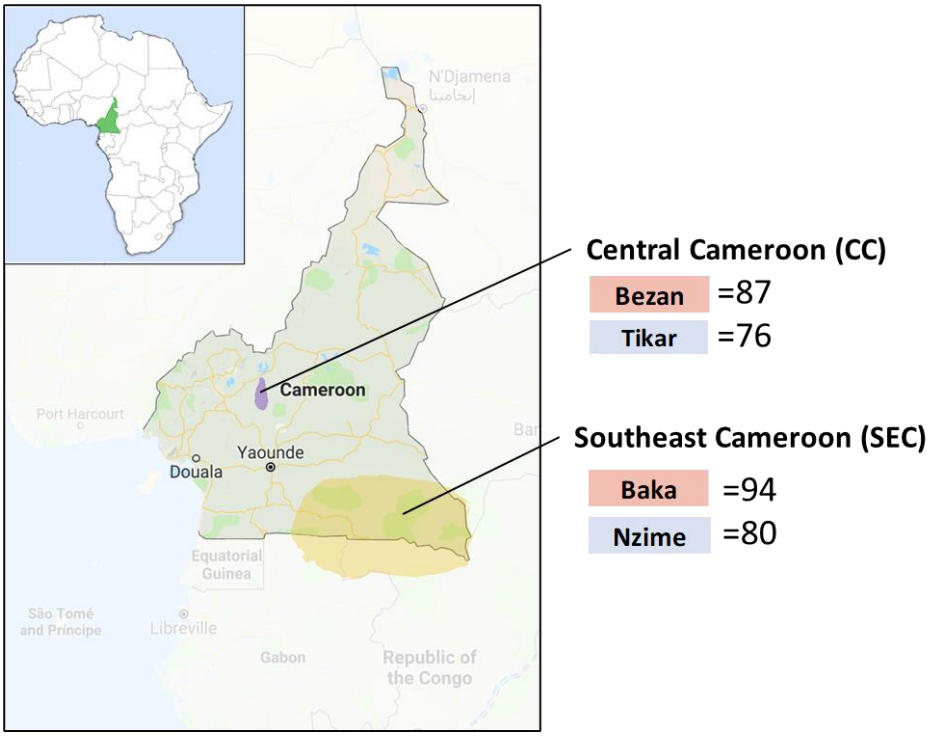


FIGURE 2.1-RANGE AND DISTRIBUTION OF CENTRAL AFRICAN PYGMY AND NON-PYGMY GROUPS STUDIED

Distributions of Central Cameroon (CC) Bezan and Southeast Cameroon (SEC) Baka Pygmy groups are shown in purple and yellow, respectively. Within each region, 2-5 Pygmy/non-Pygmy villages were sampled (See Table 2.4). Obtained sample sizes are displayed on the right for each Pygmy (red) and non-Pygmy (blue) group.
Images are modified from Google Maps and www.freeworldmaps.net

Those participating in the study were of adult age without visible shortening of the spine (a result of advanced age) and, to the best of the individual's knowledge, not pregnant. Appropriate research authorization was obtained from the University of Manitoba Research Ethics Board and the Cameroonian government. Informed verbal consent was collected from all participants and potentially relevant phenotypic information (such as blood pressure, pulse, weight, and blood glucose) was returned to the participants as a hand-written document (in French). Verbal consent was obtained either in French or the local Bantu language translated by a hired interpreter.

2.3.2 Data collection

All participants verbally completed questionnaires relating to ethnicity, place of birth and family history either in French or in the local Bantu language (via an interpreter). In consultation with anthropologists Drs. Alain Froment and Fernando Ramirez-Rozzi, we measured a total of 19 anthropometric and 7 physiometric traits (Figure 2.2), with 21 additional traits subsequently calculated from the measured values (Table 2.1). Trait selection was based on whether traits were high quality proxies for estimating body structure components, minimally invasive to the participant, previously reported in anthropometric studies of similar populations, and practical to collect in a backcountry setting.

Trait		
Type	***Name***	***Formula***
Anthropometry - generic formulas:	Trait : standing height	trait / standing height
	Trait : arm length	trait / arm length
Anthropometry:	Arm span	shoulder breadth + (2 x arm length)
	Femur length	leg length – tibia length
	Humerus length	arm length – cubit length
	Shoulder breadth : hip breadth	shoulder breadth / hip breadth
Anthropometry/ Physiometry:	Body mass index (BMI)	weight / (standing height) ^{2}
	Waist circumference : s.height (WHtR)	waist circumference / (standing height) $^{0.5}$
Physiometry:	LDL cholesterol	Calculated by *CardioChek Plus* point of care device
	Non-HDL cholesterol	
	Total : HDL cholesterol	
	LDL : HDL cholesterol	
	Mean arterial pressure (MAP)	([2 x DBP] + SBP) / 3
	Pulse pressure	SBP - DBP

TABLE 2.1-CALCULATED PHENOTYPIC TRAITS

A total of 21 phenotypic traits were calculated in this study (13 anthropometric, 6 physiometric, 2 relating to both). Two generic formulas used to calculate 9 ratio traits are defined in the first two rows and subsequent rows provide formula details for the remaining 12 traits. *Abbreviations used*: LDL-low-density lipoprotein, HDL-high-density lipoprotein, SBP-systolic blood pressure, DBP-diastolic blood pressure.

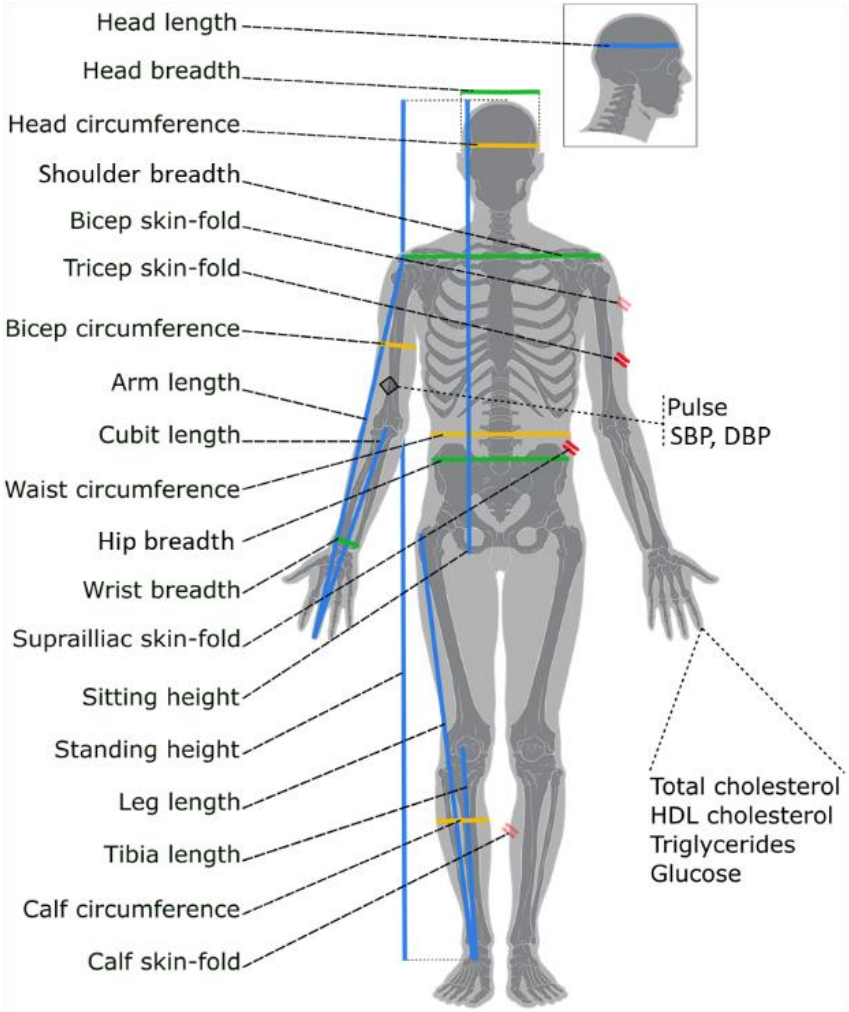


FIGURE 2.2-MEASURED ANTHROPOMETRIC AND PHYSIOMETRIC TRAITS
Anthropometric traits are shown via color coded lines: blue refers to heights or lengths, green for breadths, yellow for circumferences and red for skin-folds. All skin-folds were measured on the subject's left side while blood-pressure, pulse and anthropometric arm and leg measurements were taken on the subject's right side. A total of 19 anthropometric and 7 physiometric traits were measured.
The skeleton image has been modified to fit this work's purpose; the original work is credited to Patrick J. Lynch and C. Carl Jaffe. See here for creative commons license: https://anatomytool.org/content/anterior-and-lateral-view-skeleton-no-labels

Headwear, shoes and heavy articles of clothing were removed prior to measurement. Height, head measurements and bone lengths/widths were measured using a standard anthropometer and circumferences were measured using a plastic, non-elastic measuring tape. Weight and skin-folds were measured using a weight scale and a skinfold caliper, respectively. By convention, all anthropometric measurements were taken on the right side of the body where possible with the exception of skin-folds which were performed on the left side. All measurements were performed

according to standard biometrical procedures [169]. With the subject sitting in a relaxed position, blood pressure of the right arm was measured via an Omron-5 blood-pressure monitor; with the reported blood-pressure representing an average value of three consecutive measurements.

Blood glucose and lipid readings were obtained using test strips (eGLU and Lipid Panel) and analyzed on a CardioChek PLUS point-of-care device from PTS Diagnostics. Participants were asked to fast overnight, and physiometric measurements were collected as close to the early morning as possible. Preceding blood collection, all participants washed their hands in warm water and a fingertip on the non-dominant hand was sterilized with an alcohol swab. Blood was drawn using a single-use lancet device and collected with a glass capillary tube.

Similar to other hunter-gatherer populations, Central African Pygmies do not track their age or associate life history events with calendar years [172][173]. As such, the exact ages of the majority of Pygmy individuals in this study was unavailable. Instead, participants' age was visually assessed and subsequently treated as an ordinal variable with six categories (Table 2.2).

	CC								SEC							
	Bezan (Pygmy)				Tikar (Non-Pygmy)				Baka (Pygmy)				Nzime (Non-Pygmy)			
Age category	Females	%	Males	%	Females	%	Males	%	Females	%	Males	%	Females	%	Males	%
<19-21	17	29.8%	4	13.3%	8	24.2%	0	0.0%	12	19.4%	6	18.8%	3	8.8%	4	8.7%
22-30	15	26.3%	13	43.3%	6	18.2%	3	7.0%	24	38.7%	10	31.3%	11	32.4%	11	23.9%
31-40	12	21.1%	9	30.0%	8	24.2%	4	9.3%	9	14.5%	3	9.4%	11	32.4%	11	23.9%
41-50	3	5.3%	3	10.0%	1	3.0%	11	25.6%	6	9.7%	4	12.5%	3	8.8%	8	17.4%
51-60	4	7.0%	0	0.0%	4	12.1%	12	27.9%	4	6.5%	5	15.6%	3	8.8%	8	17.4%
61-70+	4	7.0%	1	3.3%	3	9.1%	11	25.6%	6	9.7%	4	12.5%	3	8.8%	3	6.5%
NA	2	3.5%	0	0.0%	3	9.1%	2	4.7%	1	1.6%	0	0.0%	0	0.0%	1	2.2%
sample size	*57*		*30*		*33*		*43*		*62*		*32*		*34*		*46*	

TABLE 2.2-AGE DISTRIBUTION ACROSS ETHNICITY- AND SEX- SEPARATED GROUPS
Approximate age is represented by six ordinal categories and an "NA" category, which represents individuals for whom age data was unavailable. Central Cameroon (CC) and Southeast Cameroon (SEC) regions are divided into their respective Pygmy (red) and non-Pygmy (blue) populations, each of whom is further divided by sex (females in green and males in yellow). For each age category, total counts and percentages are provided for each sub-group, where percent (%) is calculated using count of individuals within each age category and overall sub-group sample size.

2.3.3 Quality Control for Individuals

Using the self-reported family history, 42 individuals were retrospectively removed from subsequent analysis after being determined to be relatives, immigrants, or of admixed ethnicity (Table 2.3). Relatives were defined as individuals with first or second-degree relatives already present in our sample; when relatives differed in sex, those representing the minority within that ethnic group were selected. Immigrants were defined as individuals from outside of the four ethnic groups in our study (Bezan, Baka, Tikar or Nzime); and admixed as those with one parent from outside of the ethnic group in question. A total of 337 individuals remained in the study; sample sizes for ethnicity- and sex-separated groups is provided in Table 2.4.

					III) Admixed					
					Individual being removed has 1 parent from:					
			II) Relatives		the neighbouring group			outside group		*Sub-group totals*
Region	Ethnicity	I) Immigrants	#	Type	#	individual's sex	parent's ethnicity	#	individual's sex	
CC	Bezan (Pygmy)	0	3	FS	5	3 M, 2 F	Tikar	1	M	9
	Tikar (Non-Pygmy)	8	4	3 PO, 1 AV	1	M	Bezan	7	2 M, 5 F	20
SEC	Baka (Pygmy)	0	1	FS	1	M	Nzime	0	N/A	2
	Nzime (Non-Pygmy)	2	1	PO	1	M	Baka	7	4 M, 3 F	11
	Sub-group totals	*10*	*9*		*8*			*15*		***42***

Total number of individuals removed from study

23

TABLE 2.3-INDIVIDUALS REMOVED AMONG THE FOUR ETHNIC GROUPS

Using self-reported family history, 42 individuals were removed from the dataset: 29 (9+20) and 13 (2+11) from Central and Southeast Cameroon, respectively. The maximum number of individuals removed from any one ethnic group was 20 individuals from the Tikar population. I) Ten individuals were not native to any of the villages in the study and therefore categorized as immigrants. Most immigrants were identified in the Tikar group, and no immigrants were identified in either of the Pygmy groups. II) Nine individuals were removed due to an existing relative already present in the study group. If relatives differed in sex, the individual whose sex represented the minority within that ethnic group was selected. Three subtypes of familial relationships were identified: two groups of fraternal siblings (FS), four parent-offspring (PO) pairs and one avuncular (AV) pair. III) A total of 23 individuals were removed due to admixed ethnicity. Eight individuals had one Pygmy parent and one non-Pygmy parent (labelled 'the neighbouring group'); most of such cases (5 out of 8) were found among the Bezan. Fifteen individuals had one parent who belonged to a group outside those studied; such cases were found predominantly in the non-Pygmy groups. Males and females are marked as M and F, respectively.

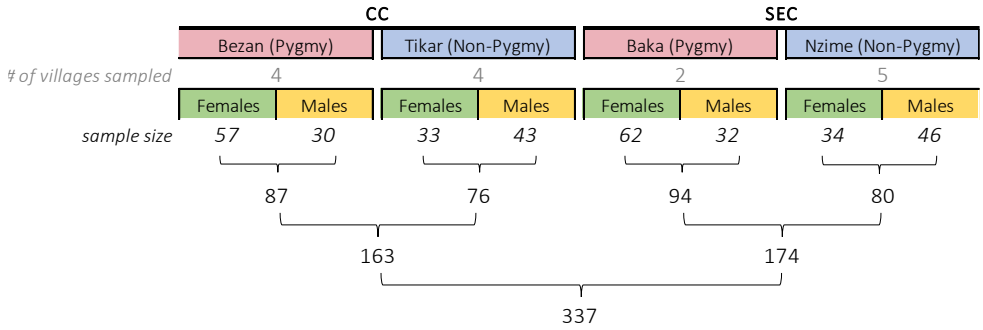


TABLE 2.4-SAMPLE SIZE BREAKDOWN OF RETAINED INDIVIDUALS

Central Cameroon (CC) and Southeast Cameroon (SEC) regions are divided into their respective Pygmy (red) and non-Pygmy (blue) populations, which are then further divided by sex (females in green, males in yellow). Sample size of each sub-group is summarized, including the number of villages sampled per ethnic group.

2.3.4 Quality Control for Measurements

No outliers were removed from the dataset, however a missingness profile for the participants and the measured traits is shown in Figure 2.3. Values for three physiometric variables, HDL cholesterol, triglycerides, and total cholesterol, in some individuals fell either below or above the device's sensitivity ranges which are 15-100 mg/dl, 50-500 mg/dl and 100-400 mg/dl, respectively. To avoid removing these values from our analysis, out of range values below the minimum cut-off were raised to the minimum value while the values above the maximum cut-off were lowered to the maximum value. For the individuals with censored HDL cholesterol, triglyceride, and total cholesterol values, their LDL and non-HDL cholesterol, which were calculated using the above three variables, were set to be missing.

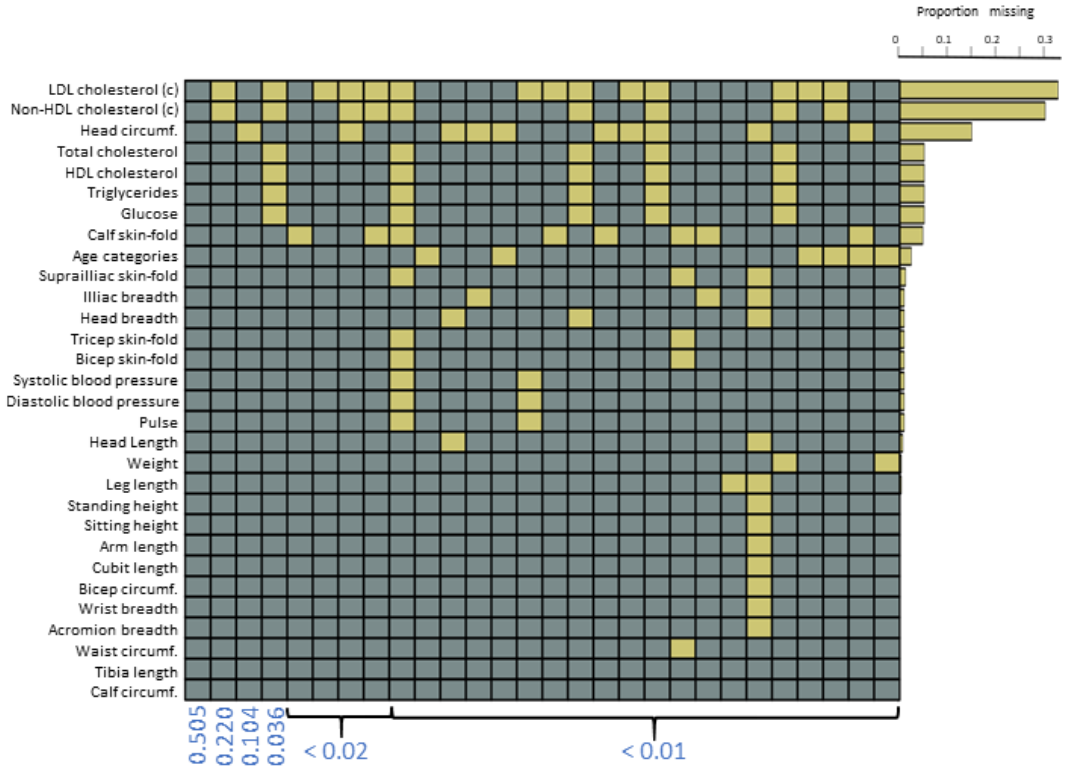


FIGURE 2.3-AGGREGATION PLOT OF MISSING VALUES

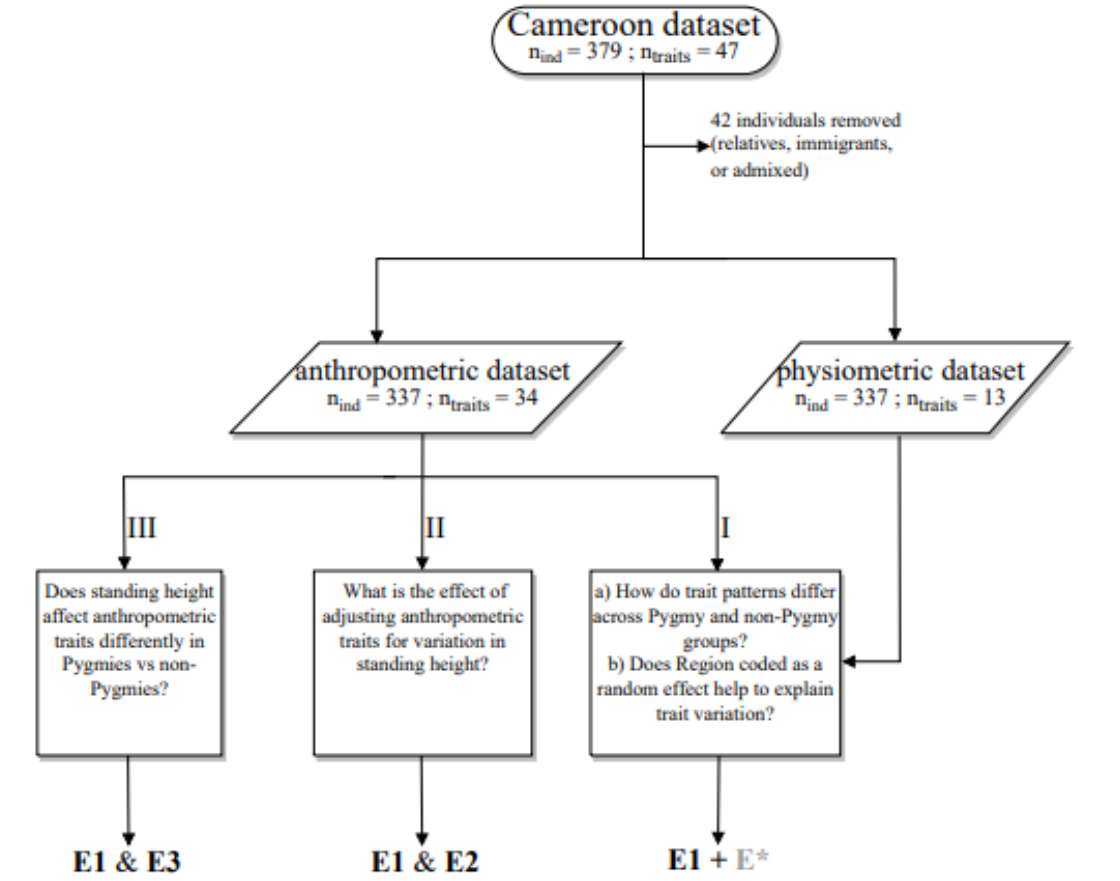
A visualization of all existing combinations of missing (yellow) and non-missing values (gray). **Each row** is a representation of the missingness profile for the phenotypic traits listed on the left-hand side and informs on the proportion of missing entries for each phenotypic trait (yellow barplot on the right). Traits are sorted by proportion of missingness in descending order. For example, the two top-most rows represent missingness profiles for LDL cholesterol and non-HDL cholesterol. Both traits have frequencies of ~0.3, translating to ~30% of participants missing entries for LDL cholesterol and non-HDL cholesterol. Note that traits shown on the left-hand side are measured raw traits, in addition to age data and two calculated values (c): LDL and non-HDL cholesterol. The latter two are calculated from total and HDL cholesterol; both of which were treated as censored values given that occasional participants had readings which fell below/above the limits of the machine (< 100, >500 mg/dl). In such an event, the censored values for total and HDL cholesterol are counted as non-missing, while traits that require either of the censored traits in their calculation are counted as missing.

Each column represents the existing missingness profiles of the participants. In other words, each column represents a specific combination of traits that a participant was either measured for (gray box) or was not measured for (ie missing, in yellow). Columns are sorted by frequency descending to the right. Each vertical column represents one missingness profile and values in blue represents the proportion of individuals which had this profile. For example, the first column is made of gray squares only and occurs at a frequency of 0.505. This translates to: ~51% of all individuals in this study had no missing values for any of the phenotypic traits measured.

2.3.5 Data analysis

All statistical analyses were performed using R version 3.6.2. Anthropometric and physiometric traits collected in this study were analyzed separately. Three linear mixed-effect models (LMM) were used, represented by equations E1, E2 and E3, with the anthropometric/physiometric trait as the dependent continuous variable (Figure 2.4). Each LMM has three to five independent fixed effect variables, depending on the purpose of the regression formula: age, sex, Pygmy status, Pygmy status x standing height (interaction variable) and either standing height or BMI (depending on whether the trait in question is anthropometric or physiometric). In addition, LMMs E1, E2 and E3 include region as a random effect predictor variable and follow the standard random intercept model. Region coded as a random effect predictor variable accounts for variability produced by environmental and genetic factors in effect within a given region. For a given phenotypic trait, the random intercept model assumes a constant slope and a variable y-intercept across Central and Southeast Cameroon, allowing for region-specific trait means [172].

Using LMM E1, E2 or E3 phenotypic trait patterns for a given predictor variable are determined using hypothesis testing, the result of which is interpreted by assessing relevant regression coefficients and 95% confidence intervals. All hypothesis testing follows a similar procedure and tests the following null hypothesis (H_0): *trait means are the same between Pygmy and non-Pygmy groups*. Refer to Figure 2.4 for visual summary of the research questions and pertinent LMMs.



Equations:

E*: $Y = \alpha + \beta_A*Age + \beta_S*Sex + \beta_{PS}*PS + TTSC + \epsilon$

E1: $Y = \alpha + \beta_A*Age + \beta_S*Sex + \beta_{PS}*PS + TTSC + Region + \epsilon$

E2: $Y = \alpha + \beta_A*Age + \beta_S*Sex + \beta_{PS}*PS + Region + \epsilon$

E3: $Y = \alpha + \beta_A*Age + \beta_S*Sex + \beta_{PS}*PS + \beta_{Height}*Height + \beta_{PSH}*(PS \times Height) + Region + \epsilon$

Legend:

n_{ind}= # of individuals in dataset ; n_{traits}= # of phenotypic traits in dataset

A = age (6 categories) ; S = sex (male vs female) ; PS = Pygmy status (Pygmy vs non-Pygmy)

TTSC = trait-type specific coefficient (anthropometric vs physiometric).

For anthropometric traits TTSC = $\beta_{Height}*Height$ and for physiometric traits: TTSC = $\beta_{BMI}*BMI$

PSH = Pygmy status x standing height

FIGURE 2.4-VISUAL REPRESENTATION OF METHODS; RESEARCH QUESTIONS AND ASSOCIATED REGRESSION EQUATIONS

Three research questions (I, II, and III) are framed in the context of three linear mixed models (E1, E1& E2, and E3, respectively). The anthropometric dataset is used across all 3 research questions, however the phenotypic dataset is used in research question I only. Note: LMM E* is included here for completion but is not part of formal analyses pertinent to the research questions. It is used exclusively for AIC calculation in Appendix Table 1 and Appendix Table 2.

Lastly, R^2 statistics (goodness of fit) are determined for LMM E1 and E3. Broadly, R^2 is a measure of linear relationship between a predictor and response variable and lies between 0 and 1; 0 representing a regression that fails to explain the observed variance in response variable while a value close to 1 explains the observed variance in response variable. Given the difficulties associated with reporting R^2 in mixed-effect modelling, this study instead used the two R^2 variance metrics proposed by Nakagawa and Schielzeth [173] and implemented in the MuMIn R package [174]. The marginal R^2 (or R^2_M) is associated variance explained by fixed predictor variables, while the conditional R^2 (or R^2_C) is the variance explained by the entire model (fixed plus random effects). Finally, the difference between R^2_C and R^2_M, or ΔR^2 is the proportion of variation explained by Region only. The larger the ΔR^2 for a given phenotypic trait, the more unexplained variation is accounted for with the addition of Region. Aikake information criterion (AIC) statistics are additionally used to obtain quantitative measures of strength for the hypotheses represented by the LMMs [175][176] and presented as supplementary material. AIC is a relative measure of model parsimony approximating the relative distance between the fitted model and the *n* "true" factors which generated the data, attempting to strike a balance between fit and complexity [177]–[180]. Here, AIC statistics are calculated in three comparative situations: **i)** a LMM with or without standing height (E1 vs E2), **ii)** with or without interaction variable Pygmy status x standing height (E1 vs E3), and **iii)** with or without region as a random effect variable (E1 vs E*). All three are conducted on the anthropometric dataset, however only iii) is completed for physiometric data. It should be noted that AIC statistics are not meant to inform on absolute model quality, but rather to provide an additional perspective on the relative goodness of fit [176].

2.3.5.1 Assessing trait differences between Pygmies and non-Pygmies

Anthropometric and physiometric trait differences across Pygmy and non-Pygmy groups were assessed using LMM E1, containing one random effect variable: region and four fixed effect variables: sex, age, Pygmy status and either standing height (for the anthropometric traits) or BMI (for the physiometric traits) (Figure 2.4). Standing height and BMI were identified as confounding variables due to their significant correlation to anthropometric and physiometric measurements, respectively. Changes in BMI correlate to changes in both blood pressure and lipoprotein profile [181]–[186] while standing height is either a direct correlate or a predictor of several body segments. The latter relationship has been known for over a century and has been used for

assembly of incomplete skeletons and estimating total stature from skeletal remains in the fields of forensics and anthropology [126][189]–[193].

Trait differences across Pygmies and non-Pygmies were assessed by extracting the regression coefficient and its 95% confidence intervals (CIs) for the Pygmy status predictor variable. The regression coefficient describes the change in the trait mean across Pygmy status (Pygmy status = 1 for Pygmies and 0 for non-Pygmies) while all other predictor variables are held constant. Interpretation of the CI lead either to a rejection/failure to reject H_0 (*trait means are the same between Pygmy and non-Pygmy groups)*. H_0 was rejected if the resulting 95% CI of the regression coefficient did not overlap zero, and a statistically significant relationship between Pygmy status and trait was concluded. Lastly, to determine which ethnic group measured significantly larger for the trait in question, the direction and magnitude of the regression coefficient was assessed. As an exploratory study, the results were not corrected for multiple testing.

2.3.5.2 Exploring the effect of standing height

2.3.5.2.1 Effect of standing height on anthropometric traits

To demonstrate the effect of standing height as a fixed predictor variable on anthropometric trait patterns across Pygmies and non-Pygmies, anthropometric traits were subjected to two LMM procedures: E1 and E2. Both LMMs include age, sex, Pygmy status, and Region predictor variables, but standing height is only included in E1 (Figure 2.4). Trait pattern differences across Pygmies and non-Pygmies resulting from both E1 and E2 were assessed following the procedure outlined in the previous section, and AIC statistics calculated. Note that R^2 statistics were not calculated for LMM E2.

2.3.5.2.2 Interaction effects- the relationship between standing height variation and ethnicity

It is unknown whether the effects of standing height on a given anthropometric trait differ across Pygmies and non-Pygmies. To determine this, anthropometric traits were subjected to LMM procedure E3 (Figure 2.4). E3 has a total of 6 predictor variables with region as the random effect variable and age, sex, standing height, Pygmy status, and lastly Pygmy status x standing height as the fixed predictor variables. As described previously, relevant regression coefficients and 95% CI were assessed, and AIC and R^2 statistics calculated.

2.4 Results

2.4.1 Sample description

From an original total of 379, 42 individuals were removed from the study (Table 2.3). Of the remaining 337, 163 individuals are from Central Cameroon (CC); 87 Bezan Pygmies and 76 Tikar non-Pygmies. The remaining 174 are from Southeast Cameroon (SEC); 94 Baka Pygmies and 80 Nzime non-Pygmies (Table 2.4). Both of the Pygmy groups had more females than males (M:F ratio 1:1.9) while the non-Pygmy groups had slightly more males than females (M:F ratio 1.3:1). The majority of all individuals in this study were between ~20 and 40 years old, however the CC non-Pygmy Tikar had the largest proportion of individuals > 50 years old compared to the other groups (Table 2.2). Across 337 individuals, 50% have no missing data. See Figure 2.3 for missingness profiles for the study participants and measured traits.

2.4.2 Anthropometry

A total of 34 anthropometric traits were analyzed in this study, 23 of which are original measurements and 11 calculated from the measured values (Table 2.1). Summary statistics for these traits are reported in Table 2.5.

	CC								SEC							
	Bezan (Pygmy)				Tikar (Non-Pygmy)				Baka (Pygmy)				Nzime (Non-Pygmy)			
	Females		Males		Females		Males		Females		Males		Females		Males	
Raw traits	*mean*	*st dev*	*mean*	*st dev*	*mean*	*st dev*	*mean*	*st dev*	*mean*	*st dev*	*mean*	*st dev*	*mean*	*st dev*	*mean*	*st dev*
Standing height	**149.13**	***5.63***	**157.32**	***7.36***	**152.52**	***7.52***	**166.42**	***10.26***	**146.92**	***5.39***	**154.30**	***5.00***	**151.27**	***8.56***	**162.61**	***8.83***
Sitting height	79.02	*2.78*	82.68	*3.39*	79.79	*3.24*	85.64	*4.34*	76.72	*2.79*	81.25	*3.21*	78.59	*4.21*	84.41	*4.34*
Leg length	84.50	*4.20*	88.68	*5.33*	86.84	*5.34*	95.60	*8.01*	83.17	*4.15*	87.36	*2.99*	86.01	*6.22*	92.83	*6.13*
Tibia length	40.05	*2.39*	42.39	*2.48*	41.14	*2.70*	45.53	*3.80*	39.63	*2.48*	41.99	*1.67*	40.90	*3.21*	44.64	*2.95*
Femur length	44.45	*2.87*	46.29	*3.35*	45.70	*3.54*	50.06	*4.55*	43.54	*2.46*	45.31	*1.87*	45.12	*3.62*	48.13	*3.63*
Arm span	167.94	*6.37*	179.68	*9.41*	171.83	*8.95*	190.55	*12.85*	164.56	*9.72*	176.12	*9.36*	168.34	*11.77*	185.23	*11.91*
Arm length	67.39	*2.79*	71.63	*4.15*	68.88	*3.71*	76.22	*5.54*	65.81	*5.70*	69.94	*5.85*	67.45	*5.91*	73.84	*5.84*
Cubit length	39.79	*1.58*	42.56	*2.37*	40.60	*2.20*	45.00	*3.04*	39.85	*3.56*	42.32	*3.64*	40.58	*3.44*	44.10	*3.16*
Humerus length	27.60	*1.54*	29.08	*2.12*	28.28	*1.99*	31.22	*2.73*	25.96	*2.39*	27.63	*2.88*	26.87	*2.78*	29.74	*3.22*
Head circumference	53.17	*1.32*	54.97	*1.45*	53.52	*1.74*	56.14	*1.92*	51.52	*5.79*	52.91	*4.92*	52.53	*5.38*	54.93	*3.77*
Head length	17.98	*0.50*	18.76	*0.56*	18.11	*0.57*	19.18	*0.72*	17.80	*0.53*	18.43	*0.70*	18.01	*0.84*	18.86	*0.99*
Head breadth	14.04	*0.43*	14.58	*0.51*	14.17	*0.46*	15.02	*0.84*	13.95	*0.42*	14.43	*0.54*	14.23	*0.67*	15.03	*0.85*
Shoulder breadth	33.16	*1.72*	36.41	*1.65*	34.07	*2.16*	38.12	*2.36*	32.94	*4.33*	36.23	*5.01*	33.44	*3.86*	37.54	*3.62*
Hip breadth	24.63	*1.28*	25.02	*1.63*	25.18	*1.69*	26.16	*1.83*	23.44	*1.79*	23.97	*2.28*	24.10	*2.11*	24.94	*2.06*
Weight	45.93	*6.40*	51.80	*7.42*	51.75	*11.48*	60.81	*12.46*	42.66	*4.67*	46.09	*5.95*	47.25	*9.99*	54.20	*9.30*
Waist circumference	75.56	*6.58*	74.60	*4.42*	78.81	*8.81*	79.22	*8.16*	74.41	*4.96*	70.54	*4.24*	76.36	*6.62*	74.64	*6.01*
Tricep skin-fold	10.56	*3.93*	6.40	*1.57*	12.54	*5.29*	7.77	*3.21*	8.73	*2.71*	4.79	*1.28*	10.03	*4.35*	5.60	*1.86*
Bicep skin-fold	4.68	*2.89*	3.53	*0.91*	5.46	*3.27*	4.18	*1.60*	3.79	*1.65*	2.81	*0.64*	4.29	*2.48*	3.10	*0.90*
Calf skin-fold	7.96	*2.59*	4.98	*1.29*	9.49	*3.83*	6.61	*2.56*	7.53	*2.31*	4.37	*1.09*	7.75	*3.12*	4.59	*1.40*
Suprailliac skin-fold	5.43	*2.24*	5.13	*1.04*	6.25	*3.09*	6.27	*2.97*	5.73	*2.36*	4.44	*1.09*	6.10	*2.73*	4.70	*1.30*
Wrist breadth	45.96	*2.46*	49.33	*3.18*	47.42	*3.36*	51.90	*3.88*	45.32	*2.71*	49.28	*2.34*	46.69	*3.34*	51.06	*4.64*
Calf circumference	29.76	*2.55*	32.06	*2.34*	31.30	*3.34*	33.65	*3.31*	30.60	*7.26*	31.75	*7.64*	31.35	*6.12*	33.62	*5.41*
Bicep circumference	23.67	*2.06*	24.39	*1.97*	25.51	*3.72*	26.37	*3.13*	23.16	*2.5*	23.71	*3.10*	24.52	*3.35*	25.99	*3.20*
Ratio traits																
Sitting height : s.height	0.53	*0.01*	0.53	*0.01*	0.52	*0.02*	0.52	*0.02*	0.52	*0.01*	0.53	*0.02*	0.52	*0.02*	0.52	*0.02*
Leg length : s.height	0.57	*0.01*	0.56	*0.02*	0.57	*0.01*	0.57	*0.02*	0.57	*0.02*	0.57	*0.01*	0.57	*0.02*	0.57	*0.01*
Tibia length : s.height	0.27	*0.01*	0.27	*0.01*	0.27	*0.01*	0.27	*0.01*	0.27	*0.01*	0.27	*0.01*	0.27	*0.01*	0.27	*0.01*
Arm length : s.height	0.45	*0.01*	0.46	*0.01*	0.45	*0.01*	0.46	*0.01*	0.45	*0.04*	0.45	*0.04*	0.45	*0.03*	0.45	*0.03*
Arm-span : s.height	1.13	*0.03*	1.14	*0.03*	1.13	*0.03*	1.15	*0.03*	1.12	*0.05*	1.14	*0.06*	1.11	*0.06*	1.14	*0.05*
Cubit length : arm length	0.59	*0.01*	0.59	*0.01*	0.59	*0.01*	0.59	*0.01*	0.61	*0.01*	0.61	*0.02*	0.60	*0.02*	0.60	*0.02*
Head length : s.height	0.12	*0.00*	0.12	*0.01*	0.12	*0.01*	0.12	*0.01*	0.12	*0.01*	0.12	*0.01*	0.12	*0.01*	0.12	*0.01*
Shoulder breadth : hip breadth	1.35	*0.07*	1.46	*0.07*	1.36	*0.08*	1.46	*0.08*	1.40	*0.12*	1.51	*0.13*	1.39	*0.11*	1.51	*0.11*
Shoulder breadth : s.height	0.22	*0.01*	0.23	*0.01*	0.22	*0.01*	0.23	*0.01*	0.22	*0.03*	0.23	*0.03*	0.22	*0.02*	0.23	*0.02*
WHtR	0.51	*0.04*	0.47	*0.03*	0.52	*0.05*	0.48	*0.04*	0.51	*0.03*	0.46	*0.02*	0.50	*0.04*	0.46	*0.03*
BMI	20.61	*2.42*	20.89	*2.37*	22.05	*3.55*	21.81	*3.10*	19.77	*1.95*	19.31	*1.93*	20.45	*2.75*	20.41	*2.20*
sample size	*57*		*30*		*33*		*43*		*62*		*32*		*34*		*46*	

TABLE 2.5-ANTHROPOMETRIC TRAIT SUMMARY STATISTICS

Mean and standard deviations (st dev) are listed for the 34 anthropometric traits across ethnicity- and sex-specific subgroups. Central Cameroon (CC) and Southeast Cameroon (SEC) regions are divided into their respective Pygmy (red) and non-Pygmy (blue) populations, each of whom is further divided by sex (females in green, males in yellow). Sample size of each sub-group is printed below each column. Units of measure are in cm for metrics relating to heights, lengths, spans, circumferences, and breadths, in mm for skin-fold thicknesses, and in kg for weights. The order of the anthropometric traits is identical to Table 2.6,Table 2.7 and Table 2.8.

2.4.2.1 Trait comparisons across ethnicity

Anthropometric trait patterns across ethnicity were assessed using LMM E1 (Figure 2.4) with the results summarized in Table 2.6 and Figure 2.5. Thirteen traits were significantly larger in the non-Pygmies compared to the Pygmies, while one trait, the cubit to arm ratio, was significantly larger in the Pygmies compared to the non-Pygmies. Refer to the paragraphs below for additional detail.

Body mass: In alignment with prior work, metrics pertaining to body mass were significantly larger in non-Pygmies (weight, BMI, both the raw and ratio waist circumference metrics (WHtR), bicep circumference and all skinfold thicknesses). Though body mass is an anthropometric measure, its effects on physiometry (such as lipoprotein levels and blood pressure) are well documented [182]. As such body mass metrics are further explored and placed into context of reference ranges in the Physiometry section (2.4.4).

Head metrics: Most head measurements were found to be significantly larger in non-Pygmies, , which contrasts previous reports [94]. Non-Pygmies had larger head circumferences and head breadths, however head length and the associated ratio were not significantly different across ethnicity.

Torso and extremities**:** In contrast to previous studies, this study reports no significant differences in anthropometric traits pertaining to torso, leg length or the associated ratios across Pygmies and non-Pygmies. Though raw arm lengths were likewise similar, both ratios pertaining to the arm showed distinct patterns: Pygmies had larger cubit to arm length ratios, while non-Pygmies instead had proportionally longer humeri in relation to overall arm length. These findings are in partial agreement with previous studies which report either a longer arm, cubit or forearm in Pygmy groups [94]. The data in the present study instead suggests that Pygmies and non-Pygmies may differ in arm proportion, without significant change in overall arm length.

	E1			
	FIXED	FIXED & RANDOM		
Raw traits:	Effect direction for: Pygmy status	R^2M	R^2C	ΔR^2
Sitting height		0.80	0.80	0.00
Leg length		0.89	0.89	0.00
Tibia length		0.82	0.82	0.00
Femur length		0.75	0.75	0.00
Arm length		0.65	0.66	0.00
Arm span		0.78	0.78	0.00
Cubit length		0.61	0.61	0.00
Humerus length		0.55	0.59	0.04
Head circumference		0.22	0.25	0.03
Head length		0.40	0.41	0.00
Head breadth		0.45	0.45	0.00
Shoulder breadth		0.47	0.47	0.00
Hip breadth		0.34	0.43	0.08
Weight		0.60	0.65	0.05
Waist circumference		0.25	0.31	0.06
Tricep skin-fold		0.32	0.44	0.12
Bicep skin-fold		0.12	0.21	0.10
Calf skin-fold		0.25	0.41	0.16
Suprailliac skin-fold		0.08	0.12	0.03
Wrist breadth		0.49	0.49	0.00
Calf circumference		0.20	0.20	0.00
Bicep circumference		0.39	0.41	0.01
Ratio traits:				
Sitting height : s.height		0.25	0.25	0.00
Leg length : s.height		0.13	0.13	0.00
Tibia length: s.height		0.02	0.03	0.01
Arm length : s.height		0.04	0.05	0.01
Arm span : s.height		0.09	0.11	0.02
Cubit length : arm length		0.06	0.20	0.15
Humerus length : arm length		0.06	0.20	0.15
Head length : s.height		0.53	0.53	0.00
Shoulder breadth : hip breadth		0.26	0.33	0.07
Shoulder breadth : s.height		0.07	0.07	0.00
WHtR		0.24	0.31	0.07
BMI		0.16	0.28	0.12

Legend

E1 = trait ~ Sex + ageCategories + PygmyStatus + standing height + (1 | Region)

Based on the 95% CI:

For PygmyStatus predictor:

Non-Pygmy > Pygmy for this trait

Pygmy > Non-Pygmy for this trait

TABLE 2.6-ANTHROPOMETRIC TRAIT PATTERNS RESULTING FROM LMM E1

This table provides a statistical summary of results obtained from a hypothesis test using LMM E1 on the anthropometric traits.

Effect direction is shown for the Pygmy status fixed effect predictor variable, color coded in blue or red. Effect direction is obtained by assessing the 95% CI and coefficient estimate; an effect is deemed present if the pertaining 95% CI does not overlap zero, and directionality is assessed via the coefficient estimate.

Goodness of fit metrics were used to assess variance explained by fixed effects (R^2_M) and by the entire model, ie both fixed and random effects (R^2_C). ΔR^2 is calculated as the difference between R^2_M and R^2_C and thus represents the change in variance explained with the addition of the random effect (Region). Goodness of fit metrics for traits significantly different between Pygmies and non-Pygmies are highlighted by a black box while ΔR^2 equal to zero are grayed out. Refer to Appendix Table 1 for a full list of coefficient estimates, standard errors, t-values and 95% CIs for the Pygmy status predictor variable and random effect metrics.

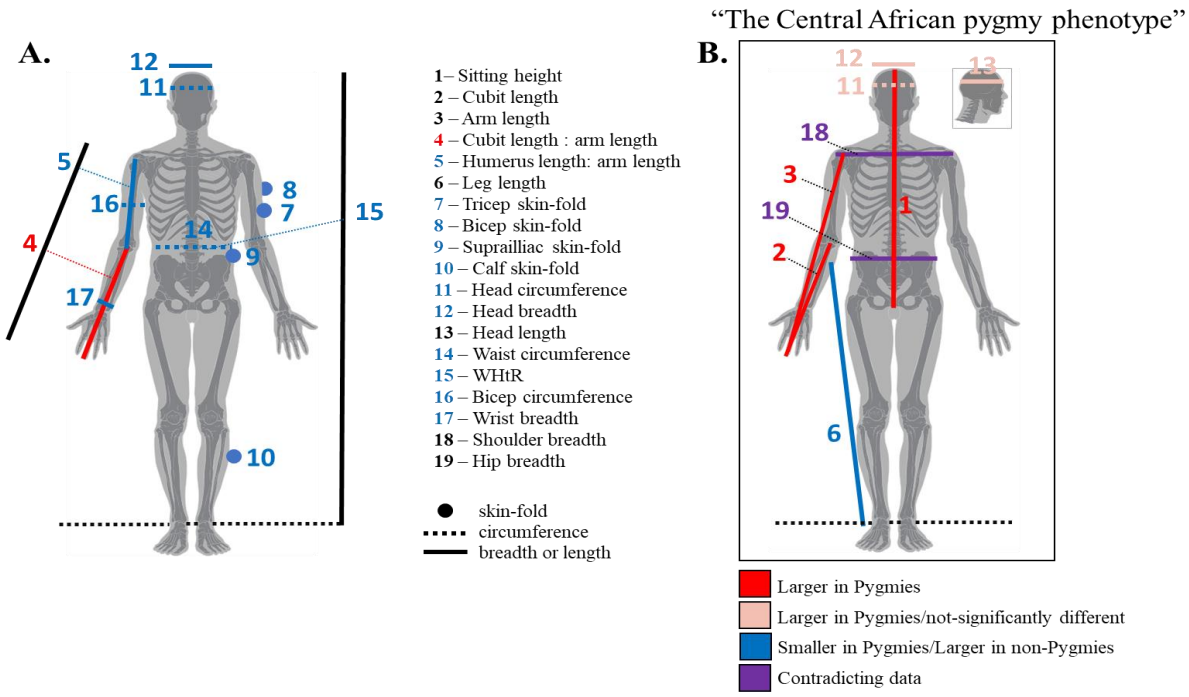


FIGURE 2.5-ANTHROPOMETRIC TRAIT DIFFERENCES ACROSS ETHNICITY

A: A visual summary of results presented in Table 2.6: anthropometric traits significantly different between Pygmies and non-Pygmies obtained in this study (weight and BMI not included). Anthropometric traits larger in non-Pygmies vs Pygmies are in blue, while those larger in Pygmies vs non-Pygmies are in red.

B: Anthropometric traits previously reported to differ between Pygmies and non-Pygmies, summarized from 6 studies spanning ~ 5 decades [94]. As in **A**, red signifies traits that are reportedly larger in Pygmies vs non-Pygmies and blue for the reverse relationship. Pink and purple represent traits with contradicting reports; traits occasionally reported to be larger in Pygmies are shown in pink.

The skeleton image has been modified to fit this work's purpose; the original work is credited to Patrick J. Lynch and C. Carl Jaffe. See here for creative commons license: https://anatomytool.org/content/anterior-and-lateral-view-skeleton-no-labels

2.4.2.2 Goodness of fit metrics for LMM E1

Anthropometric trait marginal (R^2_M) and conditional (R^2_C) variance metrics ranged considerably; with R^2_C values at 0.03 - 0.89 and ΔR^2 (the difference R^2_M and R^2_C) at 0 - 0.16. See Table 2.6 for goodness of fit values and the corresponding Appendix Table 1 for additional detail and AIC statistics. Presented below are goodness of fit measures for a) traits with significant patterns across ethnicity followed by b) general trends relating to R^2_C and ΔR^2 across all anthropometric traits.

a) For the anthropometric traits that showed significant distinctions across Pygmies and non-Pygmies (see section 2.4.2.1), R^2_C ranged widely from 0.12 to 0.65 and ΔR^2 from 0 to 0.16. The highest R^2_C value of 0.65 belonged to the weight measurement; suggesting that 65% of existing variance in weight could be explained via LMM E1. The AIC model selection predominantly preferred the LMM model which included region (E1). One peculiar exception to the latter was noted: though both arm proportions had among the largest proportion of variance explained due to the presence of region, AIC favored the LMM which excluded region (ΔAIC >40).

b) More broadly, sitting height and long bone metrics had **the largest R^2_C values and thus represent the anthropometric traits best explained by LMM E1**. R^2_C ranged from 0.59-0.79 for all raw arm metrics, 0.75-0.89 for raw leg lengths and 0.80 for sitting height. Interestingly, the latter traits also had among the smallest ΔR^2 values, suggesting that including region as a random effect did not increase explained model variance. Additionally, several anthropometric traits had particularly low conditional variance values ($R^2_C < 0.10$) and represent traits that were poorly explained by LMM E1.

Anthropometric traits with the **largest differences** across marginal (R^2_M) and conditional (R^2_C) values, or ΔR^2 ranged from 0.10—0.16 and belonged to BMI, skinfold thicknesses (tricep, bicep or calf) or the two arm ratios (humerus to arm length and cubit to arm length). Broadly speaking, larger ΔR^2 values indicate the inclusion of region as a random effect variable improved the model's goodness of fit which raises the possibility that traits with larger ΔR^2 values may have distinct patterns across Pygmies and non-Pygmies within either CC and/or SEC. The two arm proportions had among the highest ΔR^2 values of 0.15, translating to a 15% increase of the total variance explained with the addition of Region as a random effect. This is of note considering the cubit to

arm length ratio was the only trait in this study found to be significantly larger in Pygmies vs non-Pygmies. That being said, the conditional variance was comparatively low (R^2_C = 0.20) and translates to just 20% of overall variance explained, which suggests the presence of unaccounted-for factors with effects on arm proportions.

2.4.3 Exploring the effect of standing height

2.4.3.1 Effect of standing height on anthropometric traits

The effect of including standing height as a fixed variable was demonstrated by subjecting the anthropometric dataset to two LMM procedures, E1 and E2 (Figure 2.4). By comparing the outcomes from both LMMs, the effect that standing height variation has on the observed trait differences between Pygmy and non-Pygmy groups is thus highlighted (Table 2.7 and Appendix Table 2). Note that LMM E1, was previously used in section 2.4.2.1 to carry out anthropometric trait comparisons across ethnicity. Here, the anthropometric trait patterns obtained from E1 are instead contrasted against those obtained from E2. Both LMMs include age, sex, Pygmy status and Region as predictor variables, but standing height was only included in E1.

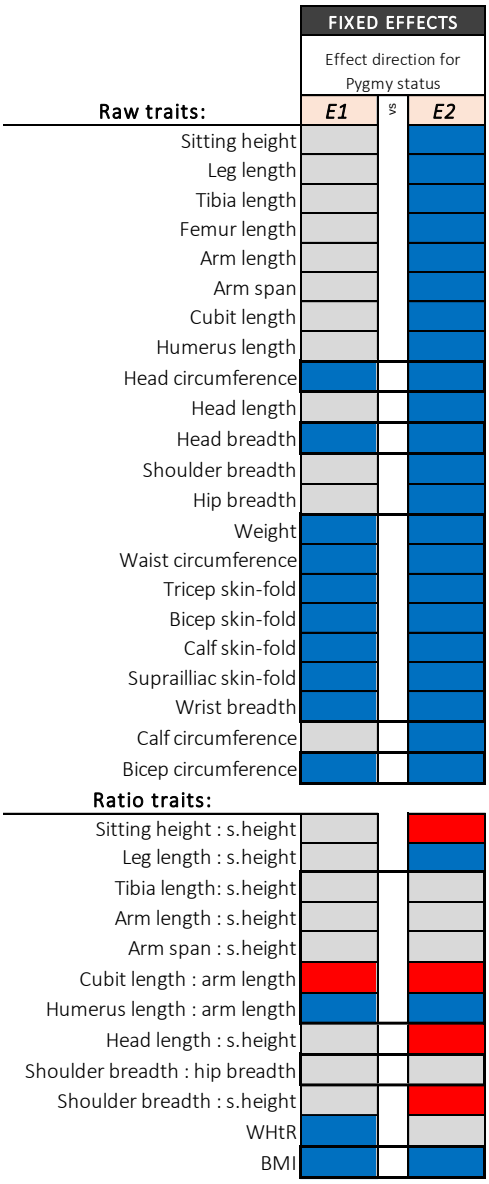
Concordance, or agreement between Pygmy and non-Pygmy trait patterns across height adjusted (E1) vs unadjusted datasets (E2) was 50% for raw traits and 58% for ratio traits. As would be expected, a greater proportion of anthropometric traits showed differences between the two ethnic groups (and had larger effect sizes) with the height unadjusted E2. Using E2, all 22 raw traits were significantly larger in the non-Pygmies as compared to their Pygmy counterparts. However, once standing height was corrected for, the following metrics were no longer perceivably different between the two ethnic groups: shoulder and hip breadth, extremity lengths (except for arm span), sitting height and calf circumference. After correction for standing height, the following ratio traits were no longer significant across Pygmies and non-Pygmies: sitting to standing height, leg length to standing height and head length to standing height.

The height adjusted E1 additionally uncovered one ratio trait pattern not observed in the height-unadjusted LMM. Using E1, waist circumference to standing height ratio was found to be larger in non-Pygmies vs Pygmies, in contrast to the results from E2, which showed no significant difference across ethnicity.

Interestingly, head dimensions displayed inconsistent patterns. While head circumference and breadth remained concordant across E1 and E2 (larger in non-Pygmies), head length and its associated ratio were instead discordant (not significantly different between Pygmies and non-Pygmies in E1). The latter suggests that head length may be more strongly correlated to standing height than either head circumference or breadth in Central African populations.

2.4.3.2 Interaction effects- the relationship between standing height variation and ethnicity

Lastly, LMM E3 was used to assess whether standing height affects anthropometric traits in Pygmies differently than it does in non-Pygmies (Figure 2.4). The interaction term, Pygmy status x standing height, was insignificant for the majority of the anthropometric traits collected in this study (Table 2.8 and Appendix Table 3). However, a significant interaction effect was noted in three traits: bicep circumference, tricep skin-fold and the ratio of head length to standing height (Figure 2.6). Though the slope directionality for bicep circumference and the ratio of head length to standing height in both Pygmies and non-Pygmies followed the same trend, the rate of change (i.e. slope) differed depending on ethnicity. In contrast, tricep skin-fold behaved differently across ethnicities. In the non-Pygmy group, tricep skin-fold size positively correlated with increasing standing height, but in the Pygmy group tricep skin-fold was instead negatively correlated with standing height.



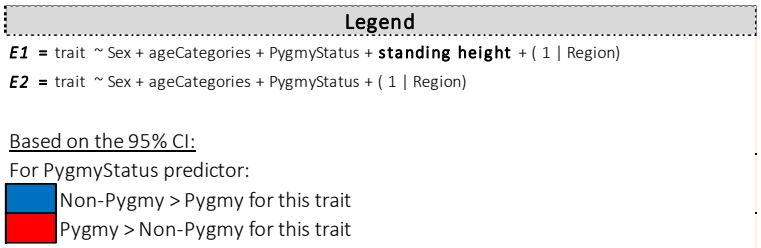


TABLE 2.7-RESULT CONCORDANCE ACROSS LMM E1 AND E2

This table provides a statistical summary of results obtained from two hypothesis test procedures conducted on anthropometric data. Effect direction is shown for the Pygmy status fixed effect predictor variable (color coded in blue or red) for both LMM E1 and E2. Results across LMM E1 and E2 are deemed concordant if effect direction of Pygmy status across both LMM E1 and E2 remained the same and are highlighted by a black rectangle.

Effect direction and directionality are determined as described for Table 2.6. Refer to Appendix Table 2 for the goodness of fit metrics for both LMM E1 and E2 as well as a full list of coefficient estimates, standard errors, t-values and 95% CIs for the Pygmy status predictor variable and random effect metrics.

Raw traits:	E3 FIXED EFFECTS — Effect direction for: Pygmy Status	PygmyStatus *standHeight
Sitting height		
Leg length		
Tibia length		
Femur length		
Arm length		
Arm span		
Cubit length		
Humerus length		
Head circumference		
Head length		
Head breadth		
Shoulder breadth		
Hip breadth		
Weight		
Waist circumference		
Tricep skin-fold		+
Bicep skin-fold		
Calf skin-fold		
Suprailliac skin-fold		
Wrist breadth		
Calf circumference		
Bicep circumference		+
Ratio traits:		
Sitting height : s.height		
Leg length : s.height		
Tibia length: s.height		
Arm length : s.height		
Arm span : s.height		
Cubit length : arm length		
Humerus length : arm length		
Head length : s.height		-
Shoulder breadth : hip breadth		
Shoulder breadth : s.height		
WHtR		
BMI		

Legend

E3 = trait ~ Sex + ageCategories + PygmyStatus + standing height **+ (PygmyStatus*standing height)** + (1 | Region)

Based on the 95% CI:

For PygmyStatus predictor:
- Non-Pygmy > Pygmy for this trait
- Pygmy > Non-Pygmy for this trait

For (PygmyStatus*standing height) predictor:
- \+ LL and UL are both positive
- \- LL and UL are both negative

TABLE 2.8-RESULTING STATISTICS FOR THE INTERACTION TERM IN LMM E3

This table provides a statistical summary of results obtained from the LMM E3 hypothesis test on anthropometric data.

Effect direction is shown for two fixed predictor variables: Pygmy status and the interaction variable Pygmy status x standing height. For each predictor variable, effect direction is obtained by assessing the 95% CI and coefficient estimate; an effect is deemed present if the pertaining 95% CI does not overlap zero, and directionality is assessed via the coefficient estimate.

Effect direction and directionality are determined as described for Table 2.6. Refer to Appendix Table 3 for the goodness of fit metrics for LMM E3 as well as a full list of coefficient estimates, standard errors, t-values and 95% CIs for the Pygmy status predictor variable and random effect metrics.

Note: LL and UL refer to lower and upper limits of the confidence interval (CI), respectively.

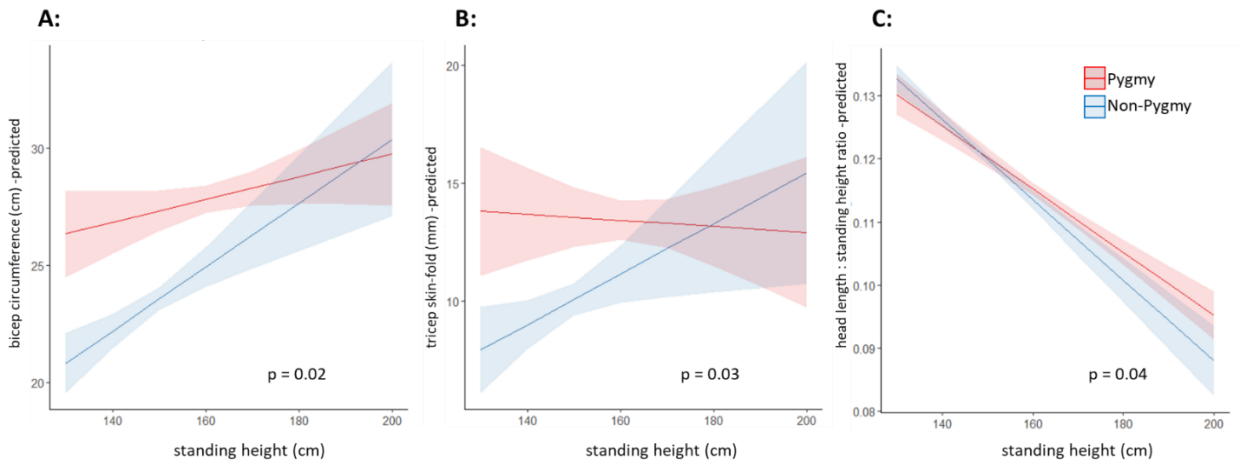


FIGURE 2.6-THREE ANTHROPOMETRIC TRAITS WITH SIGNIFICANT INTERACTION BETWEEN STANDING HEIGHT AND PYGMY STATUS

Marginal effects of the interaction term standing height x Pygmy status is shown for **A**: bicep circumference **B**: tricep skin-fold and **C**: head length : standing height. (see Table 2.8). Predicted trait values (on the y-axis) are plotted using standing height and Pygmy status as the grouping factor (red for Pygmies, blue for non-Pygmies) via the plot_model(type = "pred") function in R.
P value tests whether the LMM which includes an interaction term (E3) better fits the data than a model without an interaction effect (E1). Using a significance level of $p < 0.05$, the data for bicep circumference, tricep skin-fold and head length : standing height ratio has an improved fit with LMM E3 vs E1 (p value = 0.02, 0.03 and 0.04, respectively).
A: Predicted bicep circumference increases with standing height in both Pygmies non-Pygmies, however the rate of increase is faster in non-Pygmies. **B:** Predicted tricep skin-fold increases with standing height in non-Pygmies, and in contrast increases with standing height in Pygmies albeit at a lower rate. **C:** As would be expected, predicted head length to standing height ratio decreases with standing height in both Pygmies and non-Pygmies.

2.4.4 Physiometry

A total of 13 physiometric traits were analyzed in this study, 7 of which are measured and 6 were calculated (Table 2.1). Summary statistics are reported in Table 2.9.

	CC								SEC							
	Bezan (Pygmy)				Tikar (Non-Pygmy)				Baka (Pygmy)				Nzime (Non-Pygmy)			
	Females		Males		Females		Males		Females		Males		Females		Males	
	mean	*st dev*	*mean*	*st dev*	*mean*	*st dev*	*mean*	*st dev*	*mean*	*st dev*	*mean*	*st dev*	*mean*	*st dev*	*mean*	*st dev*
Total cholesterol	136.65	*29.38*	111.50	*18.94*	143.93	*35.56*	129.15	*33.14*	132.70	*44.55*	117.52	*29.80*	133.87	*40.66*	120.60	*28.86*
HDL cholesterol	45.35	*15.28*	31.70	*11.10*	48.89	*17.26*	40.83	*17.71*	44.43	*15.02*	41.97	*21.04*	45.44	*14.73*	41.08	*17.64*
LDL cholesterol	76.57	*20.90*	70.36	*19.09*	77.92	*24.08*	78.60	*28.32*	67.36	*22.19*	48.82	*9.04*	68.65	*24.85*	59.12	*23.52*
Total: HDL cholesterol	3.29	*1.03*	3.53	*0.92*	3.20	*0.98*	3.41	*1.23*	3.16	*0.98*	2.62	*0.76*	3.17	*1.00*	2.95	*0.82*
LDL:HDL cholesterol	1.84	*0.74*	1.95	*0.76*	1.73	*0.73*	1.93	*0.99*	1.59	*0.73*	1.03	*0.40*	1.61	*0.81*	1.40	*0.67*
non-HDL cholesterol	95.52	*26.96*	91.82	*18.41*	98.83	*29.21*	98.42	*29.79*	89.96	*23.74*	81.08	*28.21*	91.27	*25.44*	84.13	*26.35*
SBP	117.35	*16.78*	118.77	*9.60*	117.50	*16.41*	122.71	*15.59*	115.15	*14.90*	117.94	*16.78*	114.82	*14.23*	117.91	*13.82*
DBP	74.23	*10.53*	71.93	*6.83*	74.73	*10.45*	76.53	*10.52*	76.38	*9.11*	75.06	*10.20*	76.44	*9.27*	75.71	*9.90*
Pulse	76.81	*13.25*	73.50	*12.18*	78.52	*12.67*	71.10	*15.20*	81.52	*12.63*	74.74	*13.12*	81.23	*11.94*	77.34	*14.04*
Pulse pressure	43.12	*9.75*	46.83	*7.97*	42.77	*9.49*	46.18	*9.53*	38.77	*9.58*	42.87	*11.83*	38.38	*8.93*	42.20	*9.45*
Mean arterial pressure	88.60	*12.11*	87.54	*6.91*	88.99	*11.94*	91.93	*11.60*	89.31	*10.44*	89.35	*11.49*	89.23	*10.35*	89.78	*10.45*
Triglycerides	98.21	*41.52*	103.40	*47.43*	103.95	*50.37*	100.67	*46.15*	107.20	*40.16*	124.13	*88.97*	109.24	*44.09*	118.15	*69.09*
Glucose	104.35	*15.20*	107.53	*27.08*	101.51	*14.83*	102.86	*19.71*	114.42	*24.38*	112.52	*23.50*	114.40	*24.32*	108.27	*23.54*
sample size	57		30		33		43		62		32		34		46	

TABLE 2.9-PHYSIOMETRIC TRAIT SUMMARY STATISTICS ACROSS ETHNICITY- AND SEX-SEPARATED GROUPS

Mean and standard deviations (st dev) for the 13 physiometric traits. Central Cameroon (CC) and Southeast Cameroon (SEC) regions are divided into their respective Pygmy (red) and non-Pygmy (blue) populations, each of whom is further divided by sex (females in green, males in yellow). Sample size of each sub-group is printed below each column. Physiometric traits are broadly grouped by type; metrics affected by fasting status are grouped together and highlighted in gray. Units of measure are mg/dl for lipoproteins and glucose, mmHg for metrics relating to blood pressure and beats per minute for pulse.

2.4.4.1 Out of range values

A portion of total cholesterol, HDL cholesterol and triglyceride measurements collected were out of the sensitivity range of the lipid test strips used in this study (Figure 2.7). The vast majority of out-of-range values across the 3 variables were below the minimum sensitivity range. Total cholesterol was by far the most strongly affected: nearly 25% of all total cholesterol measurements sat below the readable range of 100mg/dl. Below-range total cholesterol values are seen predominantly in the males of both Pygmy groups: 19 (59%) of the CC Bezan males and 17 (53%) of the SEC Baka males. For HDL cholesterol and triglycerides, only a small number of individuals fell out of the readable range: 2.6% for HDL cholesterol and 4% for triglycerides, and no one ethnic group had disproportionate numbers of out-of-range individuals as observed for total cholesterol.

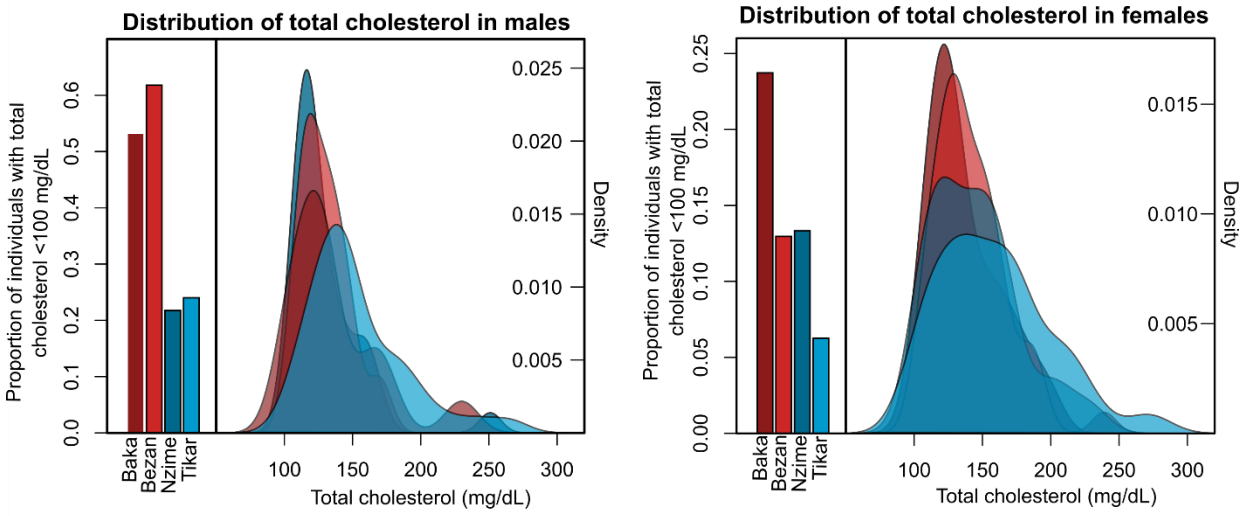


FIGURE 2.7-DISTRIBUTION OF TOTAL CHOLESTEROL SEPARATED BY SEX AND ETHNICITY
The y-axis on the left pertains to the bar-plot and represents the proportion of individuals with total cholesterol values under 100mg/dL while the y-axis on the right represents density. Pygmy groups are labelled in shades of red and non-Pygmy groups are labelled in shades of blue. Groups from Central Cameroon (Bezan and Tikar) are colored in a lighter tone than groups from Southeast Cameroon (Baka and Nzime).

2.4.4.2 Trait comparisons across ethnicity

Physiometric trait comparisons across Pygmies and non-Pygmies were assessed using LMM E1 (Figure 2.4), following the same procedure described in section 2.4.2.1 except BMI is used as one of the five covariates instead of standing height (Table 2.10 and Appendix Table 4).

No observable differences in lipoprotein or blood pressure metrics were noted apart from HDL cholesterol and pulse pressure. HDL cholesterol was significantly higher in non-Pygmies while pulse pressure was higher in Pygmies. This in partial agreement to previous studies: HDL cholesterol trends align with previous reports [125][140] however previously reported blood pressure metrics note markedly lower blood pressure in Pygmies vs their non-Pygmy counterparts [135], which our study found no evidence of.

	E1			
	FIXED	RANDOM		
	Effect direction for: Pygmy Status	R^2M	R^2C	ΔR^2
Total cholesterol		0.11	0.11	0.00
HDL cholesterol		0.08	0.08	0.00
non-HDL cholesterol		0.10	0.11	0.01
LDL cholesterol		0.08	0.14	0.06
LDL : HDL cholesterol		0.03	0.06	0.03
Total : HDL cholesterol		0.04	0.04	0.00
SBP		0.12	0.12	0.00
DBP		0.08	0.09	0.01
Pulse		0.06	0.08	0.03
Pulse pressure		0.08	0.13	0.04
MAP		0.10	0.10	0.00
Triglycerides		0.02	0.06	0.04
Glucose		0.02	0.11	0.09

Legend

E1 = trait ~ Sex + ageCategories + PygmyStatus + BMI + (1 | Region)

Based on the 95% CI:

For PygmyStatus predictor:

Non-Pygmy > Pygmy for this trait

Pygmy > Non-Pygmy for this trait

TABLE 2.10-PHYSIOMETRIC TRAIT PATTERNS RESULTING FROM LMM E1

This table provides a statistical summary of results obtained from a hypothesis test using LMM E1 on physiometric traits and follows the same format as Table 2.6.

Effect direction is shown for the Pygmy status predictor variable and is color coded in blue or red. Effect direction is obtained by assessing the 95% CI and coefficient estimate; an effect is deemed present if the pertaining 95% CI does not overlap zero, and directionality is assessed via the coefficient estimate. Goodness of fit metrics R^2_M, R^2_C, and ΔR^2 are used to assess variance explained by fixed effects, the entire model, or random effects, respectively. Goodness of fit metrics for traits significantly different between Pygmies and non-Pygmies are highlighted by a black box while ΔR^2 equal zero are grayed out. Refer to Appendix Table 4 for a full list of coefficient estimates, standard errors, t-values and 95% CIs for the Pygmy status predictor variable and random effect metrics.

Due to unforeseen circumstances, physiometric measurements influenced by recent intake of food or drink encountered an unexpected limitation: an unknown proportion of the study group failed to complete an overnight fast, which introduced considerable bias into the hypothesis testing scenario for traits most strongly affected by a non-fasting state: glucose and triglycerides. Given that the outcomes of the hypothesis tests for these latter two traits are unlikely to represent unbiased physiological trends they are not presented here; but for completion are included in Table 2.10 and Appendix Table 4. It is worth noting that goodness of fit metrics for physiometric traits were poor, with R^2_C variance metrics ranging from 0.04 to 0.14 and ΔR^2 from 0 to 0.06. This is not altogether unexpected given that lipoprotein levels and blood pressure are strongly influenced by a plethora

of genetic and environmental factors including individual physiology, health status, activity levels and nutrition.

2.4.4.3 Physiometry in the context of reference ranges

This section explores physiometric trends by placing unadjusted physiometric measures for males and females within each ethnic group into the context of available reference ranges. Given the inter-relationship between physiometry and body mass, BMI and waist to height ratio trends are also presented. Finally, unlike the previous section, glucose and triglyceride levels are additionally described alongside both fasting and non-fasting reference values.

Lipoproteins: Since total, HDL and LDL cholesterol are not significantly affected by food or drink, the categories reported in Table 2.11 provide reference ranges for optimal, borderline and high lipoprotein measurements without differentiating between fasting vs non-fasting states [192]–[194]. The vast majority of Cameroonian individuals studied fell into the normal ranges for total cholesterol. It is worth noting that a considerable proportion of individuals had total cholesterol values < 100 mg/dL, which as previously noted fell below the readable range of the CardioChek Plus point-of-care instrument. Most of these individuals (70%) were represented by Pygmy males. Breaking it down by region, 59% of all CC Pygmy Bezan males and 53% of all SEC Pygmy Baka males had total cholesterol levels < 100 mg/dL (Figure 2.7). Less than 5% of individuals in each sex-separated ethnic group had total cholesterol levels exceeding 200 mg/dL, apart from CC non-Pygmy Tikar females, 15% of whom presented with borderline-high levels (200-239 mg/dL). Additionally, the CC non-Pygmy Tikar males and females had the highest proportions of individuals in the high cholesterol group (>240 mg/dL).

		CC				SEC			
		Bezan (Pygmy)		Tikar (Non-Pygmy)		Baka (Pygmy)		Nzime (Non-Pygmy)	
Total cholesterol		Females	Males	Females	Males	Females	Males	Females	Males
Optimal:	<200	98.1%	100.0%	81.8%	95.2%	98.3%	96.8%	96.6%	97.6%
Borderline high:	200-239	1.9%	0.0%	15.2%	2.4%	0.0%	3.2%	3.4%	0.0%
High:	>240	0.0%	0.0%	3.0%	2.4%	1.7%	0.0%	0.0%	2.4%
LDL cholesterol									
Optimal:	<100	88.1%	90.9%	80.0%	75.0%	88.9%	100.0%	79.2%	93.5%
Near optimal:	100-129	9.5%	9.1%	16.7%	18.8%	11.1%	0.0%	20.8%	3.3%
Borderline high:	130-159	2.4%	0.0%	0.0%	3.1%	0.0%	0.0%	0.0%	3.2%
High:	160-189	0.0%	0.0%	3.3%	3.1%	0.0%	0.0%	0.0%	0.0%
Very high:	>190	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
HDL cholesterol									
HD risk:	<40 (M) & <50 (F)	63.4%	70.0%	45.4%	40.5%	70.0%	54.8%	55.2%	52.4%
Normal:	40-59 (M) & 50-59 (F)	23.1%	30.0%	18.2%	35.7%	11.7%	25.8%	24.1%	40.5%
HD Protective:	>60	13.5%	0.0%	36.4%	23.8%	18.3%	19.4%	20.7%	7.1%
BMI									
Underweight	<17	8.9%	3.3%	0.0%	0.0%	3.2%	12.5%	6.1%	4.3%
Thin but normal	17-18.4	10.7%	6.7%	3.0%	9.3%	19.4%	15.6%	9.1%	4.3%
Optimal	18.5-24.9	75.0%	83.3%	60.6%	65.1%	74.2%	71.9%	72.7%	89.2%
Overweight	25-29.9	5.4%	6.7%	21.2%	23.3%	3.2%	0.0%	9.1%	2.2%
Obese	30+	0.0%	0.0%	15.2%	2.3%	0.0%	0.0%	3.0%	0.0%
WHtR									
Low	<0.4	1.8%	0.0%	0.0%	4.7%	0.0%	3.1%	0.0%	0.0%
Optimal	0.4-0.5	42.1%	80.0%	28.1%	67.4%	50.0%	90.6%	51.5%	91.3%
Borderline high	0.51-0.6	52.6%	20.0%	56.3%	25.6%	48.4%	6.3%	48.5%	6.5%
Elavated	>0.6	3.5%	0.0%	15.6%	2.3%	1.6%	0.0%	0.0%	2.2%

TABLE 2.11-LIPOPROTEIN AND OBESITY METRIC PROPORTIONS ACROSS ETHNICITY- AND SEX-SEPERATED GROUPS

Central Cameroon (CC) and Southeast Cameroon (SEC) regions are divided into their respective Pygmy (red) and non-Pygmy (blue) populations, each of whom is further divided by sex (females in green, males in yellow). For each of the latter sub-groups, proportion of individuals in each level (low, normal or high) of lipoprotein/obesity metric is expressed as a percentage. Since total, HDL and LDL cholesterol are not significantly affected by food or drink, the levels reported do not differentiate between fasting/non-fasting. Note that cut-off points for HDL cholesterol levels is sex-specific (males -M, females -F).

75-100% of all individuals fell into the normal ranges for non-fasting LDL cholesterol. In SEC, proportionally more females had LDL cholesterol levels at 100-129 mg/dL as compared to their male counterparts. However, high LDL cholesterol was seen only in the CC non-Pygmy Tikar group; ~3% of non-Pygmy Tikar regardless of sex had LDL cholesterol levels exceeding 160 mg/dL. Only ~30-60% of individuals across the four groups fell into the normal or protective range for HDL cholesterol. Both CC and SEC Pygmy groups had higher proportions with abnormally low HDL cholesterol than their respective non-Pygmy counterparts.

Blood pressure: As per JNC 8 recommendations, hypertension was defined as SBP ≥140 or DBP≥90 mmHg in individuals <60 years and SBP of ≥150 or DBP of ≥90 mmHg for >60 [195] and was observed in 12% of individuals. As per previous recommendations, the hypertension cut-off age standard for this study cohort was lowered to 50. Notably, sample size for individuals <50 was significantly larger across all ethnic groups versus those >50, and thus the findings from the latter group may be unrepresentative. The only exception is the CC non-Pygmy Tikar, whose sample size for individuals aged <50 and > 50 were 41 and 30, respectively. Interestingly, the highest proportion of hypertensive individuals was found amongst the Tikar at ~37%. In individuals <50 years old, the highest proportions of individuals with hypertension (high readings for SBP and/or DBP) were found among the SEC Pygmy Baka at 12%, while SEC non-Pygmy Nzime had the lowest proportions with 3% (data not shown).

Weight-related metrics: ~60-90% of all individuals in this study fell into the optimal range for BMI of 18.5-24.9 while 28-91% fell into the optimal range for WHtR of 0.4-0.5 (Table 2.11). Of note, BMI and WHtR metrics across the four groups were generally discordant, with the exception of the CC non-Pygmy Tikar group, who had the highest proportion of individuals in the 'obese' category of BMI and the 'elevated' category of WHtR.

Triglycerides and glucose: Since the fasting state of participants is unknown, Table 2.12 and Table 2.13 provide reference cut-off values for both fasting and non-fasting states for each of the two measurements. Of the four lipoproteins, triglyceride levels most affected by intake of high-fat foods or alcohol. Hypertriglyceridemia is defined as 200 mg/dl if fasting, and 175 mg/dl if non fasting [196]–[198]. Table 2.12 splits triglyceride measurements into 5 categories and lists the respective proportion of individuals from each sex-separated ethnic group that falls into each category. >80% of individuals from all sex-separated ethnic groups had triglyceride levels <174 mg/dl, considered normal regardless of fasting status. The last two rows of Table 2.12 are the most informative, since triglyceride levels above 200 mg/dl are considered elevated regardless of fasting status. Proportionally more males from all four ethnic groups fell into the high triglycerides group (>200 mg/dl) in contrast to females of the same group. Additionally, one SEC Pygmy Baka male had triglyceride levels >500 mg/dl.

Glucose is also highly affected by a non-fasting state: normal fasting blood glucose levels are 72-99 mg/dl in most healthy individuals versus up to 140mg/dl 2 hours after a meal [199]. Table 2.13 splits glucose measurements in four categories and lists the respective proportions of individuals in each category from sex and ethnicity specific sub-groups. As before, the last row of Table 2.13 is most informative, since glucose levels above 140 mg/dl are considered high regardless of fasting status. Two noteworthy patterns are observable in CC. First, high blood glucose levels (144-199 mg/dl) were proportionally much higher in the CC Pygmy Bezan males vs their female counterparts, and second, CC non-Pygmy Tikar males and female groups had the highest proportion of individuals with glucose readings falling between 70-99mg/dL, and were the only two groups whose glucose measurements did not exceed 140mg/dL.

Triglycerides			CC				SEC			
Assumption:			Bezan (Pygmy)		Tikar (Non-Pygmy)		Baka (Pygmy)		Nzime (Non-Pygmy)	
Non-Fasting	Fasting	*mg/dl*	Females	Males	Females	Males	Females	Males	Females	Males
Normal	Normal	*<150*	84.6%	83.3%	75.8%	85.7%	88.3%	77.4%	79.3%	85.7%
		150-174	5.8%	6.7%	12.1%	7.2%	0.0%	6.5%	10.3%	2.4%
Elevated		*175-199*	7.7%	3.3%	9.1%	0.0%	8.4%	6.5%	6.9%	4.8%
	High	*200-499*	1.9%	6.7%	3.0%	7.1%	3.3%	6.5%	3.5%	7.1%
Very high		*>500*	0.0%	0.0%	0.0%	0.0%	0.0%	3.2%	0.0%	0.0%
		sample size	57	30	33	43	62	32	34	46

TABLE 2.12-TRIGLYCERIDE LEVELS ACROSS ETHNICITY AND SEX ON A FASTING/NON-FASTING SCALE

Central Cameroon (CC) and Southeast Cameroon (SEC) regions are divided into their respective Pygmy (red) and non-Pygmy (blue) populations, each of whom is further divided by sex (females in green, males in yellow). For each of the latter sub-groups, the proportion of individuals at each triglycerides range is expressed as a percentage, and sample size of each sub-group is printed below each column. Given that there is no way to differentiate between individuals who did/did not fast, both 'fasting', and 'non-fasting' cut-off values are shown. The last two rows are most informative given that triglyceride levels above 200 mg/dl are considered elevated regardless of fasting status.

Glucose			CC				SEC			
Assumption:			Bezan (Pygmy)		Tikar (Non-Pygmy)		Baka (Pygmy)		Nzime (Non-Pygmy)	
Non-Fasting	Fasting	*mg/dl*	Females	Males	Females	Males	Females	Males	Females	Males
Normal	Normal	*70-99*	46.2%	46.7%	51.5%	61.9%	31.7%	38.7%	31.0%	54.8%
	Elevated	*101-125*	42.3%	40.0%	48.5%	35.7%	48.3%	35.5%	44.9%	28.6%
	High	*126-140*	9.6%	0.0%	0.0%	2.4%	5.0%	9.7%	10.3%	7.1%
High	Very high	*141-199*	1.9%	13.3%	0.0%	0.0%	15.0%	16.1%	13.8%	9.5%
		sample size	*57*	*30*	*33*	*43*	*62*	*32*	*34*	*46*

TABLE 2.13-GLUCOSE LEVELS ACROSS ETHNICITY AND SEX ON A FASTING/NON-FASTING SCALE
Central Cameroon (CC) and Southeast Cameroon (SEC) regions are divided into their respective Pygmy (red) and non-Pygmy (blue) populations, each of whom is further divided by sex (females in green, males in yellow). For each of the latter sub-groups, the proportion of individuals at each range of glucose is expressed as a percentage, and sample size of each sub-group is printed below each column. Given that there is no way to differentiate between individuals who did/did not fast, both 'fasting', and 'non-fasting' cut-off values are shown. The last two rows are most informative given that glucose levels above 140 mg/dl are considered high regardless of fasting status. Note: In Canada, glucose is not measured in mg/dL.

2.5 Discussion

2.5.1 Anthropometry

1) One of the main questions this study aimed to answer is whether Central African Pygmies are proportional reductions of non-Pygmies, or whether they display unique morphological characteristics. Scaling relationships for Central African populations are poorly defined and it is unknown to what extent scaling patterns are maintained, and across which anthropometric measurements.

There are two known reports pertaining to allometries in Central African Pygmies. A 1977 metanalysis combined and contrasted anthropometric data from 5 separate anthropometric reports (for males only) published between 1948 and 1976. This data suggested that some East Central African Pygmy groups are allometrically reduced in contrast to their neighbours, while the West Central African Pygmy group studied (Binga) instead displayed allometric discordance when compared to their non-Pygmy neighbours (Mbimu) [125]. Next, a 1996 study evaluated body proportions of East Central African Pygmies, and reported findings similar to the earlier study: Central East African Pygmies share common patterns of growth allometry to other African groups [124]. Regrettably, neither of the aforementioned studies are directly comparable to the present work. Though the 1977 study compares scaling patterns between the Central West African Pygmies and non-Pygmies, the reporting was limited to overall differences between ethnic groups

rather than raw trait values.

To allow comparability across individuals of variable standing height, the LMMs used for this purpose included standing height as a predictor variable. In effect, this equalized Pygmy and non-Pygmy standing heights and allowed for an unbiased comparison of body proportions across ethnic groups. Though procedures involving adjustment for standing height have been discussed since the end of the last century [124], to the best of the authors' knowledge they do not appear routine in morphological analyses involving Central African Pygmies and non-Pygmies. Thus, this report provides a novel perspective on the proportions of Cameroonian Pygmies and their non-Pygmy neighbours.

First, roughly half of the anthropometric trait patterns between Pygmies and non-Pygmies changed when the LMM in-use included standing height as a predictor variable (E1), presumably indicating a higher correlation to standing height than the traits whose patterns did not change (concordant patterns). In the present study, discordant trait patterns were observed for metrics relating to the shoulder, hip, sitting height and extremity lengths. In contrast, concordant patterns were observed for skin-fold thicknesses and body mass.

In this context, if Central West African Pygmies were allometric reductions of non-Pygmies, inclusion of standing height as a predictor variable would be expected to remove skeletal component differences observed between the two groups. Though many patterns relating to skeletal components did disappear after adjustment for standing height, differences across Pygmies and non-Pygmies remained for wrist breadth, bicep circumference, head breadth, head circumference, and both arm proportions. As such, Central West African Pygmy groups in this study do not represent allometrically reduced correlates of their non-Pygmy counterparts. See **3)** for further discussion of phenotypic distinctions between the two ethnic groups.

2) We additionally examined whether anthropometric traits are uniquely affected by standing height within Pygmy and non-Pygmy groups by including an additional fixed predictor variable in the form of an interaction term between standing height and Pygmy status. Our study findings indicate that the majority of anthropometric traits correlate to standing height in a similar fashion

regardless of ethnic group. However, there are 3 exceptions which to the best of our knowledge are novel findings and suggest that bicep circumference, tricep skinfold and the ratio of head length to height are differentially affected by standing height depending on whether the individual in question is a member of the Pygmy or non-Pygmy ethnic group.

3) Anthropometric differences between Central African Pygmies and non-Pygmies can be summarized via two statements: **i)** non-Pygmies have significantly larger body mass metrics, skin-fold thicknesses (representing subcutaneous fat) and most head measurements and **ii)** Pygmies and non-Pygmies differed in their relative proportions of the cubit and humerus without change in overall arm length: Pygmies had proportionally longer cubits and non-Pygmies instead had proportionally longer humerus lengths. Apart from the proportionally longer cubit in Pygmies, the results of this study do not support previous observations of similar Central African Pygmy populations, which report differences for sitting height, extremity lengths, shoulder and hip breadths [94]. Two potential explanations for the above discrepancy are provided below.

First, it is unclear whether previous anthropometric studies of Central African Pygmies and non-Pygmies corrected for standing height variation, which we have shown to have a non-trivial effect on anthropometric trait patterns across ethnic groups. Though it is unknown whether the above-mentioned studies corrected for height, it is nonetheless interesting to note that anthropometric trait patterns between Pygmies and non-Pygmies from the LMM which does not include standing height as a predictor variable (E2) more closely align with previously reported Pygmy-specific trait patterns. **Secondly**, region-specific distinctions in anthropometry could be present and potentially align more closely with previous reports. The term “West Central African Pygmy” is a collective term for at least 6 distinct ethnic groups whose range spans four countries. Though Central West African Pygmies are assumed to share a common ancestor ~3,000 ya, extant groups cannot be considered truly uniform. Indeed, this is readily recognized in the aforementioned studies which stress that any reported differences across Pygmies and non-Pygmies are at best generalities given they are based on geographically, culturally and genetically distinct populations [200]. In the present study, anthropometric data was collected from individuals living in two broad regions: Central and Southeast Cameroon, with ~600 km between the two sampled regions. To account for the possible independence of data points within Central or Southeast Cameroon, all

LMM equations included Region coded as a random effect variable. Broadly speaking, anthropometric traits which benefited from the inclusion of region (as indicated by a larger ΔR^2 value) may display intra-regional patterns not picked up by the present statistical analysis. Though out of scope for this report, it is nonetheless important that future studies consider the possibility of region-specific effects within either Central or Southeast Cameroon on the part of either ethnic group.

2.5.2 An evolutionary perspective on the observed trait differences in extant Pygmy and non-Pygmy populations

It has been known for over two centuries that variation in human body size and proportion across different temporal regions is at least in part a result of thermoregulatory adaptation [28][29][33]. Worldwide populations broadly follow predictable patterns consistent with known ecogeographical principles, which describe the response to thermal stress predominantly as changes in body surface area to body mass ratio [29]. Previous studies have shown that populations from warm climates such as Sub-Saharan Africans and Australian Aborigines have longer distal bones of both upper and lower extremities as compared to populations from colder climates [49][50], and arm spans in African Americans are proportionally longer as compared to Caucasian Americans [201]. Given that extant Central African Pygmies and non-Pygmies populations inhabit in the same climate, phenotypic distinctions between the two could indicate differential selective pressures experienced by their ancestral populations. Though the exact selective pressures are unknown, the Central African Pygmy phenotype is thought to result from adaptation to hunter-gatherer lifestyles in tropical rainforests. The thermoregulatory adaptation scenario is one of the four classic hypotheses attempting to explain Central African Pygmy stature. Assuming that ancestral Pygmy populations were subject to thermoregulatory pressures of a tropical rainforest for a substantially longer period than the non-Pygmies (whose ancestral population were Bantu farmers arriving to the Central African belt just ~5,000-3,000 ya [104]) Pygmies would be expected to have body proportions more efficient at heat dissipation.

In this study, overall arm length did not differ between Pygmies and non-Pygmies, however arm segments were proportionally distinct: the cubit was proportionally longer in the Pygmy group vs the non-Pygmy group. The latter observation has been previously noted in Central African

Pygmies and is in line with thermoregulatory principles predicting longer distal limb bones in tropical populations [124]. It is important to point out that the cubit is a composite measurement of the forearm (ulna and radius) and the hand bones (carpals, metacarpals, and phalanges). Thus, which of the latter component(s) resulted in a proportionally longer cubit is unknown. However, assuming that the advantage of a proportionally longer cubit is thermoregulatory in nature, a longer forearm is far more likely given its larger surface-area.

Though the thermoregulatory principle does provide a convincing rationale for a proportionally longer cubit length in Pygmies, it is unlikely that it was the sole environmental pressure or selection mechanism which led to the changes observed in modern Pygmy and non-Pygmy groups. Both populations have had distinct demographic histories over the last 30,000-150,000 years, the estimated time since their last common ancestor [154]. Thus, the number and strength of selective pressures and the extent of the resulting accumulation of genetic differences is potentially vast.

One hypothesis is that underlying molecular mechanisms may be reflected in observed proportional differences between the two groups. For example, a proportionally longer humerus in non-Pygmies does not have an obvious rationale, but it is tempting to speculate that this apparent balance of arm proportions across ethnicities could be demonstrative of intrinsic proportions mediated by presently unknown biological mechanisms and/or reflective of functional thresholds dictating total arm length. Interestingly, a small number of studies have reported the presence of intrinsic patterning or fundamental proportions in the human body [202][203]. However, given the i) lack of large-scale studies on intrinsic body patterning across worldwide human populations and ii) a poor understanding of underlying mechanism(s) responsible for maintaining proportional growth of long bones, the above hypothesis is difficult to test formally at the present time.

Growth is a tightly regulated process capable of integrating cues both internal and external to the growing organ. This allows the growth response to be adaptive and ensures that organ size (such as a limb segment) is coordinated with the rest of the body [204]. Growth cues are mediated by numerous interrelated signalling pathways with multiple regulatory inputs. Hypothetically, genetic or epigenetic changes in either population could conceivably result in altered long bone size and

proportion if they happened to impact the sensitivity or signalling profile of the growth plates, surrounding tissues or distal organs [205][206].

2.5.3 Physiometry

2.5.3.1 Lipoproteins

Previous lipoprotein profiles of Central African Pygmies and non-Pygmies report lower levels of LDL and total cholesterol in Pygmies, and a higher and more favorable HDL cholesterol profile in non-Pygmies [137][123][139][144][142]. Though the present study saw no differences in LDL cholesterol levels, trends for total and HDL cholesterol were similar to those previously reported. Our analyses noted no ethnicity-specific differences for total cholesterol; however this was likely due to a methodological limitation rather than a true lack of differences. As shown in Figure 2.7, a markedly larger proportion of Pygmies fell below the readable range of total cholesterol of 100 mg/dL. In the interest of including these censored values in subsequent data analysis, values under 100mg/dL were re-coded to 100mg/dL which as expected reduced overall data resolution.

HDL cholesterol levels were significantly higher in non-Pygmies, however it is worth noting that HDL cholesterol was surprisingly low across both ethnic groups. Normal HDL cholesterol values range between 40-49 mg/dL for males and 50-59 mg/dL for females. Values lower than 40 or higher than ~116 mg/dL are associated with increased cardiovascular risk while values 60 -116 mg/dL are instead protective [207]–[209]. When split by ethnicity and sex, 41-70% of individuals have abnormally low HDL cholesterol values. HDL cholesterol is broadly seen as an important biomarker in cardiovascular and metabolic health [210], however a 2016 meta-analysis suggested that HDL cholesterol may be a more appropriate metric for overall poor health, rather than being cardiac specific [211]. Levels of HDL cholesterol levels are the result of numerous interactions between inherited and lifestyle factors. Generally, factors like sedentary lifestyle, obesity, smoking, a diet rich in saturated fats, alcohol consumption as well as diseases like metabolic syndrome, COPD, diabetes and HIV are known to contribute to a low HDL cholesterol [212][211].

The causes behind low HDL cholesterol in Central African populations are not well understood, however as mentioned earlier, other studies report similar findings. Two studies on Pygmy and non-Pygmy populations from Southern Cameroon report non-sex separated HDL cholesterol averages range between 24 - 30.16 mg/dL for Pygmy groups and 44 - 49mg/dL for non-Pygmy

groups [139][215]. Importantly, several relatively recent studies point out that dyslipidemia characterized by low HDL cholesterol is prevalent in several African, Middle Eastern and Asian populations [214][215][216].

2.5.3.2 Blood pressure

Blood pressure and pulse pressure are both prognostic values of arterial stiffness [217]. Broadly, hunter gatherer populations reportedly have lower blood pressures in contrast to populations subsisting by other means. Central African Pygmies are reported to have lower blood pressures when compared to neighbouring non-Pygmy groups [125][137][139]–[141][143][146] but other hunter-gather populations tend to have lower blood pressure metrics than Central African Pygmies [135]. In contrast to previous reports, the present study reports a larger pulse pressure in Pygmies, but no apparent difference in either systolic or diastolic blood pressure metrics between the Pygmy and non-Pygmy groups. That being said, the highest proportion of hypertensive individuals from all age categories was found in the CC non-Pygmy Tikar at ~21%, while the highest proportion of young hypertensive individuals (<50) was found in SEC Pygmy Baka at ~12%.

Blood pressure metrics are highly sensitive to environmental factors, and recent lifestyle changes on the part of the Pygmies involving increased use of tobacco, salt and alcohol has been documented over the last few decades [94]. A lack of significant differences between Pygmy and non-Pygmy groups could thus be due to shared environmental variables, however an in-depth study of lifestyle and nutrition factors would be required to explore this further.

2.5.3.3 Triglycerides and glucose

Proportionally more males had high triglyceride levels regardless of ethnic group, possibly indicative of alcohol intake prior to blood draw. High glucose levels were observed more often in the SEC groups. In CC, high glucose levels occurred more often in Pygmy Bezan males vs females. Perhaps most interestingly, the CC non-Pygmy Tikar group was the only group which did not present with blood glucose levels >140mg/dL, despite having the highest proportion of obesity. Hypothetically, such a pattern could have resulted if a higher proportion of Tikar individuals fasted than the other groups.

2.5.3.4 Body fat Metrics

Previous studies of both East and West Central African Pygmies have reported lower BMI in Pygmies versus their non-Pygmy counterparts [130][94] and this study is in agreement with these previous reports. In the last several decades, BMI has been subject to increasing criticism resulting from its limited ability to capture body fat location or differentiate between muscle and fat [218], however to allow comparability to previous phenotypic studies of Central African Pygmies, BMI is reported here. Additionally, WHtR is also presented as it is considered a more accurate measure of central adiposity and has better predictive power for cardiovascular and cardiometabolic health risks [219]–[224]. It is also important to note that though BMI and WHtR metrics are presented alongside standard reference ranges, the latter are based on predominantly Caucasian populations [225][226] and therefore not necessarily appropriate for Central African populations.

2.5.4 Limitations

Several study limitations should be noted. First, it was not possible to randomize participant recruitment since only those present in the village during sampling had the opportunity to participate in the study. A notable sex bias towards women existed in both Pygmy groups, whereas both non-Pygmy groups were more evenly spread (Table 2.4). Though an effort was made to give at least a 24-hour notice before our arrival, data collection was conducted during the day, and as such it is probable that at least a proportion of healthy able-bodied men (and to some extent women) were absent during this time. That being said, all LMM equations did adjust for the latter biases and included both age and sex as covariates.

Physiometric variables in particular had a number of limitations. Total cholesterol, HDL and triglyceride measurements were censored due to the limited range of the CardioCheck Machine, and glucose and triglyceride readings were rendered unreliable as a result of inconsistent fasting states on the part of the study participants. Additionally, two limitations pertaining to blood pressure were noted: 1) The circumference of the Omron blood pressure monitor cuff was too large for many Pygmy individuals (potentially resulting in artificially low values) and 2) blood pressure measurements were conducted in a loud and busy environment. Though efforts were made to

provide a relatively quiet area in which to collect blood pressure measurements, it was not practical sequester individuals entirely.

2.6 Conclusions and Future work

This study is among the largest collection of anthropometric and physiometric measurements available for Central West African Pygmy and non-Pygmy populations. Unique morphologies are highlighted and placed in context of previous reports from similar populations, and where applicable, potential evolutionary selection mechanisms. Additionally, this work provides a novel perspective on the relationship between standing height and anthropometric variables and on scaling patterns and body proportions of Cameroonian Pygmies and their non-Pygmy neighbours. The significance of this work extends beyond the characterization of morphological distinctions as it is the necessary precursor to the identification of underlying developmental, molecular and genetic mechanisms and provides a valuable starting point for future genetic studies. Though DNA samples were collected in tandem with the phenotypic measurements discussed here, genomic sequencing has regrettably not yet been possible. Needless to say, genetic analyses are the natural extension of this work. Once available, whole genome sequencing (WGS) data paired with a comprehensive phenotypic summary of Central African Pygmies and non-Pygmies (this study) will provide a narrowed field in which to focus deep sequencing efforts. Arguably the full potential of this work may be reached via genome wide association studies (GWAS). The mechanisms underpinning intrinsic body patterning and growth in humans is not well understood, and GWAS in the current context has the potential to uncover genetic variants associated proportional growth of long bones, organ size coordination in humans and the scaling patterns observed between Central African Pygmies and non-Pygmies. Elucidation of genetic variants associated with long bone lengths composing the arm, for example, can contribute to the broader understanding of how growth of adjacent organs is harmonized and how terminal organ size is determined.

Once phenotypic and genetic data are incorporated, this study will represent the most detailed biological survey of Central West African Pygmies and non-Pygmies collected to date. The latter has potential to inform on a considerable knowledge gap in the evolutionary history of Central

West African Pygmies, provide novel insights into the genetic and molecular mechanisms of growth regulation and guide future therapeutic treatments of growth-related disorders.

Chapter 3 : A weighted likelihood inference method to detect autozygosity patterns in genotypic data

3.1 Literature review

3.1.1 Small gene pools in human prehistory

Throughout the majority of human evolution, a series of significant bottlenecks and founder effects repeatedly imposed severe limitations on the genetic diversity of ancestral human groups. Current paleontological and genetic evidence strongly suggests that the human species has had a low effective population size (N_e) for the majority of its existence; where N_e is defined as the number of individuals contributing to the gene pool that produces the next generation [227]. The latter, from the perspective of evolution results in more pronounced effects of genetic drift (see section 3.1.2-The effect of genetic drift and endogamy on allele frequencies).

Human ancestors and other members of the *Homo* genus lived in small groups spread over a vast area, with restricted mating opportunities either within the immediate group or with surrounding neighbours, if any existed [228][229]. Restriction in mate choice has thus shaped the course of human evolution, with inbreeding (the mating between closely related individuals) an inevitable component of human evolutionary history [230].

3.1.1.1 Paleontological evidence

Numerous paleontological fossil discoveries over the last few decades have uncovered an unusual number of Pleistocene fossils of the *Homo* genus with a high number of developmental abnormalities [230][233]–[236]. A compendium of these fossils and their associated abnormalities was published in 2018, summing to 75 skeletal and dental anomalies in 66 individuals ranging in age from infant to adult [228]. Fossils were uncovered from Europe, Asia or the Middle East and dated to predominantly the Late Pleistocene period, or ~200 kYa. ~ 40% of these abnormalities are associated with rare or very rare (<0.1 and <0.01%, respectively) genetic variants among modern humans, while ~18% have no known etiology [228]. Though environmental causes could have been responsible for some of the abnormalities, they cannot explain the majority of abnormalities documented. Assuming that the uncovered fossils skeletons are typical of the

population they belonged to; the high frequency of inherited conditions and congenital abnormalities could indicate high consanguinity levels in Pleistocene populations.
Additionally, several studies note morphological uniformities in skeletal features across individuals from the same location, which suggests close familial relationships and reinforces the premise of low genetic diversity among ancient hominin groups [228].

3.1.1.2 Genetic evidence

Ancient DNA (aDNA) analysis has additionally uncovered high inbreeding coefficients, low mtDNA diversity and long stretches of genomic homozygosity in hominin populations from the Upper Paleolithic period, all indicative of high levels of inbreeding [230][237]–[239].

Genetic studies estimate that hominin ancestor groups living 1.2 million years ago (Mya) had an N_e of ~18,500-26,000 while those migrating out of Africa 200-65 Kya had N_e of just ~700-10,000 [232][240]–[243]. The latter estimates are particularly low; especially for a population which, by that time had spread across the entire Old World. Comparative genomic studies of the great apes additionally report the genetic diversity of modern humans to be among the lowest of all other extant primate species, with patterns suggestive of strong, repeated selection sweeps and/or bottleneck/founder events [242]. Here it is additionally interesting to note that a rapid increase in human populations began ~12,000-10,000 ya after the last Ice Age and emergence of agriculture and the subsequent Industrial revolution ~250 ya. The latter initiated an unprecedented and exponential growth in human numbers, which reached 1 billion ~200 years ago (by the early 1800's) and 7.8 billion in 2020 [243]. This increase in N_e in turn has non-trivial implications on human genomic architecture, some of which include increased accumulated mutations/polymorphisms, higher deleterious loads, reduced probability of allele fixation, and an accelerated rate of evolution [244][245].

3.1.2 The effect of genetic drift and endogamy on allele frequencies

Evolution, the process of gradual change in allele frequencies is mediated by mutation, gene flow, genetic drift, and selection processes. For a more in-depth perspective on the latter topics, refer to the introductory chapter. The purpose of this section is to discuss genetic drift and endogamy (a special case of non-random mating), as they relate to the N_e of human groups. Examples of genetic

drift and endogamy in human groups are highlighted, and in the case of endogamy, the preference and prevalence of this behaviour in past and present human groups is placed in context of demographic, socio-economic and psychological considerations.

3.1.2.1 Genetic drift

Genetic drift is a stochastic process of allele frequency fluctuation. Genetic diversity is reduced via genetic drift by two mechanisms: bottleneck or founder effect. Bottleneck and founder effects are both examples of genetic drift and result in decreased allele frequencies. A bottleneck is defined as an abrupt reduction in population size as a result of natural disasters and/or human-mediated events while founder effect is a gene pool reduction due to the separation of a sub-group from a larger population, for example as a result of migration [246]. In the literature, the term bottleneck is often used inclusively to refer to either of the above processes. From a genetic pattern perspective, demographic/historical information is required in order to differentiate between the two.

A series of founder effects and bottlenecks occurred numerous times throughout human demographic history, beginning with the Out of Africa migration(s), to the subsequent peopling of the Old and New worlds. Some of the best evidence for this is the decrease of neutral genetic variation with increased distance from Africa [242][249][250]. Though the specifics of when, where and how many of these events actually occurred continues to be debated, these events are thought to have been at least partially precipitated by ecological changes [155]. Current data provides strong evidence for at least two major bottleneck events which shaped the course of human genetic diversity: the Out Of Africa migration(s) and the Bering Strait crossing into the Americas [248]. Genetic studies estimate that the most recent bottleneck in human history may have occurred ~20 Kya, after which population expansion began [239].

Many modern examples exist for both bottleneck and founder effects mostly from the perspective of rare disease discovery. The Pingelap islanders of Micronesia represent one of the most famous examples of a bottleneck in the modern world. Approximately 5-10% of the Pingelap population is affected by complete achromatopsia, or absence of color vision. To put this into context, this condition is otherwise extremely rare, affecting ~ 0.003% people worldwide [249]. The markedly

increased prevalence of achromatopsia on Pingelap island is due to a population bottleneck, occurring as a result of a typhoon and subsequent famine at the end of the 16th century. Today, all known carriers and affected individuals on the island can trace their ancestry back to one of the survivors, who was a carrier for the autosomal condition [250][251]. It should be noted that the disease carrier happened to be the ruler of the time, which presumably conferred non-random advantages to both his survival and large number of descendants.

The founder effect is also well demonstrated in the Anabaptist populations, more commonly known as the Amish, Mennonite and Hutterite groups. With foundational roots in Europe during the 16th century, religious persecution by the Catholic majority resulted in Anabaptist migration event(s) to North America. Many of these groups maintain extensive genealogical records, which have made it possible to retrace their demographic history with relative accuracy. The North American Hutterites are among the best studied founder populations. Over 40,000 Hutterites living across North America can trace their ancestry back to just ~89 founding members [252]. Resulting from the unique demographic variables and stochastic processes that affect allele frequencies, the Hutterites are at an increased risk for ~ 30 autosomal recessive disorders, with some of the mutations identified entirely unique to those of Hutterite ancestry and as of 2016, a diagnostic carrier screening chip has been made available [253][254][252].

3.1.2.2 Endogamic practices

In addition to genetic drift, societal marriage customs represent powerful drivers of genetic diversity in a given population, and also have the potential to reduce N_e.

Endogamy is the custom of marrying within a specific ethnic, religious, caste or social group, which includes but is not limited to consanguineous marriage. **Consanguinity** refers to a close-kin relationship where the individuals in question share a recent common ancestor. Broadly, **consanguineous marriage** is the union between closely-related individuals, however the precise definition of consanguinity is highly culture-specific and can be quite variable. Most commonly, consanguineous marriage refers to unions between first, second or third cousins or between uncle and niece, but can also extend further down the family tree [230].

3.1.2.2.1 History

Isolation, small founding group sizes and the restriction in mate choice are fundamental aspects of hominin evolutionary history [230]. It is equally important to consider that societal standards of acceptable kinship structures or proscribed mating behaviours can similarly have profound effects on genomic variation. However, unlike founder effects and bottlenecks, socially dictated behaviours regarding mate choice are non-random.

There is a long history of preference for close-kin unions both in human history and prehistory [255][256], though it is unknown when these behaviours arose in the course of hominin evolution. Similar to the unknown timelines and origins of human language, culture or behavioural modernity, questions regarding these processes are regrettably untestable as they do not leave a mark in the physical record [257]. Definitive evidence for endogamic practices is at most ~ 3,500 years old, in the form of written texts from ancient Egyptian, Greek Ptolemaic, Incan and Zoroastrian civilizations [255]. The earliest evidence of possible endogamy dates to ~9 Kya. A 2013 archeological study of an early-agriculturalist site in Jordan found an unusually high incidence of a rare congenital dental defect despite compelling evidence for the presence of a well-established economic exchange with neighbouring communities. This suggests that the signatures of inbreeding, as evidenced by the high rate of congenital defects, did not occur as a result of geographic isolation [258].

3.1.2.2.2 Modern worldwide populations

Worldwide, consanguineous marriages are most prevalent in North Africa, sub-Saharan Africa, the Middle East and Central, West and South Asia; with especially high rates commonly seen in Arab and South Indian communities [232][259][262][263].

As of 2010, ~20% of worldwide populations prefer consanguineous marriages which in 2019 translates to over a billion people. Within communities practicing consanguinity, marriages between kin accounts anywhere from 20 to >50% of all marriages [259][263]–[266].

Attitudes toward consanguinity are highly variable across the major world religions and societies. Worldwide, the most common consanguineous pairings are between first cousins and

uncle and niece, though there is considerable variety in the type of unions preferred, prohibited or considered taboo [261][264]. For example, Muslims in the Middle East have a strong preference for paternal parallel first cousin marriages (two brothers have children who marry). Double first cousin marriages (the couple shares both sets of grandparents) are also acceptable but less popular, however niece-uncle unions are strictly prohibited by the Koran [260][261]. Numerous variations exist across different Muslim communities residing in different countries. For example, Muslims residing in India and Ethiopia do not follow the same consanguinity patterns described above, the latter strongly against marriage between relatives altogether [264][260]. In contrast to practices in the Middle East, Hindus in South India strongly prefer uncle-niece (as well as cross first cousin) marriages [265][260], while North Indian Hindus and Sikhs prohibit marriage between relatives altogether, but follow extensive caste endogamy [265].

3.1.2.2.3 Epidemiological and social *considerations*

Irrespective of country, consanguineous marriage is positively associated with low socioeconomic status, rural areas, low maternal education, and young parental age. In contrast, it has also been noted that the ruling class among worldwide societies commonly favour consanguineous practices predominantly for the purpose of preserving land and wealth [256][266][260][263].

In the Western world, consanguineous marriage customs are typically discussed from the perspective of disease risk, however to better understand the overall health perspectives of this practice it is worth considering the influence of non-genetic factors and their broad implications. Endogamy is a complex phenomenon influenced by social, economic, demographic and psychological variables, which in many cases can provide benefits that outweigh the inherent biological risks associated with inbreeding [264].

Major reasons for consanguineous marriages include maintenance of family property, significantly reduced dowry or bride-wealth payments, increased family unity and preservation of traditional customs and family values. In communities where consanguineous marriage is common, there is a belief that such unions reduce uncertainties, increase family compatibility and lead to more stable marriages and stronger family bonds. In addition, consanguineous marriages have also been reported to have beneficial outcomes specifically for women; improved family unity generally

provides women with greater autonomy, an increased support network and a significantly reduced risk of domestic violence [262][263][266][269]–[273]. Interestingly, it has been suggested that improved stability, organization, and cooperation within endogamic families could be particularly advantageous during periods of societal change or turmoil. A similar theory has also been applied to ancient populations; some anthropologists hypothesize that the high degree of relatedness within ancient hominin groups may have improved their "invasive potential" [271].

3.1.3 Biological perspectives of inbreeding

Going forward, **inbreeding** will be used in the broadest sense to refer to populations with low N_e, regardless of the underlying factors which led to it (bottlenecks, founder effect, genetic drift or endogamy).

3.1.3.1 Disease

Inbreeding depression is the reduced fitness of either an individual or a population resulting from the accumulation of recessive harmful mutations [272], first outlined by Darwin in 1876, who noticed that cross-fertilized plants had improved vigor over those that were self-fertilized [273]. In 1902, Garrod observed increased incidences of rare conditions like albinism and alkaptonuria in children of consanguineous parents, hypothesizing that children of related parents are more likely to inherit two identical disease gametes [274]. A standard measure for level of inbreeding is the inbreeding coefficient, or F, which represents the probability that alleles at a given locus are identity by descent (IBD) [259]–[261]. From a statistical standpoint, marriage between second cousin or closer (where F equals 1/64 or 0.0156) is a relatively standard cut-off point for defining inbreeding. However, it should be noted family trees in highly inbred populations tend to exhibit multiple inbreeding loops across generations which can inflate F between more distant relatives to exceed the 0.0156 cut-off [261].

Inbreeding increases the probability of offspring being homozygous at a given genetic locus due to the increased likelihood that both parents happen to carry the same recessive allele via IBD [263][275]. As a result, rare recessive diseases are more likely to appear in populations with limited gene-pools; which has been indispensable in the discovery of rare recessive diseases and underlying molecular mechanisms responsible for observed clinical phenotypes [261].

The association between inbreeding and increased risk for monogenic disorders has been known for well over a century, during which time hundreds of rare and novel mutations have been reported in inbred communities worldwide [261]. Very broadly, monogenic disorders commonly observed include mutations resulting in blood disorders, enzyme dysfunctions, atrophies, deafness/blindness, behavioural/cognitive abnormalities or improper growth of bones, cartilage or organ tissues [278][279]. That being said, population allele frequencies are the product of complex interplays between evolutionary, demographic and socio-cultural variables. Thus, monogenic disease prevalence and distribution in a given community, inbred or otherwise, is highly variable and population specific [278][275].

It has been observed that some complex traits and diseases are governed by rare variants of small effect size [279], meaning that inbreeding would likewise increase the likelihood that these complex traits and diseases would appear more often in populations with low N_e. To that effect, numerous associations have been made between higher levels of inbreeding and susceptibility to infectious disorders [280]–[282], cardiovascular diseases [264][286]–[289], hypertension [283], neurodegenerative disorders [287], intellectual disability [285][290]–[292], psychiatric conditions [291]–[293], Alzheimer's disease [294], cancer [295]–[299], diabetes [303][304] and osteoporosis [302]. Positive associations between inbreeding and complex traits like height, weight, intelligence, blood pressure and cholesterol levels have also been noted [303]–[311]. Somewhat unexpectedly, parental inbreeding has also been associated with increased incidence of Down syndrome in the offspring, in some cases spanning multiple generations [312]–[314]. Though there is some controversy regarding this last finding [315][316], it nevertheless suggests that chromosomal non-disjunction events could be genetically determined.

3.1.3.2 Epidemiology of disease

Disease risk for the child of a consanguineous couple is highly dependent on the population and family history of both parents, falling anywhere between 0 and >25%. In the absence of known genetic disease present in either the family or population history, studies estimate the risk of infant mortality and congenital disease in the offspring of first-cousins to increase by 3.5% and 1.5-3.5%, respectively, in contrast to offspring of unrelated parents [232][266][279].

The effects of inbreeding on reproductive biology are not well understood, likely due to the difficulties involved in disentangling true biological fertility and fecundity rates from confounding social and economic variables [276][317]. For example, some studies report that consanguineous marriages are associated with higher fertility rates. However, since consanguineous marriages are also positively associated with younger maternal age at marriage, higher completed fertility rates could be due to a longer reproductive period rather than any biological feature. Others report a lack of difference in fertility rates; however, this too is subject to considerable bias since the inbred couple could have made a conscious decision to have more children to compensate for previous infant mortalities [318][265][263].

A genealogical study published in 2008 by the deCODE consortium studied the relationship between kinship and fertility using 165 years worth of genealogical data for over 160,000 Icelandic couples [317]. Being one of the most socioeconomically and culturally homogenous worldwide populations, the usual socio-economic confounders affecting human fertility would be minimized in the Icelandic population. Unexpectedly, their data showed that couples who were either 3rd or 4th cousins experienced the greatest reproductive success, in contrast to more distantly related couples. Couples who were 2nd cousins or closer instead experienced a drop in reproductive success. Though the mechanisms responsible remain to be elucidated, these findings indicate that kinship status may affect biological mechanisms associated with reproductive success.

3.1.3.3 Inbreeding and selection for positive variants

Though damaging variants are more likely to be recessive, and long homozygous regions tend to be enriched for deleterious variants [319], inbreeding does not always produce detrimental effects. To the contrary, several animal studies have reported increased fitness in offspring with higher levels of inbreeding. Though the underlying reasons are not understood, hypotheses for such observations include selection for a “proven genotype” in a given ecological scenario as for a wild population of endangered black robins [320] or of purging selection, as hypothesized for a captive gazelle species [321]. To that effect, it has additionally been demonstrated that inbreeding knock-out mice over several generations progressively produced healthier offspring within <4 generations, possibly via epigenetic mechanisms [322].

Interestingly, a 2005 human study found a homozygous nonsense mutation was associated with increased resistance to norovirus infections in a Swedish population [323]. More broadly, epidemiological studies in the last 2 decades have reported associations between higher levels of inbreeding and decreased incidence of multifactorial diseases like multiple sclerosis in Iranians [324] and breast cancer in Tunisian and Arab populations [325]–[327] These findings are in contradiction to other studies which instead report positive associations between level of inbreeding and multiple sclerosis in a Dutch isolate [328] and breast cancer in a Pakistan and Croatian population [329] [330]. The reasons underlying these findings are unknown but possible explanations for the observed discrepancies could involve inherently different genetic predispositions for the diseases in question or unaccounted for environmental effects.

3.1.4 Stretches of homozygosity

The majority of animals are diploid organisms, meaning their somatic cells carry two sets of homologous chromosomes inherited from maternal and paternal lines. **Homologous** chromosomes are similar in size and genetic sequence and are able to pair, align and recombine during meiotic gamete formation. A given locus on homologous chromosomes can either be identical (**homozygous**) or non-identical (heterozygous).

An inbred population has a reduced gene pool which increases the likelihood that the subsequent generation will be homozygous for a given homologous genetic locus. Homozygosity at a given locus can be either identical by-descent (**IBD**) or by-state (**IBS**). Loci that are IBD were inherited from a recent common ancestor shared by both maternal and paternal lines, while IBS loci are identical by chance, possibly resulting from mutation or recombination. Occasionally, IBS is also used to refer to homozygous loci that have an unknown mode of inheritance, in which case IBD alleles can also be considered IBS, but not vice versa. A related term is **autozygosity**, which refers to homozygosity that is IBD. A locus in IBD is defined in contrast to a reference population and for a defined timescale [331]–[334]. Refer to Figure 3.1 for a visualization of IBD.

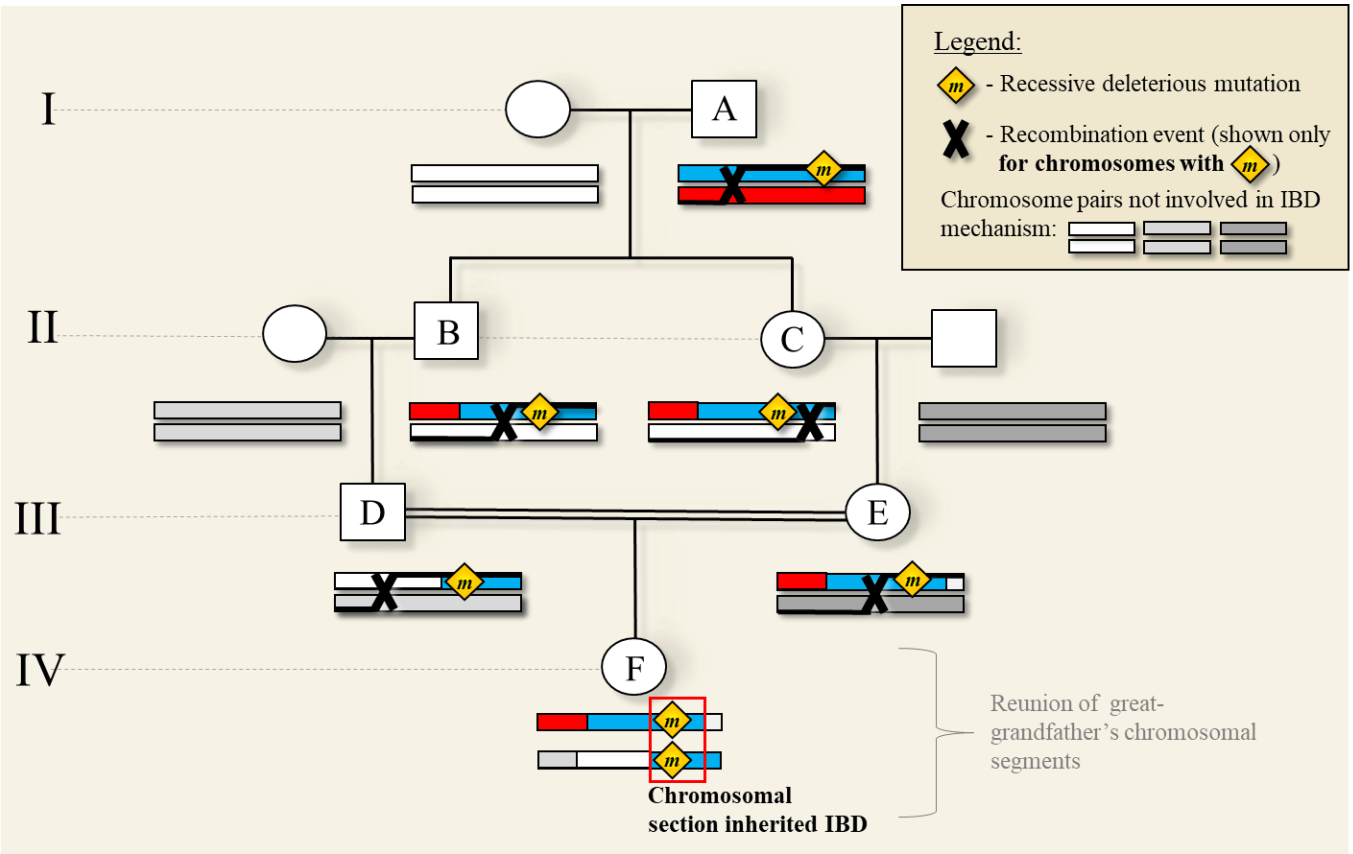


FIGURE 3.1-IDENTITY BY DESCENT; A RECESSIVE DELETERIOUS MUTATION IN A FIRST COUSIN MARRIAGE
The concept of identity-by-descent (IBD) is illustrated via a hypothetical first-cousin union. The pedigree consists of nine individuals across 4 generations (I-IV). For simplicity, only chromosomes directly involved in the IBD-mechanism are shown, with recombination events marked via a black X.
A deleterious mutation event, m (yellow diamond) occurs in one of individual A's chromosomes. Assuming a recessive pattern of inheritance, individual A is now a carrier for a deleterious mutation. In one plausible scenario, individual A's children, grandchildren, and great-grandchildren (individuals B-F) inherit a chromosomal segment containing m. Like their forefather, individuals B-E are also carriers but remain unaffected. Individual F on the other hand, inherits both m copies from each of her parents and is homozygous for the recessive deleterious mutation.

It is important to note that extended tracts of homozygosity along a chromosome can arise for reasons other than inbreeding, including mutation, DNA repair errors, uniparental disomy (both copies of a chromosome come from the same parent), low rates of recombination or linkage disequilibrium [335].

Stretches of homozygosity are referred to by various names in the literature, some of the most common ones include: long contiguous stretch of homozygosity (LCSH), run of homozygosity (**ROH**) run of autozygosity (**ROA**), loss of heterozygosity (LOH) and absence of heterozygosity (AOH), though the latter two refer specifically to processes involving removal of once-present heterozygosity, for example as occurs in disseminated metastatic tumours [335]–[337].

For the purpose of clarity, runs of homozygosity will be referred to as ROH for the remainder of this manuscript and when applicable, homozygous runs known to be IBD will be termed ROA.

3.1.4.1 ROH Features

Three main ROH features are common to all human populations: ROH are **1)** omnipresent **2)** unevenly distributed and **3**) non-randomly distributed across the genome. Though ROH are ubiquitous in the genome, their prevalence and length does correlate with increased inbreeding levels. Regions where ROH are infrequent (reduced homozygosity) are called ROH coldspots while regions where ROH are particularly abundant (increased homozygosity) are called ROH hotspots or ROH islands. ROH hotspots exist across all human populations, and are the predominant ROH form in outbred populations [280][341]–[343].

3.1.4.1.1 Linkage disequilibrium, recombination and selection

ROH hotspots are most prevalent in regions of low recombination, high linkage disequilibrium (LD) and low genetic diversity. As a result, some ROH hotspots have remarkably consistent borders from individual to individual and even across populations, resulting from ancestral haplotypes persisting in areas of low recombination [277][339].

Linkage disequilibrium (LD) is the non-random probability of two or more alleles across different loci occurring together in a population [341][342]. LD is distinct from genetic linkage (the

tendency for nearby loci along a chromosome to be inherited together) but recombination processes are inversely proportional to LD, and thus affect the size and location of LD blocks [343]–[345]. LD patterns are additionally affected by selection, genetic drift and demographic processes such as population history, structure and N_e. Over spans of generations, recombination during each progressive meiotic event shortens LD segments. The smaller the N_e, the less generations required to coalesce to a common ancestor [346]. The oldest LD blocks will be shortest due to the effect of repeated recombination events in each generation, while the most recent LD haplotypes will be the longest [280][342][347][349]–[351].

Apart from low recombination regions, ROH hotspots also form in and around genomic regions under positive selection pressure and again, can be highly population specific [280][341]–[343]. For example, the lactase (LCT) gene is located in an ROH island in Europeans, and has been demonstrated to have been under very strong positive selective pressures [340] [277]**.**

3.1.4.1.2 Worldwide patterns of autozygosity

Considering the small N_e of human founding populations, it is perhaps unsurprising that ROH are such a common feature in human genomes. In organisms with biparental inheritance, the number of ancestors for a given individual doubles each generation and travelling backwards G generations would theoretically equal 2^G ancestors [280][352][353]. However, as one travels back in time, this exponential growth taken together with a progressively smaller human population quickly leads to an impossibly large ancestor tree. For example, it takes just 33 generations travelling backwards to result in over 8 billion (2^{33}) ancestors, which exceeds the current worldwide total. This inconsistency can be explained by the concept of pedigree collapse: the number of ancestors does not double each generation because not all ancestors are unique [349]. In fact, the farther backwards one travels on the pedigree tree, the smaller the N_e and the more likely that a given ancestor appears more frequently.

Individual ROH patterns tend to cluster within a population. Using the human genome diversity project dataset (HGDP), a 2010 study distinguished four population classes based on the sum and length of ROH [347]: **1)** South and West Asia, the Middle East, other Muslim communities are typically distinguished by very long ROH indicative of consanguinity **2)** Oceanian populations

present with a large number of short ROH but few long ROH, indicating reduced N_e in their past, but no recent inbreeding. **3)** Native American populations present with greater numbers of both short and long ROH compared to other worldwide population which suggests both ancient and recent inbreeding. **4)** The last category is best associated with large N_e common to outbred urban populations (Europeans, East Asians and Africans) which are defined by short and infrequent ROH. On average, sub-Saharan Africans have the least ROH across all length categories, consistent with being the most genetically diverse human population. That being said, significant intra-continental variation in ROH patterns exists among hunter-gatherer and agriculturalist populations; for example, Biaka and Mbuti Pygmies have significantly larger ROH totals compared to the less isolated Bantu groups [277][347][339]. Additionally, shorter ROH, which represent the most ancient haplotypes strongly correlate with distance from Addis Ababa in similar fashion to other markers of genetic diversity [351], which is thought to reflect serial bottlenecks throughout the Out of Africa expansion(s) [277][347].

A 2016 paper suggested a 5^{th} population class that represent admixed groups, defined by the on-average shortest but also the most variable ROH among all length categories. Interestingly, the same paper points out that first or second- generation offspring from mixed race parents have the fewest ROH worldwide [277].

3.1.5 Brief history of autozygosity mapping

The inbreeding coefficient, developed by *Wright* in 1922 [352] is arguably the first measure of autozygosity and predecessor to homozygosity mapping. In 1953, several years before the introduction of the first molecular marker, *Smith* noted the theoretical potential of mapping out autozygous genomic regions for the purpose of studying disease-affected children from inbred families. Interestingly, due to the technological limitations at the time associated with constructing full linkage maps, *Smith* had concluded such an approach is likely impractical [353].

Homozygosity mapping (HM) was more formally proposed in 1987 by *Lander and Botstein* [354], whose method proposed using restriction length polymorphisms (RFLPs) to look for homozygosity by descent regions in disease-affected children from inbred communities. The inbreeding coefficient and the logarithm of odds (LOD) score was used to determine whether the

disease marker is linked to surrounding regions of homozygosity. The LOD score is a traditional measure used in parametric linkage analysis originally developed by *Morton* in 1955 [355], calculated as $\log_{10}$ of the following relative ratio: the likelihood of the alternative hypothesis of linkage / the likelihood of the null hypothesis of no linkage) [356]. *Lander and Botstein* hypothesized that once a human genome RFLP linkage map become available, HM would extend the range of detectable recessive diseases compared to traditional linkage mapping.

In 1999, HM was first put to practice by *Broman and Weber* using short tandem repeat polymorphism (STRP) markers from 8 outbred Centre d'Etude du Polymorphisme Humaine (CEPH) reference families [357]. The study reported a surprising number of long homozygosity tracts, providing one of the first pieces of evidence that long homozygosity segments are common across all human groups. In 2009, *Wang et al* extended homozygosity mapping to a genome wide association study (GWAS) and examined the strength of shared autozygous segments in 270 individuals with Parkinson's disease versus 251 healthy controls using ~400,000 SNP chip data [358]. Both of the latter studies included a similar error and mutation model as well as a somewhat modified LOD score. Instead of using LOD to distinguish between linkage and non-linkage, their LOD score distinguished between autozygosity and homozygosity for all possible subsets of continuous markers for all individuals (see formula below). *Wang et al* additionally described the development of an algorithm quantifying the strength of autozygosity for a given chromosomal segment, *i*, using a sliding window framework. In other words, a window of a given size, *w*, moves from one end of a chromosome to the other in pre-defined increments of size *b* (both *w* and *b* are measured in Mb). In each *i* segment, there are K_i SNP markers. At any given SNP marker, the autozygosity status of genotype, G_k is defined by X_k. If SNP *k* is autozygous, X_k will equal to 1; otherwise $X_k = 0$. The probability of observing a specific genotype (G_k) is determined by the autozygosity status (X_k) which in itself is a function of the allele frequency at marker *k* under Hardy-Weinberg equilibrium (HWE). Assuming linkage equilibrium between markers, the strength of autozygosity for chromosomal segment, *i*, is calculated via the following LOD score [358]:

$$LOD(j,i) = \sum_{k=1}^{K_i} log_{10}\left(\frac{\Pr(G_k \mid autozygous\ at\ i)}{\Pr(G_k \mid not\ autozygous\ at\ i)}\right) = \sum_{k=1}^{K_i} log_{10}\ \frac{\Pr(G_k \mid X_k{=}1)}{\Pr(G_k \mid X_k{=}0)}$$

For individual *j*, the null hypothesis that chromosomal segment *i*, is autozygous is compared with the alternate hypothesis that segment *i* is not autozygous. Genotype probabilities incorporate population specific allele frequency estimates and an assumed rate of genotyping error and mutations (equivalent to Equation 1).

In 2012, *Pemberton et al* adapted the method of the *Wang et al* to identify ROH in 1,839 individuals from 64 worldwide populations using ~600,000 SNP obtained from the Genome Diversity Panel-CEPH Cell Line Panel and International Haplotype Map Projects. In contrast to *Wang et al*, the 2012 study used the LOD score to infer individual homozygosity status rather than across groups. This allowed them to define population-specific LOD-score thresholds, investigate ROH patterns across worldwide populations and provide a resource of baseline ROH frequencies for 26 worldwide populations. Using a Gaussian-mixture approach, they additionally classified ROH into three length groups for each population, providing an advantage over most other HM methods which assume that distribution of ROH lengths is constant [338].

3.2 Purpose of study

The purpose of this study is to **a)** improve upon the original LOD-score method proposed by *Pemberton et al* in 2012 [338], **b)** compare the performance of wLOD against other autozygosity detection programs currently available and **c)** evaluate how ROA inference by wLOD is influenced by marker density.

3.3 Methods

3.3.1 Weighted likelihood autozygosity estimator

The LOD score [338] is an ROA inference approach based on a number of earlier methods [357][358]. To obtain a LOD score, a likelihood-based autozygosity estimator is applied in a sliding window framework where window size is defined as a fixed number of SNPs [338] (see section 3.1.5 -Brief history of autozygosity mapping). In this approach, within window *w* in individual *I* from population *j*, the LOD score of autozygosity is then calculated across the *K* SNP markers within window *w*, where the observed genotype G_k at the k^{th} SNP has state X_k, which equals 1 if the SNP is autozygous and 0 otherwise:

EQUATION 1

$$LOD(w,i) = \sum_{k=1}^{K} log_{10}\left(\frac{\Pr(G_k|X_k = 1)}{\Pr(G_k|X_k = 0)}\right)$$

The per-SNP likelihoods of autozygosity and non-autozygosity are based on Hardy-Weinberg proportions (Table 3.1) and include population-specific allele frequencies and an assumed rate of genotyping errors and mutations ε. Missing genotypes are ignored in this algorithm; that is, they have a log-likelihood of zero. The log likelihood of autozygosity for homozygous SNPs is positive and decreases exponentially as a function of allele frequency (Figure 3.2). The log-likelihood of autozygosity for heterozygous SNPs is instead negative and equal to $\log_{10}(\varepsilon)$, thus acting as a penalty for the presence of heterozygous genotypes within a window.

G_k	$\Pr(G_k \mid X_k = 1)$	$\Pr(G_k \mid X_k = 0)$
AA	$(1-\varepsilon)f_{A,j} + \varepsilon f_{A,j}$	$f_{A,j}^2$
AB	$2\varepsilon f_{A,j} f_{B,j}$	$2f_{A,j} f_{B,j}$
BB	$(1-\varepsilon)f_{B,j} + \varepsilon f_{B,j}$	$f_{B,j}^2$
Missing	1	1

TABLE 3.1-PER SNP LIKELIHOODS OF AUTOZYGOSITY AND NON-AUTOZYGOSITY
For a given genotype (G_k), per-SNP likelihoods of autozygosity ($X_k = 1$) and non-autozygosity ($X_k = 0$) are determined by Hardy-Weinberg proportions ($p^2 + 2pq + q^2$). The latter likelihoods include population-specific (for population *j*) frequencies of alleles A and B denoted by $f_{A,j}$ and $f_{B,j}$. Missing genotypes are given a value of 1 and are therefore ignored by the LOD and wLOD algorithms ($\log_{10}[1] = 0$). This has been previously reported in [359].
Table 1 in original publication [360].

To address the shortcomings of the LOD score method, a weighted LOD-based method (wLOD) was developed. wLOD accounts for non-independence among SNPs and the probabilities of recombination and mutation within window *w*:

EQUATION 2

$$wLOD(w,i) = \sum_{k=1}^{K} \left[log_{10}\left(\frac{\Pr(G_k|X_k=1)}{\Pr(G_k|X_k=0)}\right) \times \mathrm{Corr}(p_k,[p_1,p_K]) \times \Pr(no\ recombination|[g_{k-1},g_k]) \times \Pr(no\ mutation|\mu,[p_{k-1},p_k]) \right]$$

The approach of *Chen et al* [361] is adapted as a means to incorporate LD information into the wLOD(*w*,*i*) estimator, weighting the log-likelihood of SNP k by the reciprocal of the sum of pairwise LD between SNP k and all other SNPs within window w calculated as:

EQUATION 3

$$\mathrm{Corr}(p_k,[p_1,p_K]) = \frac{1}{\sum_{l=1}^{K} LD_{k,l}}$$

and bounded in the interval [1/*K*,1]. An intuitive explanation for this correction is that when a number of SNPs are highly correlated, they provide redundant information. By weighting the log-likelihood for SNP *k* as a function of its correlation with all other SNPs within window *w* it contributes only the unique autozygosity information it possesses to wLOD(*w*, *i*).

LD reflects historical recombination and mating patterns in a population and is largely insensitive to the effects of mating patterns within the last few generations that can, through recombination events onto very similar haplotype backgrounds, lead to false-positive autozygosity signals [362]. If a recombination event occurred in the recent past, we would want to place the ROA boundary at that position. However, the direct detection of such events becomes increasingly challenging as the distance between genotyped positions increases, particularly when their genotype patterns are highly correlated. Thus, wLOD also weights the log-likelihood of SNP *k* by the probability of no recombination events having occurred within the genomic interval bounded by SNP $k-1$ and SNP *k* in the last *M* generations, calculated based upon their genetic map position g (in Morgans) as previously described [363][364]:

EQUATION 4

$$\Pr(no\ recombination|[g_{k-1}, g_k]) = e^{-2M(g_k - g_{k-1})}$$

Here, M reflects the expected minimum number of meioses since the most recent common ancestor (MRCA) for a pair of IBD haplotypes [365], tuning the sensitivity of the wLOD estimator to the expected age of IBD haplotypes underlying ROA without limiting it to only ROA of that age. In a population-genetic context, *M* can be set based upon effective population size estimates and the probability that a pair of individuals share a common ancestor *M* generations in the past [366], while in a disease-genetic context *M* can instead be set based on known relationships between affected individuals.

Finally, wLOD accounts for the potential presence of unobserved genetic differences within the genomic interval bounded by SNP $k-1$ and SNP k by weighting the loglikelihood of SNP k by the probability of no unobserved mutation events having occurred within the genomic interval in the last M generations, calculated based upon their physical map position *p* (in bp) and a per-base mutation rate μ using an approach adapted from *MacLeod et al* [367]:

EQUATION 5

$$\Pr(no\ mutation|\mu, [p_{k-1}, p_k]) = e^{-2M\mu(p_k - p_{k-1})}$$

As shown in Figure 3.2B, the recombination and mutation weightings reduce the log likelihood of SNP k as a function of its distance from SNP $k-1$, reflecting the increased likelihood of recombination and mutation events going undetected as the size of the uninterrogated region increases. At $M = 7$, the log-likelihood of SNP k is reduced by ~ 50% at an inter-marker distance of ~ 300 kb and ~ 85% at ~ 100 kb. When the wLOD method is applied to high-density SNP genotype datasets, such as that created by WGS, the recombination and mutation weighting will have a minimal effect on the wLOD score of window *w*. However, when applied to lower-density SNP genotype data, such as that created by genotyping microarrays—particularly those available for non-human species such as dogs and cows—the much larger inter-marker distances in these datasets will lead the recombination and mutation weighting to have a much larger effect on the wLOD score of window *w*. It can also be seen that as *M* decreases the magnitude of the change in the weighting with increasing distance also decreases; thus, wLOD scores in populations with small effective population sizes or in disease studies where affected individuals share a more recent

common ancestor (smaller *M*) will be adjusted to a lesser extent than those with larger effective population sizes or where affected individuals share a more distant common ancestor (larger *M*).

3.3.2 Genotype datasets

The dataset used in this study is The 1000 Genomes Project Phase 3, release v5a (henceforth TGP) [368] was accessed March 29th 2015. The TGP dataset consists of phased genotypes at 84,801,880 genetic variants in 2,504 individuals from 26 worldwide human populations sequenced using a low-coverage whole genome sequencing (WGS) approach. Occasional missing genotypes were imputed during genotype phasing, as such this dataset contains no missing data. We first developed a subset of this WGS dataset in which to perform individual-level quality control prior to developing different subsets in which to evaluate the performance of the wLOD method. All dataset subsets underwent a common set of quality control procedures previously described by *Pemberton et al* to remove low-quality variants [369].

3.3.3 Individual-level quality control

To independently verify the putative unrelatedness and population labeling of individuals reported by TGP, we developed a preliminary Omni dataset comprised of the 2,165,831 autosomal, 48,458 X-chromosomal, and 543 Y-chromosomal SNPs in the TGP dataset to match those present on the Illumina HumanOmni2.5-8 BeadChip. Across the 1,693 individuals for which genotypes derived using the HumanOmni2.5-8 BeadChip were also available, genotype concordance between the WGS- and BeadChip-derived genotypes lay between 0.99431 and 0.99986 (mean = 0.99953, SD = 0.00041).

3.3.4 Preparation of final datasets

A total of 68 individuals were removed as a result of incorrect population/sex labelling or due to close relative status as previously described in *Pemberton et al* [369]. 47 individuals did not genetically cluster with other individuals sharing the same population label, 6 whose reported sex is likely erroneous, and 15 pairs of relatives more closely related than first cousins (1 intra- and 14 inter-population pairs), from which one individual was removed.

Four subsets of the TGP dataset were developed using autosomal biallelic variants: **a**) The WGS dataset was developed using all single nucleotide variants (SNVs) which passed quality control criteria, totaling 75,071,695 SNVs. Next, three data subsets were created in order to mimic: **b)** a whole-exome sequencing dataset as captured by the Roche Nimblegen SeqCap EZ Human Exome Library v3.0 system (WES henceforth, composed of 1,830,512 SNVs) **c)** an Illumina HumanOmni2.5-8 BeadChip (**Omni2.5;** 2,166,414 SNPs) and **d)** an Illumina OmniExpress-24 BeadChip (**OmniExp;** 676,445 SNPs). ~96% of markers present on the OmniExpress-24 BeadChip are present on the HumanOmni2.5-8 BeadChip, the OmniExp dataset was created by taking a subset of the Omni2.5 dataset.

Since the wLOD method explicitly accounts for LD among genotyped positions within a given window (see Equation 3) LD pruned datasets are not considered here. Additionally, since the wLOD estimator is most informative via homozygous alleles low or rare frequencies (Figure 3.2A), minor allele frequency (MAF) pruned datasets are not considered.

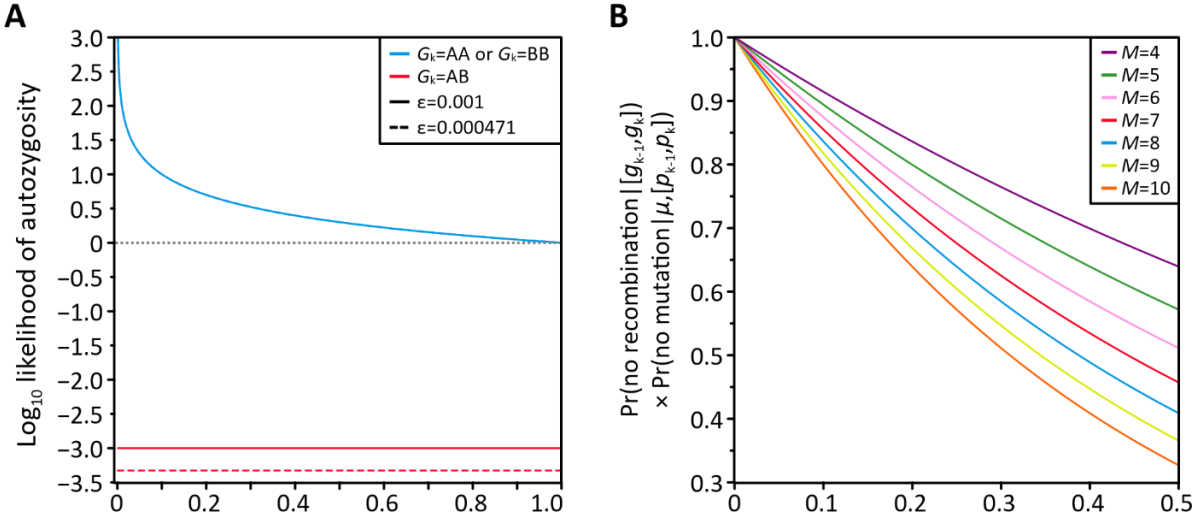


FIGURE 3.2-PROPERTIES OF THE LIKELIHOOD-BASED ESTIMATOR

A) The relationship between allele frequencies (x-axis) and the log-likelihood of autozygosity via Equation 1 (y-axis). The log-likelihood of autozygosity (LOD score) is shown for homozygous (G_k =AA or BB) and heterozygous (G_k =AB) genotypes which are determined by Hardy Weinberg proportions (Table 3.1). The error rate ε is equal to 0.001 [338] and 4.71×10^{-4} (this study). Note that the line for homozygous genotypes (G_k = AA or BB, in blue) with $\varepsilon=4.71\times10^{-4}$ overlaps the line with ε=0.001.

B) Several effective population sizes (M = 4 -10) are plotted to illustrate the relationship between the distance separating a pair of SNPs in centimorgans (x-axis) and the magnitude of the product of the probabilities of recombination and mutation (Equation 4). A fixed crossover rate of one centimorgan (cM) per one million base pairs is assumed with a sex-averaged mutation rate (μ) of 1.18×10^{-8} [373]. *Figure S1 in original publication* [360].

3.3.5 Geographic distances

The geographic distance of each population from Addis Ababa, Ethiopia, was calculated as in *Rosenberg et al*. [370] with the use of waypoint routes, based on the sampling location reported by the TGP [368].

3.3.6 Simulation of genetic and true ROA datasets

Note: Simulation of genetic and true ROA datasets was completed by Dr. Trevor Pemberton.

Fifty independent replicates of genetic datasets for two demographic scenarios were generated using a forward in-time process as previously described by *Kardos et al* [371]. In their original approach, prior to performing the simulation steps they placed N predetermined polymorphic SNV onto the chromosome's genetic map by randomly sampling N unique genetic map positions in the range 0 to g_{max} (the user-defined genetic map length of the simulated genome), only converting genetic map positions to physical map positions based upon a fixed user defined recombination rate to physical map distance relationship when writing the simulated datasets to file. Here, we modified their approach to instead create a non-uniform distribution of recombination rates across the simulated chromosome and allow any base pair to mutate during the simulation.

Genetic map position, g_p, is assigned to physical map position p and is equal to the base-pair count from the beginning of the chromosome. Based on the user-defined values for g_{max} and recombination rate θ, all values of g lie within the interval $[0, g_{max}]$ and all values of p lie within the interval $[1. . (g_{max}/\theta) \times 1,000,000]$. To begin, we created a backbone of genetic and physical map positions onto which we will place all other positions, randomly drawing $(g_{max}/\theta)+1$ values of g and assigning them in increasing order to p in the range $[1. . (g_{max}/\theta)]$ (i.e. every Mb). Next, we randomly chose N values of p to be predetermined polymorphic SNVs, and then randomly assigned each a value of g based upon the backbone interval in which it was located, again ensuring that values of g always increase as a function of p. Finally, all values of p that were not among the set of predetermined SNVs were assigned a value of g through interpolation onto the construct created by the values of p and g assigned to the predetermined SNVs. This approach created a non-uniform relationship between physical and genetic map distance along the simulated chromosome that is similar to that observed on real human chromosomes.

To extend the method of *Kardos et al*. to enable any base pair on the simulated chromosome to mutate, for each individual in each generation, the number of mutations that occur during each meiosis was drawn from a Poisson distribution with mean $\mu \times [(g_{max}/\theta) \times 1,000,000]$, where μ is mutation rate. The base pairs to be mutated were then chosen at random from all $(g_{max}/\theta) \times 1,000,000$ possible positions without replacement. Mutations were tracked and then incorporated into the genotypes of individuals in the analyzed dataset; all monomorphic positions were removed during dataset construction.

In all simulations, we set g_{max} to 325 cM, θ to 1.3 cM/ Mb [372], and μ to $1.18 \times 10{-}8$ [373], and scaled θ and μ by a factor of 10 to increase genetic diversity in the final generation [374].
N was chosen separately for each simulated scenario such that the final number of polymorphic SNVs in the dataset (both predetermined and de novo) was ~750,000; N = 725,000 for scenario 1 and N = 650,000 for scenario 2. Because predetermined polymorphic SNVs can become fixed over the course of the simulation, their numbers in the analyzed datasets lay between 679,256 and 717,855 (25,788–31,503 de novo SNVs) for scenario 1 and between 633,582 and 638,675 (103,871–110,077 de novo SNVs) for scenario 2. The simulated WGS datasets used in the analyses contained 50 individuals randomly chosen from among the 75 present in the final generation with genotypes for 709,862–746,963 polymorphic SNVs in scenario 1 and 737,957–748,572 polymorphic SNVs in scenario 2.

To better mimic real genetic datasets, genotyping errors were randomly introduced into each of the simulated datasets at a rate of ε 0.001. This is a conservative value similar to but slightly higher than the average rate of genotype discordance across 1693 individuals between genotypes in their WGS data and those obtained at the exact same SNVs with the Illumina HumanOmni2.5 BeadChip [375].. Analysis of the simulated pedigrees found the parents of individuals in the final generation to have a common ancestor on average three generations in the past for scenario 1 (all between 1 to 5 generations) and four generations in the past for scenario 2 (all between 1 to 7 generations) and M was set to these average values when analyzing their respective datasets.

3.3.7 Parameters used in the application of PLINK, LOD, and BCFtools methods

For PLINK, we allowed a maximum of 2% of SNPs to have heterozygous genotypes and 5% of SNPs to have missing genotypes for a window to be inferred to be autozygous [376]. The LOD method and BCFtools were applied using the same allele frequency estimates and error rate ε as the wLOD method, while BCFtools additionally incorporated genetic map positions and performed Viterbi training with initial transition probabilities between autozygous and non-autozygous states and vice versa of 6.6×10^{-8} and 5.0×10^{-9}, respectively, to optimize its underlying model prior to ROA inference [377].

3.3.8 Application of the wLOD estimator to real data

To minimize the number of variables that varied in within-dataset comparisons, we used a single set of allele frequencies when calculating wLOD scores at all window sizes considered. To account for sample-size differences among populations, we used a resampling procedure to estimate the allele frequencies, sampling 100 non-missing alleles with replacement and calculating allele frequencies from these 100 alleles. As a consequence of the resampling procedure, it was possible for an individual to possess an allelic type whose frequency was estimated to be 0 in the sample of 100 alleles. SNV positions at which this scenario was encountered were treated as missing when calculating wLOD scores for all windows containing the positions in individuals that had the allelic type of frequency 0.

As our datasets contained phased genotypes, in the LD correction applied in the wLOD estimator (Equation 3), LD was estimated with the correlation coefficient r^2 [378] using a resampling procedure to account for the possible influence of sample size on homozygosity-based LD statistics [379]. For each pair of SNPs, we randomly sampled 55 individuals (the smallest population sample size in our dataset; see Table 3.2) without replacement and the LD computation was performed using those 55 individuals. Note that we used a single set of LD estimates when calculating wLOD scores at all window sizes considered.

In the recombination rate correction applied in the wLOD estimator (Equation 4), the genetic map position of each marker in the Omni2.5 dataset and its subsets were downloaded from the Laboratory of Computational Genetics at Rutgers University (*http://compgen.rutgers.edu*). The

genetic map position of each marker in the WES and WGS datasets was determined by interpolation onto the Rutgers linkage-physical map [380] based on the physical map position in build hg19 of the University of California – Santa Cruz (UCSC) reference genome assembly. Due to computer memory requirements for Gaussian kernel density estimation, the wLOD score distributions used to determine the autozygosity score thresholds in the WGS dataset considered only 20 individuals chosen at random. Based on our investigation into the effect of sample size on score threshold, we do not expect this approach to have biased our detection of ROA in the WGS dataset. All genome-wide windows were, however, considered when determining optimal window sizes in the Omni2.5, OmniExp, and WES datasets.

3.3.9 Classification of ROA

We ran unsupervised Gaussian fitting of the ROA length distribution using the mclust package (v.5.2) [381] in the R statistical software (v.3.3.3)[382], allowing component magnitudes, means, and variances to be free parameters. Bayesian Information Criterion (BIC) likelihoods with increasing number of components (*G*) were calculated using the function mclustBIC, while final classification under the five-component model was performed using the function Mclust. Violin plots [383] of total lengths of ROA in individual genomes were produced separately for each length class using the vioplot function from the vioplot package in R.

3.3.10 Genomic distribution and geographic patterns of ROA

The frequency at which each SNV was present in ROA in each population was calculated as described previously in *Pemberton et al* [338]. To compare the genomic distribution of ROA across populations, we calculated mean ROA frequencies in non-overlapping 50 kb windows across all SNVs polymorphic in that population that were within the window and excluding windows that lay within the centromere and telomeres. To evaluate the similarity of ROA frequency patterns among populations, we performed classical (metric) multidimensional scaling (MDS) separately for each ROA length class based on a matrix of ROA frequency dissimilarities between all pairs of populations, calculated as one minus the Pearson correlation coefficient (r) of their mean ROA frequencies across windows. We then applied MDS to this matrix using cmdscale in R.

We compared population patterns in the MDS based on ROA frequencies to an MDS based on a matrix of pairwise F_{ST} among populations calculated with our WGS dataset and the method of *Hudson et al* [384] according to the recommendations of *Bhatia et al* [385]. The similarity of patterns in our MDS of ROA dissimilarities and those in the MDS of F_{ST} was evaluated with the Procrustes method [386].

3.3.11 Relationship between ROA and genomic variables

For each ROA length class, we investigated recombination rate and the normalized haplotype-based selection statistic or nS_L [387] for correlations with ROA frequency across the autosomes. Population-based recombination-rate estimates were obtained from TGP [368], and nS_L values for each of the 26 populations were calculated in the WGS dataset considering only SNVs with MAF > 0.05 and normalization of unstandardized scores in 100 genome-wide frequency bins with selscan [388]. Comparisons between ROA frequency and recombination rate and nS_L were performed as described in *Pemberton et al* [338] considering the mean value of each variable in non-overlapping 50 kb windows, excluding windows within the centromere and telomeres, calculated across all SNV within the window for which the variable was available. Admixed Afro-European (ASW and ACB) and Mestizo (CLM, MXL, PEL, and PUR) populations and the geographically imprecise CEU (Utah residents of Northwestern European ancestry) group were omitted from geographic analyses but were included in the scatterplots.

3.4 Results

3.4.1 Properties of the wLOD estimator

To investigate how the LD, recombination, and mutation corrections implemented in the wLOD estimator influence per-window autozygosity likelihoods, we compared them with those of the LOD estimator using the TGP dataset that contains phased genotypes for 84,801,880 genetic variants discovered using a low-coverage WGS approach in 2,436 unrelated individuals from 26 human populations (Table 3.2) [368].

To approximate a typical microarray-based SNP genotyping study, we created a subset of the WGS dataset called Omni2.5, which contains 2,166,414 autosomal SNPs and approximates a commonly

used microarray (see section 3.3-Methods). In all analyses, μ was set to 1.18×10^{-8} [373] and ε was set to 4.71×10^{-4}, the average rate of discordance across samples between genotypes in our Omni2.5 dataset and those obtained for 1,693 of the 2,436 individuals directly with the Illumina HumanOmni2.5 BeadChip [368]. Unless otherwise stated, M was set to 7, a conservative value broadly reflecting the average of effective population size estimates for populations included in the TGP [366], [368], [389]. Window size was varied in an arbitrary interval [K_0, K_n] in which K is increased in 10 SNP increments (i.e. $K_n = K_0 + [10 \times n]$).

The genome-wide distribution of wLOD scores for all windows in the Omni2.5 dataset is bimodal and centered around 0, with wLOD scores under the left-hand mode favoring the hypothesis of non-autozygosity, whereas those under the right-hand mode favor the autozygosity hypothesis Figure 3.3A). The area under the right-hand mode decreases as a function of window size as ROA are progressively covered by fewer but longer windows. In addition, while the location of the right-hand mode does not change appreciably with window size, there is a noticeable shift toward lower wLOD scores in the left-hand mode with increasing window size, likely reflecting the larger number of heterozygous SNPs in non-autozygous compared with autozygous regions and their greater cumulative effect on wLOD scores with increasing window size. This shift progressively increases the distance between the non-autozygous and autozygous modes until either the autozygous mode disappears (non-autozygosity, whereas those under the right-hand mode favor the autozygosity hypothesis (Figure 3.3B) or the intermodal distance begins to decrease instead (Figure 3.4); both potentially reflecting the point above which window lengths exceed those of the majority of ROA in the sample set. In this scenario, as window size increases, autozygous windows increasingly overlap non-autozygous regions flanking shorter ROA leading them to encompass greater numbers of heterozygotes within these non-autozygous regions, deflating their wLOD scores. Whether the autozygous mode disappears or shifts toward lower wLOD scores is likely determined by the abundance of ROA and their levels of support in the sample set: sets with fewer ROA and ROA with generally lower wLOD scores lose their autozygous mode while those with large numbers and higher wLOD scores have it shift toward the non-autozygous mode. Nevertheless, the location of the minimum between the two modes does shift subtly toward higher wLOD scores with increasing window size, reflecting the expected increase in scores for autozygous windows as a function of the number of SNPs within the window. The periodicity

observed in the genome-wide score distribution of the original LOD estimator [338] is absent with the wLOD estimator, indicating that this property was a reflection of LD among SNPs within the window.

Population			Geographic region	Consanguinity frequency [a]
ID	Name	Total number		
ESN	Esan	94	Africa	–
GWD	Gambian	**109**		–
LWK	Luhya	96		–
MSL	Mende	80		–
YRI	Yoruban	107		51.20%
GBR	British	89	Europe	0.40%
CEU	European American	97		0.20%
FIN	Finnish	98		0.17%
IBS	Iberian	107		1.99%
TSI	Toscani	104		–
BEB	Bengali	84	Central/South Asia	5.00%
GIH	Gujarati	101		4.90%
PJL	Punjabi	96		0.90%
STU	Sri Lankan Tamil	96		38.20%
ITU	Telugu	101		30.80%
CDX	Dai	92	East Asia	21.30%
JPT	Japanese	103		4.80%
KHV	Kinh	94		–
CHB	Northern Han	101		1.16%
CHS	Southern Han	102		3.43%
ASW	African American	**55**	Admixed	–
ACB	Afro-Caribbean	94		–
CLM	Colombian	89		2.83%
MXL	Mexican American	62		0.80%
PEL	Peruvian	84		1.90%
PUR	Puerto Rican	101		3.30%

TABLE 3.2-POPULATIONS INCLUDED IN THE 1000 GENOMES PROJECT, PHASE 3
Population ID, name, sample size and geographic region is provided for the human populations included in the 1000 Genomes Project Phase 3 dataset (grouped by color using the same coloring scheme as in Figure 3.5 and Figure 3.13). There is a total of 2,436 unrelated individuals from 26 human populations. The smallest sample size is found in African-Americans (admixed group) at 55 individuals, while Gambians (Africa) are the largest at 109.
[a] Consanguinity frequencies were obtained from http://www.consang.net
Table 2 in original publication [360].

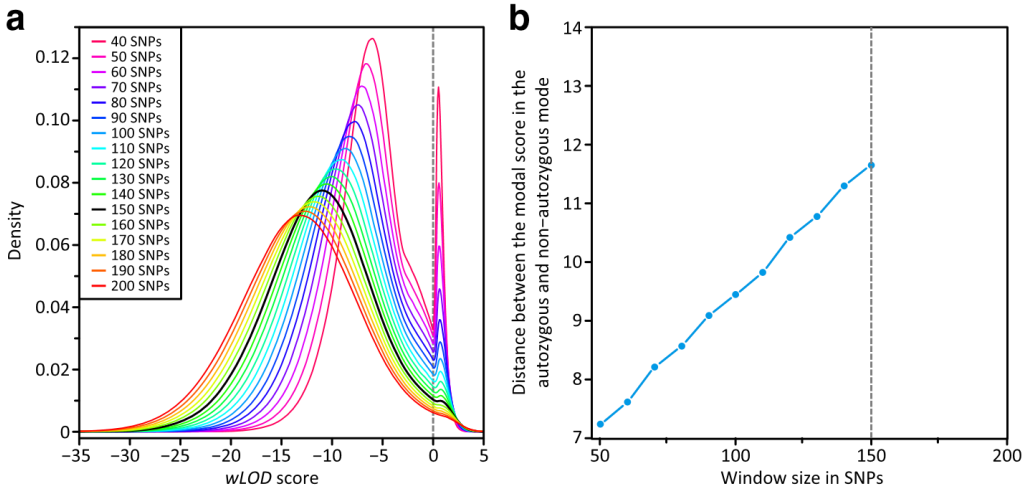


FIGURE 3.3-DISTRIBUTION OF GENOME-WIDE WLOD SCORES IN EUROPEAN AMERICANS

A) Using the European Americans (CEU) in the Omni2.5 dataset, Gaussian kernel density estimates of the pooled wLOD scores (for the 97 CEU individuals) are plotted for all window sizes between 40-200 SNPs moving at 10 SNP increments. The largest window size that produced a clear bimodal distribution (150 SNPs) is shown in black. **B)** The intermodal distance between autozygous and non-autozygous modes increases with increasing window size. Note: Though A and B analyses use the CEU population, the patterns shown above are representative for the other 25 populations.

Figure 1 in original publication [360].

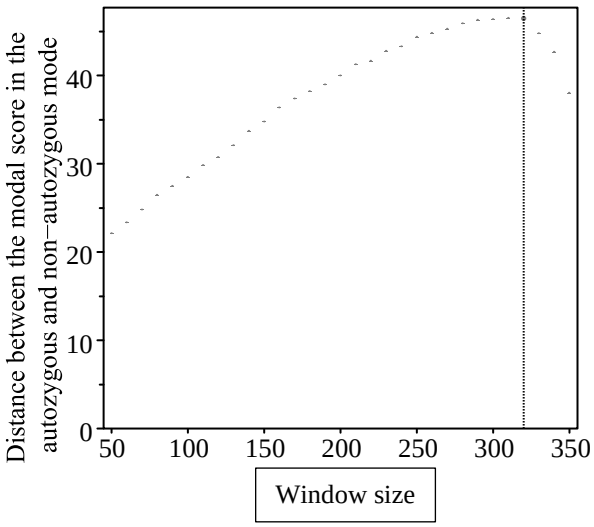


FIGURE 3.4-INTERMODAL DISTANCES IN THE PUERTO RICAN POPULATION

The relationship between window size and intermodal distance (between autozygous and non-autozygous modes) is shown for the Puerto Rican population (PUR) using the Omni2.5 dataset. The vertical dotted line indicates the window size at which the maximum intermodal distance was attained (320 SNPs).

Figure S2 in original publication [360].

To evaluate how the improvements incorporated into the wLOD estimator (Equation 2) influence per-window scores as compared to the original LOD estimator (Equation 1), we compared wLOD and LOD scores in the Omni2.5 dataset with a window size of 150 SNPs (Figure 3.5A), the largest value that produced a clear bimodal wLOD score distribution in all 26 populations. Across populations, per-window wLOD scores differed from their corresponding LOD scores by between −103.87 and 454.07 with the range and average differences between wLOD and LOD scores increasing as a function of a population's geographic distance from East Africa ($\rho = 0.8460$ with $P = 5.029 \times 10^{-6}$ and $\rho = 0.8846$ with $P = 4.961 \times 10^{-7}$, respectively), reflecting increasing LD [390][391] and decreasing genetic diversity [392]–[396]─leading to larger distances between polymorphic SNPs with increasing distance from Africa. Among the six admixed populations included in Phase 3 of the TGP, those of mixed African and European ancestry (ACB and ASW) had smaller ranges and averages of wLOD and LOD score differences than those of mixed of Amerindian and European ancestry (CML, MXL, PUR, PEL), consistent with the lower LD [397]–[399] and higher genetic diversity [396][400] of admixed African-European populations compared with Amerindian-European populations. Across populations, 5.15–47.93% of all windows in the right-hand "autozygous" mode with the LOD estimator were present in the left-hand "non-autozygous" mode with the wLOD estimator (Figure 3.5C) potentially reflecting false-positive autozygosity signals reported by the LOD estimator. False positive signals can arise due to the non-independence among homozygous SNPs which cumulatively give the mistaken impression of autozygosity. The proportion of windows was lowest in African populations and highest in most European populations, increasing incrementally through Central/South Asian and East Asian populations. This pattern can be explained by population differences in the location of the autozygous mode and its shift toward lower scores with the wLOD estimator. Modal LOD and wLOD scores in the autozygous mode are generally smallest and most similar in European populations and highest and least similar in African populations. Thus, for a given unit decrease in score between the LOD and wLOD estimators, an autozygous LOD window has a greater chance of transitioning to the non-autozygous wLOD mode in Europeans populations than in African populations. Consistent with this hypothesis, the magnitude of the difference between modal LOD and wLOD scores in the autozygous mode and the location of the minima between the autozygous and non-autozygous modes is significantly negatively correlated with the proportion of autozygous LOD windows that transition to the non-autozygous wLOD mode.

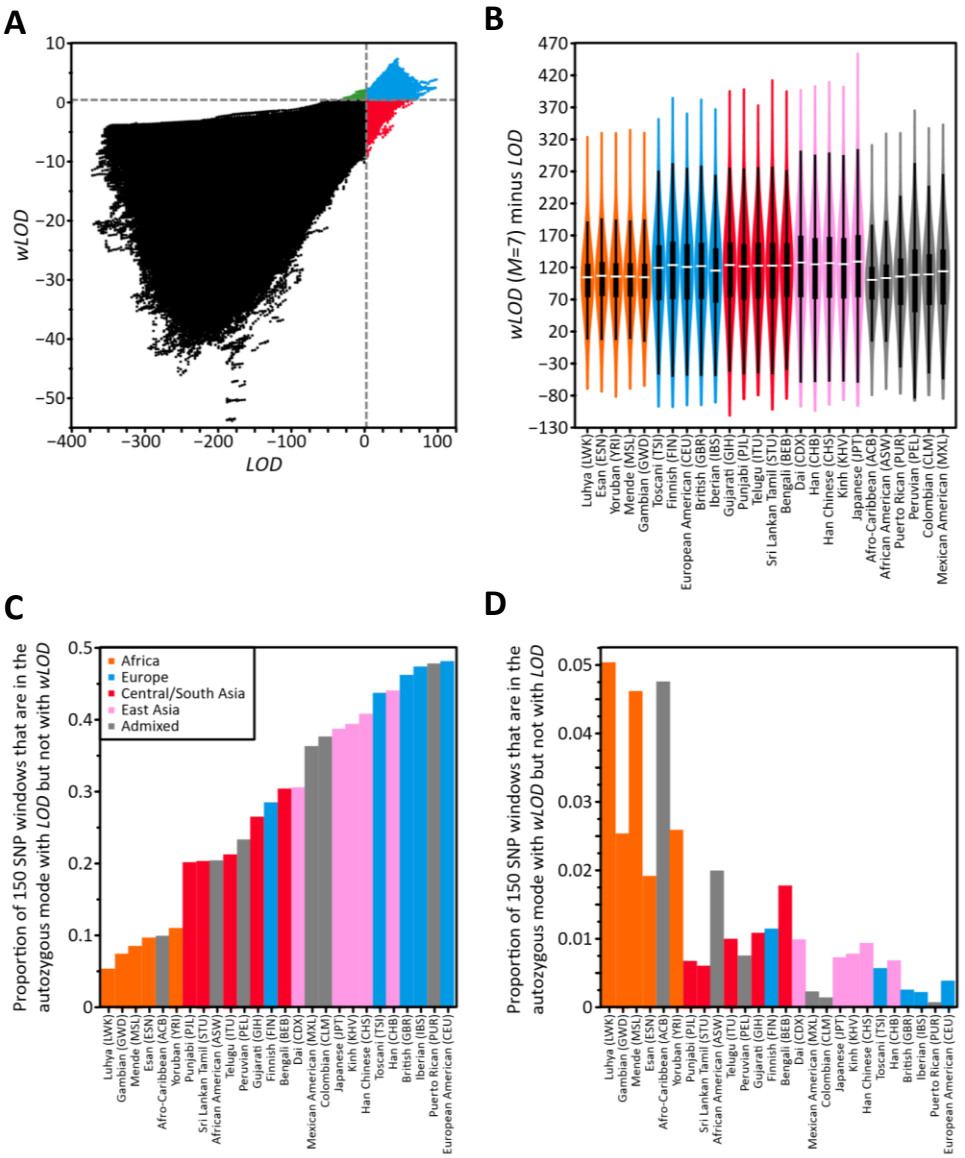


FIGURE 3.5-DIFFERENCE IN PER-WINDOW SCORES BETWEEN THE LOD AND WLOD ESTIMATORS

A) Scatterplot compares per-window LOD and wLOD scores across all individuals in the European American (CEU) population at a window size of 150 SNPs in the Omni2.5 dataset. Points colored black or blue represent windows for which LOD and wLOD methods agreed: the majority of considered windows were in the non-autozygous mode (black) and 2,675,059 windows were in the autozygous mode (blue). Next, points colored in green or red are windows for which LOD and wLOD did not agree: 9,885 windows were in the non-autozygous mode using LOD but in the autozygous mode with wLOD (green) and 2,462,843 windows were in the autozygous mode with LOD but the non-autozygous mode with wLOD (red). **B)** Violin plots represent the change in per-window score between wLOD and LOD across all individuals in each population for 150 SNP windows in the Omni2.5 dataset. Each 'violin' contains a vertical black line (25%–75% range) and a horizontal white line (median), with the width depicting a 90º-rotated kernel density trace and its reflection, both colored by the geographic affiliation of the population [388]. For each population, barplots represent the proportion of 150 SNP windows in: **C)** the autozygous mode with LOD but in the non-autozygous mode with wLOD and **D)** in the non-autozygous mode with LOD but in the autozygous mode with wLOD.
Figure S3 in original publication [360].

In contrast, across populations only 0.055–5.015% of all windows in the non-autozygous mode with the LOD estimator were present in the autozygous mode with the wLOD estimator (Figure 3.5D), potentially reflecting false-negative autozygosity signals reported by the LOD estimator as a consequence of heterozygotes in high LD with a larger number of homozygotes that, in one possibility, might reflect genotyping errors. The proportion of windows was highest in most African populations and lowest in most European populations, with broadly similar values observed in Central/ South and East Asian populations. This pattern is the opposite of that observed with the putative false-positive windows above and can also be explained by population differences in the location of the autozygous mode and its shift toward lower scores with the wLOD estimator.

Overall, the much larger numbers of windows transitioning from the autozygous to the non-autozygous mode than vice versa between the LOD and wLOD estimators accord with the expectation that the LOD estimator frequently overestimates the amount of information available in the data leading it to falsely report autozygosity signals particularly in genomic regions with higher levels of LD, while it underestimates the amount of information much less frequently.

3.4.2 ROA inference with the wLOD estimator

To infer ROA with our wLOD method, we must select an appropriate wLOD score threshold above which genomic windows are considered autozygous. A natural threshold above which to consider a window as autozygous is the location of the minimum between the non-autozygous and autozygous modes in its wLOD score distribution [338]. Sample size was not observed to appreciably influence the location of the minimum between the non-autozygous and autozygous modes. However, across 100 random samples of individuals greater consistency in its determination was observed with increasing sample size, particularly compared with sample sizes of less than 10 individuals, indicating that 10 or more individuals should be used to ensure a robust estimate of the threshold is obtained. All windows with wLOD scores above threshold are considered autozygous [338] and overlapping autozygous windows are joined to define ROA.
As each SNP is included in multiple windows (i.e. a SNP is included in 50 different windows at a window size of 50), SNPs near the boundaries of true ROA will be included in both autozygous and non-autozygous windows as the sliding window enters and leaves the ROA. To improve the

accuracy of ROA inferences when using a sliding-window approach, we require a SNP to be covered by a certain proportion of autozygous windows before it is placed within an ROA [365], with an overlap fraction of 0.25 previously recommended for use with the original LOD ROA inference method [401].

3.4.3 Accuracy of the wLOD estimator

Note: The simulation of genetic and true ROA datasets was completed by Dr. Trevor Pemberton.
To evaluate the sensitivity and specificity of the wLOD method to detect ROA in dense genotype data, we simulated 50 independent replicates of genetic data under two demographic scenarios as previously described [371], except that we considered a non-uniform distribution of recombination rates across the simulated chromosomes and allowed all base pairs to be mutable (see 3.3-Methods). Scenario 1 considered a small partially isolated population of constant effective size (Ne = 75) that receives approximately one migrant per generation, simulated for 150 generations (4350 years for a generation time of 29 years [402]. Scenario 2 considered a medium sized closed population (Ne = 500 simulated for 100 generations [2900 years]). Each simulated dataset is 250 Mb long and consistent with the SNV density and length of chromosome 1 in TGP WGS data containing ~750,000 polymorphic single-nucleotide variants (SNVs). For each simulated dataset, the wLOD estimator was applied to windows between 50 and 500 SNPs (in 10 SNP increments), and a proportion of overlapping autozygous windows was used to construct ROA of between 0 and 50% (in 1% increments).

We then calculated three metrics to determine how well wLOD inferred ROA agreed with true ROA reported by the simulation program. wLOD's power was calculated as the proportion of true ROA length overlapped by inferred ROA and the false positivity rate as the proportion of inferred ROA that do not overlap with true ROA. Finally, for the true ROAs detected via wLOD, we calculated the ratio of inferred ROA length to true ROA length. For the latter, ratios greater than one indicate a tendency to over-call ROA by falsely including non-autozygous regions near the boundaries of true ROA, while ratios below one indicate a tendency to under-call an ROA by falsely excluding true autozygous regions near the boundaries of true ROA [401].

A large numbers of false positive ROA calls are made by the wLOD method with a window size of 50 SNPs, however, these decrease markedly with increasing window size and proportion of overlapping windows required during ROA construction (Figure 3.6A). These patterns are consistent with the observation that false positive ROA calls are very small—on average 16.97 kb (standard deviation [SD] = 3.85) with a window size of 50 SNPs—and therefore delineated by a few erroneous autozygous windows that progressively fail to meet the required threshold during ROA calling as the window overlap fraction increases. Once window size reaches ~90 SNPs, the wLOD estimator can distinguish autozygosity from homozygosity-by-chance with good precision. Conversely, numbers of false negative ROA calls increase as a function of window size and overlap fraction (Figure 3.6B). These patterns are consistent with the expectation that as window size increases smaller ROA increasingly go undetected, likely due to the ROAs being spanned by progressively fewer but larger windows. The autozygosity signal for the latter would therefore be increasingly masked by the inclusion of non-autozygous flanking regions in the wLOD score calculation. Similarly, higher overlap fractions also lead to small ROA spanned by just a small number of autozygous windows increasingly going undetected as they fail to meet the required threshold. Nevertheless, overall power to detect ROA with the wLOD method only decreases slightly as window size and overlap fraction increase (Figure 3.6C), consistent with the expectation that at larger window sizes and overlap fractions the sliding window approach will have increasing difficulty in detecting smaller ROA but nonetheless retains high power to detect longer ROA. Finally, ratios of inferred to true ROA length increase as a function of window size and decrease as a function of overlap fraction (Figure 3.6D), reflecting the tendency of the wLOD method to overcall the boundaries of smaller ROA at larger window sizes and smaller overlap fractions with those of longer ROA affected to a much lesser extent. Together, these patterns suggest that the smallest window size and overlap fraction combination where no false-positive ROA are inferred and the average ratio of inferred to true ROA length is approximately one is a suitable point within the parameter space at which to evaluate the sensitivity and specificity of the wLOD method.

To evaluate how SNV density influences the sensitivity and accuracy of ROA inference with the wLOD method, we created three subsets of the simulated WGS datasets that reflect the SNV densities of commonly used human microarray-based genotyping platforms: Illumina's

HumanOmni2.5-8 (125,000 SNVs) and OmniExpress-24 (50,000 SNVs) BeadChips and Affymetrix's GenomeWide Human SNP 6.0 Microarray (80,000 SNVs).

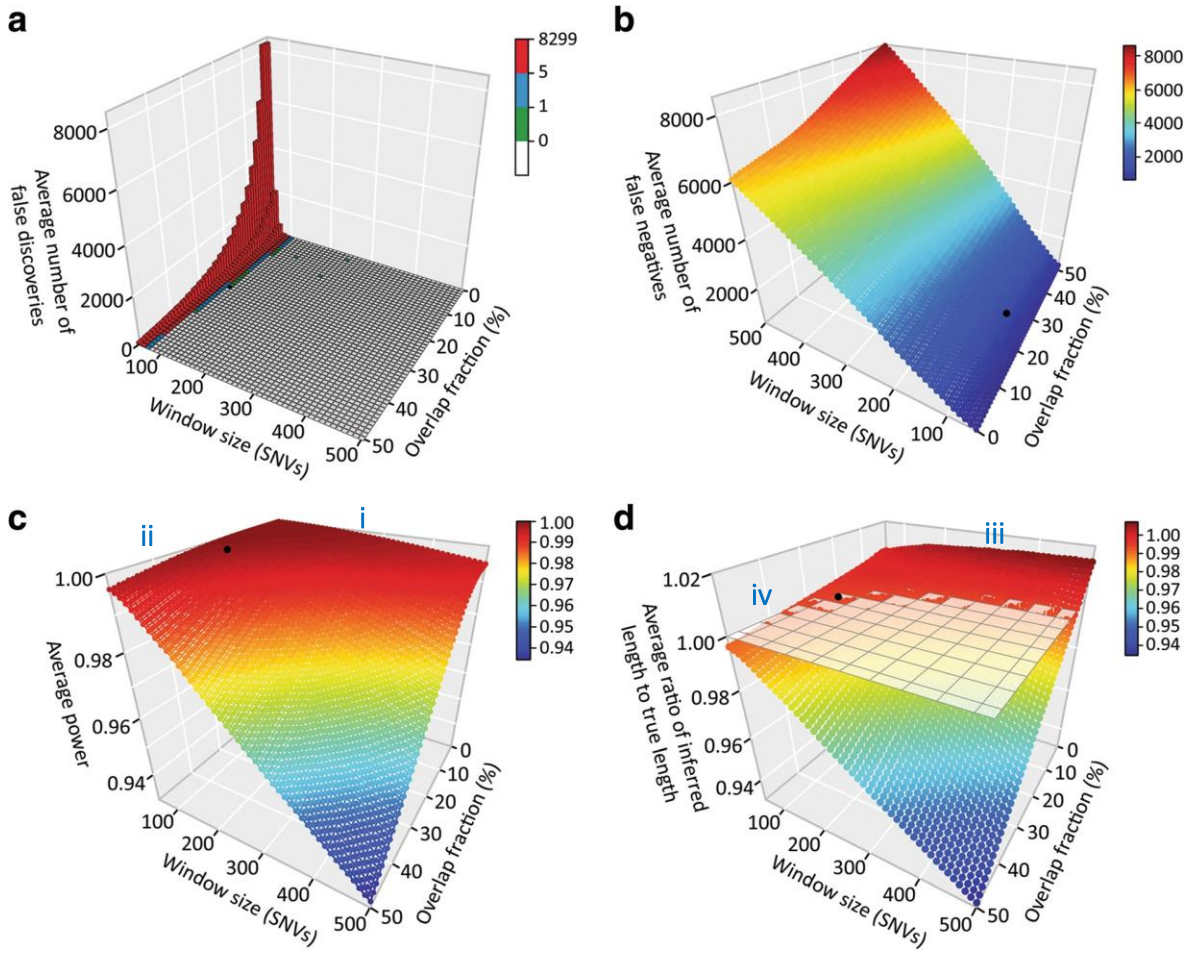


FIGURE 3.6-PERFORMANCE OF THE WLOD METHOD ACROSS DIFFERENT WINDOW SIZES AND OVERLAP FRACTIONS

Three dimensional plots illustrating the performance of wLOD using Scenario 1 simulated data and a 750,000 polymorphic SNV dataset. **A:** A 3D barplot demonstrates the decrease in false-discovery calls made by wLOD (false positives) with increasing window size and % overlap fraction (proportion of overlapping windows required during ROA construction).
B: This 3D scatterplot demonstrates the increase in false-negative calls made by wLOD with increasing window size and % overlap. The layout is similar to **A** except the y-axis represents false negatives (instead of false positives) and the axes for window size and % overlap are reversed.
C: This plot demonstrates the slight reduction of average power with increasing windows size (see i) and % overlap (see ii). **D:** This plot demonstrates the relationship between inferred : true ROA length ratio with window size and % overlap: ratios of inferred to true ROA length increase with increasing window size (see iii) and decrease with increasing overlap fraction (see iv).
In each plot, the point in black marks the approximate location where window sizes and overlap fractions are at their smallest and the inferred: true ROA length ratio is equal to 1 (ie, the average number of falsely discovered ROA is 0).
Figure 2 in original publication [360].

In addition, we included subsets with SNV densities consistent with the genotyping platforms used by ROA studies in cattle and dogs: Illumina's Bovine HD (80,000 SNVs) and Canine HD (18,000 SNVs) BeadChips. ROA inference accuracy was evaluated exactly as described above for the WGS datasets at the optimal window size and overlap fraction determined separately for each SNV density and demographic scenario (Appendix Table 5). Interestingly, optimal window size varied only slightly across the different SNV densities, lying between 60–130 SNPs and 70–120 SNPs for scenarios 1 and 2, respectively, but nevertheless increasing as a function of SNV density. The optimal window overlap fraction did however vary more widely, increasing as a function of SNV density and lying between 7–37% and 5–32% for scenarios 1 and 2, respectively.

As would be expected, the power of the wLOD method to detect ROA increases as a function of ROA length and the density of SNV in the genetic dataset (Figure 3.7). While ROA longer than 1 Mb are captured extremely well (> 99.7%) at all SNV densities explored, the detection of ROA shorter than ~1 Mb decreases appreciably as a function of SNV density. Nevertheless, even with only ~ 18,000 SNVs (1 SNV every ~ 14 kb) wLOD captures 96.3% and 89.0% of ROA under scenarios 1 and 2, respectively, with this increasing to 99.9% for both scenarios with 750,000 SNVs (1 SNV every ~ 333 bp).

False discovery rates, on the other hand increase dramatically with decreasing SNV density, particularly for smaller ROA (Figure 3.7) where they jump from 0.0045 to 0.0069 with 750,000 SNVs to 0.0445 and 0.1362 with 18,000 SNVs for scenarios 1 and 2, respectively, while longer ROA are much less affected: 0.0010 and 0.0001 with 750,000 SNVs increasing to 0.0200 and 0.0495 with 18,000 SNVs for ROA $\geq$ 5 Mb, respectively. It should be noted that these false discovery rates are solely the result of over-calling true ROA and not erroneous ROA calls. This is reflected in the ratios of inferred to true ROA length (Figure 3.7) that increase with decreasing SNV density, particularly for smaller ROA, and approach—but never quite reach—one with increasing ROA length.

Overall, these findings indicate that the wLOD method is well powered to detect ROA with high sensitivity and good specificity at a wide range of SNV densities that are consistent with WGS as well as popular microarray-based platforms that are commonly used in human and non-human

studies of ROA, and in particular long ROA that are of interest in studies of Mendelian and complex diseases and traits. In the simulations, both the optimal window size and the optimal overlap fraction increased logarithmically as a function of SNV density (R^2 = 0.9814 and R^2 = 0.8868, respectively, when considering their averages across scenarios). Fitting these averages against the natural logarithm of average SNV density *D* across all 50 replicates of their respective SNV subset, this suggests that as a rule of thumb future studies apply the wLOD method at a window size equal to $16.400 \times \log_e(D) + 218.020$ and an overlap fraction equal to $0.0736 \times \log_e(D) + 0.8063$. Based upon these equations, and calculating SNV density as the number of autosomal SNPs on the microarray divided by the total length of the target species' autosomal genome, guideline settings for window size and overlap fraction with the commonly used human and non-human genotyping microarrays are: 111 SNPs (33%), 103 SNPs (29%), 85 SNP (21%), 81 SNPs (19%), and 59 SNP (9%) for Illumina's HumanOmni5, HumanOmni2.5, Bovine HD, OmniExpress, and Canine HD BeadChips, respectively, and 85 SNPs (21%) for the Affymetrix Genome-Wide Human SNP 6.0 Microarray. Considering the range of autosomal SNVs observed in the WGS data available for the 26 populations in Phase 3 of The 1000 Genomes Project (12–24 million SNV [368]) a window size of 128–140 SNPs and an overlap fraction of 40–45% would be recommended for WGS datasets.

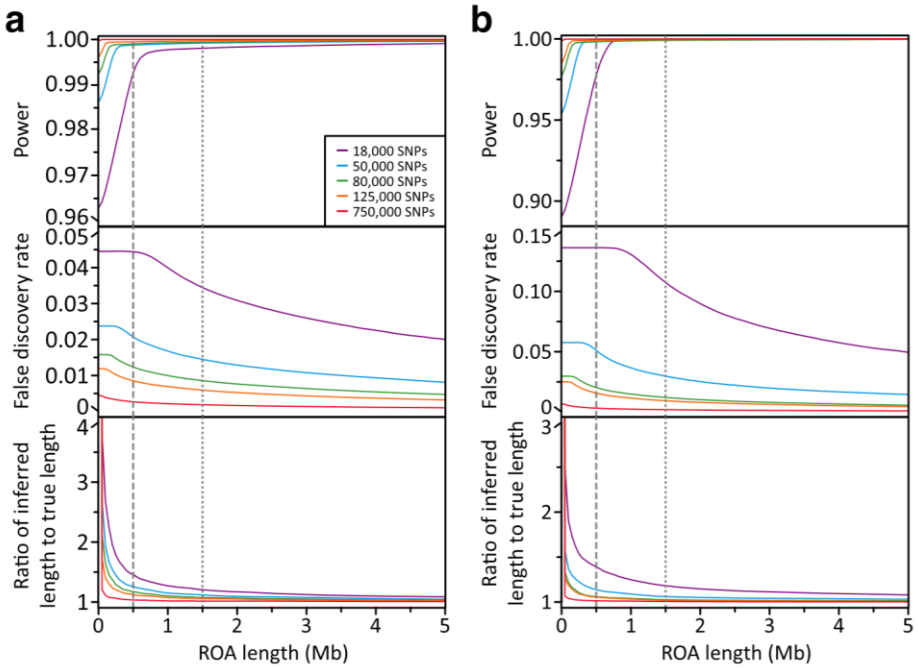


FIGURE 3.7-PERFORMANCE OF THE WLOD METHOD AT DIFFERENT SNV DENSITIES
These plots illustrate the relationship between increased ROA length and: power, false discovery rate and ratio of inferred and true ROA length (top, middle and bottom panels, respectively) across 50 replicate genetic simulation Scenarios 1 (**A**) and 2 (**B**) at five densities of data (see legend). Each comparison was performed at the optimal combination of window size and overlap fraction for that scenario and SNV subset (Appendix Table 5). The gray vertical lines denote 500kb (dashed) and 1.5 Mb (dotted), which represent frequently applied length thresholds used to categorize ROA. ROA lengths < 500kb arise due to LD while ROAs > 1.5Mb in length are indicative of more recent inbreeding (>1.5Mb) in humans [464].
Figure 3 in original publication [360].

3.4.4 Performance of wLOD against existing ROA detection methods

We next evaluated how the power and false discovery rate of the wLOD method compared with those of three current methods designed specifically to detect ROA in dense genotype data using the datasets simulated above: the original LOD method, the commonly used genotype counting method implemented in PLINK [365], and the hidden Markov model (HMM) method of BCFtools [377]. Though numerous others exist, PLINK and BCFtools have been previously reported to outperform several methods [376][377] and as such are selected here to represent different mathematical approaches to the detection of homozygosity.

Since the false discovery and boundary placement properties of the sliding-window-based LOD and PLINK methods would be expected to differ from those of the wLOD method due to their different underlying models, separately for each dataset we identified the optimal window size and overlap fraction for the LOD method (not shown), PLINK (always a window size of 50 SNPs and overlap fraction of zero) and BCFtools (applied using default settings). It should be noted that we apply PLINK here to the entire dataset to provide an "all else being equal" comparison with the other methods; however, it is generally recommended to apply it to minor allele frequency (MAF) and LD-pruned datasets to minimize the number of spurious ROH calls [376][365]. Thus, its power and false-positive rates of ROA detection reported here are expected to be inflated and deflated, respectively, relative to those after MAF- and LD-pruning.

For both simulated scenarios, all four methods were able to detect an average >99.5% of ROA using 750,000 SNVs which is representative of the density of SNVs observed in WGS data (Figure 3.8A and D). Nevertheless, the wLOD method outperformed both the original LOD method as well as PLINK and BCFtools, particularly at shorter ROA lengths. Interestingly, power with BCFtools became increasingly erratic at longer ROA lengths, most noticeably in scenario 1 (small partially isolated populations). However, while the wLOD method had a lower false discovery rate than the LOD method, it was notably higher than that of BCFtools and PLINK. It should be noted that this elevated false positive rate solely reflects the overcalling true ROA due to the sliding-window approach employed and not erroneous ROA calls. This overcalling can be reduced using a more stringent overlap fraction (at the expense of power to detect short ROA). Despite this, average ratios of inferred to true ROA length were broadly similar across the wLOD, LOD, and BCFtools methods, where they are highest for extremely short ROA and decrease exponentially with increasing ROA length until they approach—but never quite reach—one, although ratios with BCFtools were marginally lower than those with the wLOD and LOD methods in scenario 2. Conversely, average ratios with PLINK decreased noticeably as a function of ROA length—reaching 0.47 in scenario 1 and 0.81 in scenario 2—consistent with the expectation that as a consequence of its naïve model, PLINK will have a tendency to under-call ROA or return fragmented ROH calls across their span as a function of the distribution of heterozygous genotypes within the ROA, which would be expected to be most numerous near its boundaries.

Overall, these observations would suggest that model improvements implemented in the wLOD estimator (Equation 2) that account for the confounding effects of LD, recombination, and mutation in the autozygosity likelihood calculation provide improved sensitivity and specificity in ROA calling over the original LOD estimator (Equation 1). Additionally, they indicate that the wLOD method's sliding window approach, which combines evidence for autozygosity across multiple SNVs, provides improved sensitivity to detect ROA compared with the HMM method of BCFtools, albeit with slightly decreased accuracy in ROA boundary placement.

When we consider simulated datasets consistent with those of genotyping microarrays we observe similar patterns to those observed with 750,000 SNVs (Figure 3.8 and Figure 3.9). For both scenarios 1 and 2, the wLOD method consistently outperforms the LOD method as well as BCFtools and PLINK in terms of power, particularly at shorter ROA lengths. False discovery rates with the wLOD method are consistently lower than those with the LOD method but remain slightly higher than those with BCFtools, while ratios of inferred to true ROA length remain similar across the wLOD and LOD methods and BCFtools. As SNV density decreases from 750,000 SNVs down to 18,000 SNVs several patterns emerge. First, the difference in power between the wLOD and LOD methods decreases as a function of SNV density (Figure 3.9B and D), disappearing faster under scenario 2 (large, closed populations) than under scenario 1 (small partially isolated populations). These patterns are consistent with the view that in datasets containing fewer SNVs, LD confounds the inference of ROA appreciably less than in datasets containing many SNVs. Consequently, the LD correction implemented in the wLOD estimator (Equation 2) increasingly becomes less important as SNV density decreases, leading the LOD and wLOD estimators to provide broadly similar autozygosity likelihoods. Nevertheless, false discovery rates with the wLOD method are consistently lower than those with the LOD method, in agreement with the expectation that as SNV density decreases the probabilities of unobserved recombination and mutation events between genotyped SNVs increases, with the recombination and mutation corrections implemented in the wLOD estimator (Equation 2) enabling it to better account for these events than the LOD estimator (Equation 1). Second, ratios of inferred to true ROA length with the PLINK method become more similar to those of the other three methods with decreasing SNV density. This pattern is consistent with the expectation that as SNV density decreases, the

number of heterozygous genotypes within ROH will also decrease, allowing PLINK to increasingly detect the entire ROA.

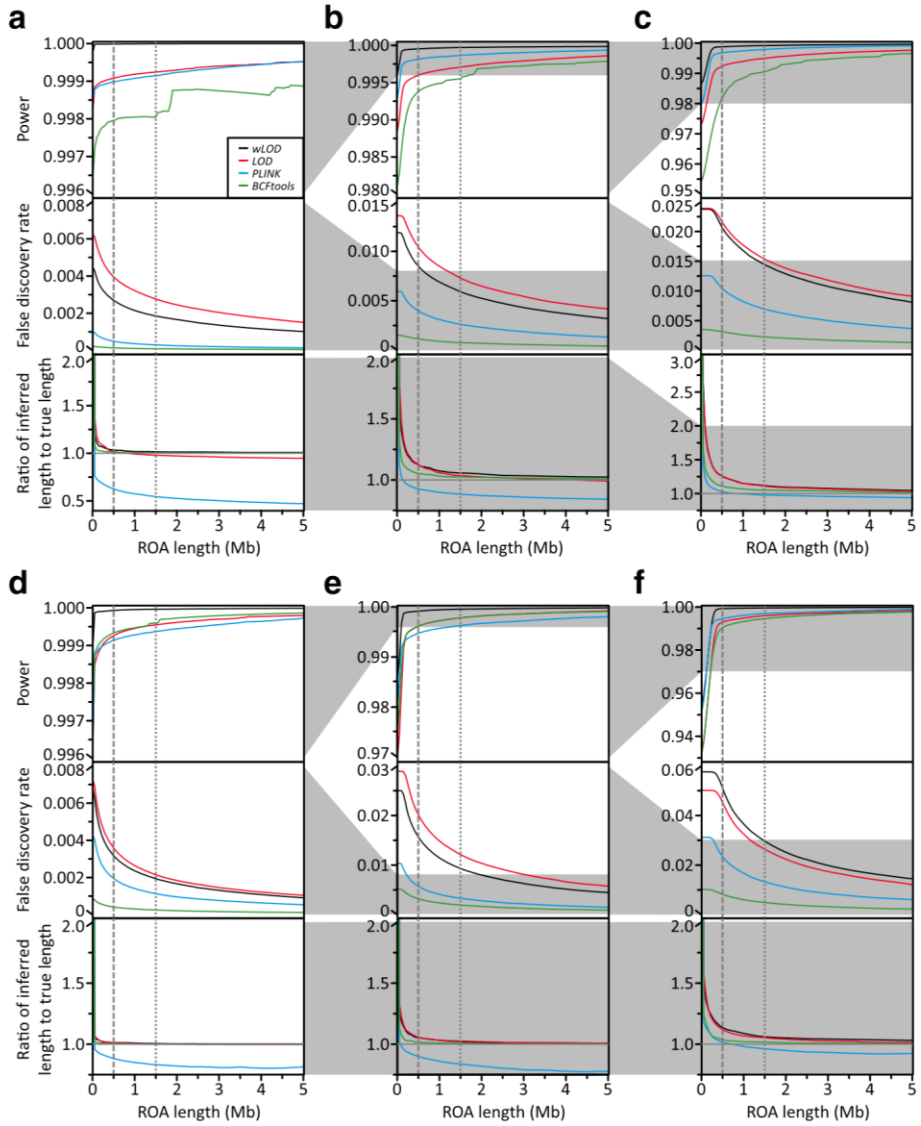


FIGURE 3.8-PERFORMANCE OF WLOD AGAINST EXISTING METHODS (50,000 - 750,000 SNV DENSITY DATASETS)

Six line-graphs illustrate the relationship between increased ROA length and: power, false discovery rate and ratio of inferred and true ROA length (top, middle and bottom panels, respectively) across 50 replicate genetic simulations for three different densities of data. Simulation Scenario 1 is used in **A-C** and Scenario 2 in **D-F**. **A** and **D** use 750,000 SNVs (consistent with WGS), **B** and **E** use125,000 SNVs (Illumina HUmanOmni2.5-8 BeadChip) and **C** and **F** use 50,000 SNVs (Human OmniExpress-24 BeadChip). The gray vertical lines denote 500kb (dashed) and 1.5 Mb (dotted), which represent frequently applied length thresholds used to categorize ROA. ROA lengths < 500kb arise due to LD while ROAs > 1.5Mb in length are indicative of more recent inbreeding (>1.5Mb) in humans [464]. *Figure 4 in original publication* [360].

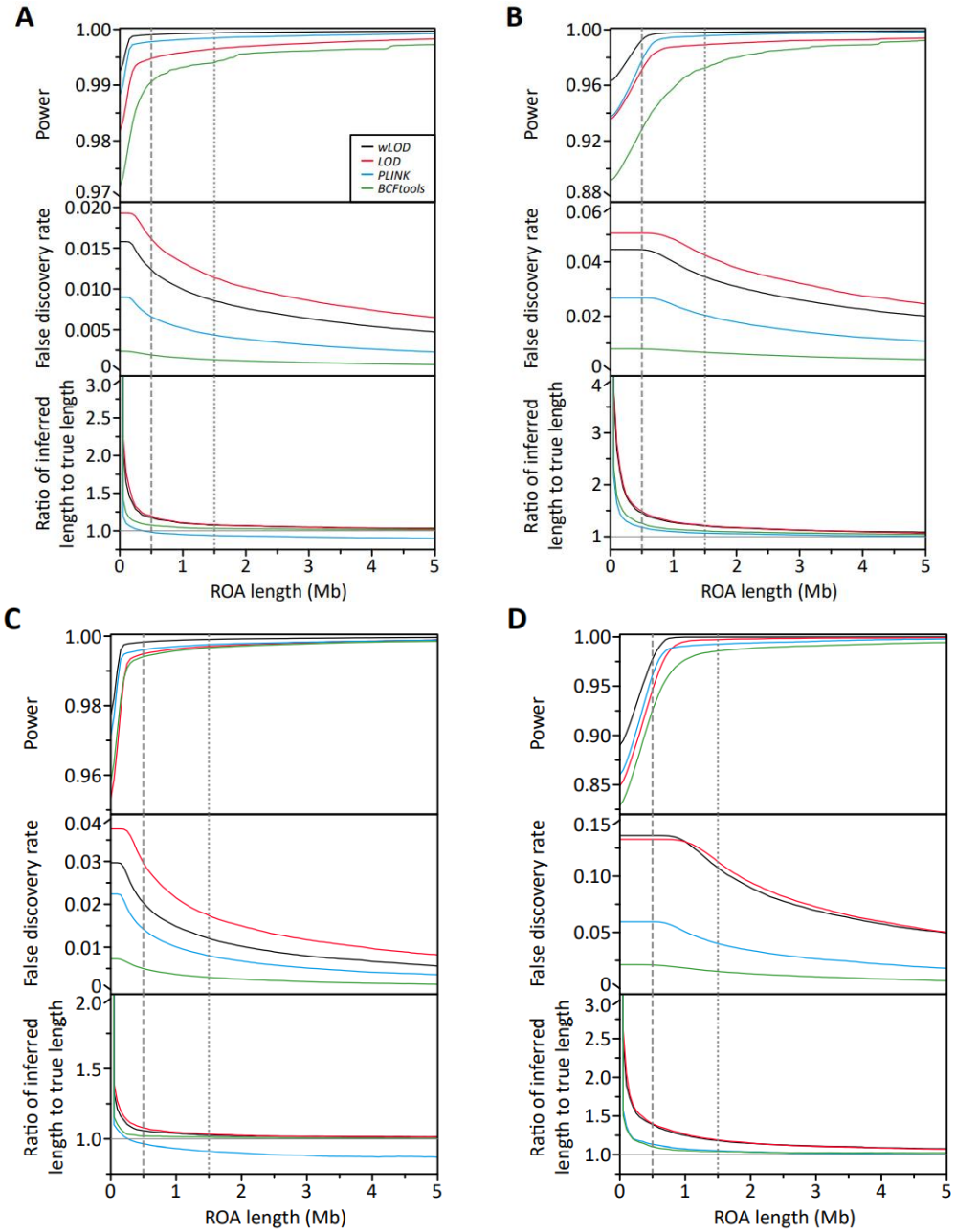


FIGURE 3.9-PERFORMANCE OF wLOD AGAINST EXISTING METHODS (18,000 - 70,000 SNV DENSITY DATASETS)

Four line-graphs illustrate the relationship between increased ROA length and: power, false discovery rate and ratio of inferred and true ROA length (top, middle and bottom panels, respectively) across 50 replicate genetic simulations for two different densities of data. Simulation Scenario 1 is used in **A** and **B** and Scenario 2 is used in **C** and **D**. **A** and **C** use 18,000 SNVs (consistent with Illumina CanineHD BeadChip) while **B** and **D** use 70,000 SNVs (Genome-Wide Human SNP 6.0 Microarray). The gray vertical lines denote 500kb (dashed) and 1.5 Mb (dotted), which represent frequently applied length thresholds used to categorize ROA. ROA lengths < 500kb arise due to LD while ROAs > 1.5Mb in length are indicative of more recent inbreeding (>1.5Mb) in humans [464].

Figure S11 in original publication [360].

Finally, the performance of BCFtools decreases as a function of SNV density, although an appreciable loss of power only manifests when we reach 18,000 SNVs and is more pronounced in scenario 2 than in scenario 1 (Figure 3.9B and D), suggesting that its HMM is sensitive to the effects of extended LD among sparsely distributed SNVs, a situation frequently encountered in closed populations due to elevated levels of general inbreeding. It should be noted, however, that BCFtools was designed for next-generation whole-genome and -exome data analysis and not for sparser microarray-derived genotype datasets, so its decline in performance in such datasets is to be somewhat expected.

Contrary to expectations based on frequent discrepancies in the autozygosity status of windows with the wLOD and LOD estimators in the TGP Phase 3 populations (Figure 3.5), in our simulated datasets the wLOD method only provided modest improvements in power and false discovery rate over the original LOD method (Figure 3.9). While the proportion of windows autozygous with the wLOD estimator but not the LOD estimator in the simulated datasets is similar to that observed in the TGP Phase 3 populations (Figure 3.5D), the proportion of windows autozygous with the LOD estimator but not the wLOD estimator is about two orders of magnitude lower than the values observed in TGP Phase 3 populations (Figure 3.5C). Thus, while we observe the expected sensitivity gain (via a reduction in the contribution of occasional heterozygotes within ROH with the wLOD estimator) we do not observe the expected inflation in LOD scores due to the confounding effects of LD among genotyped positions that leads to increased false positive ROA calls. Based on their underlying models, we would expect the LOD (Equation 1) and wLOD (Equation 2) estimators to provide highly similar inferences in situations where autozygosity patterns align almost perfectly with LD patterns among genotyped SNVs and are investigated with a sufficiently high density of SNVs that the probabilities of unobserved mutation and recombination events are effectively zero. The most parsimonious explanation for the surprisingly high similarity of ROA calls made by the LOD and wLOD methods in the simulated datasets is therefore that LD patterns in these simulated datasets do not faithfully recapitulate the complexity of those found in real populations who have experienced much more complex histories than those simulated here, limiting the impact of the LD correction (Equation 3) incorporated into the wLOD estimator. We therefore expect to observe appreciably greater improvements in the sensitivity and

specificity of ROA calls with the wLOD method compared with the LOD method in real genetic data than in our simulated datasets.

3.4.5 Effect of genotyped SNV density on ROA inference in real data

We have shown the wLOD method to be well powered to detect ROA in genetic datasets consistent with WGS and microarray-based genotyping, and to outperform a number of existing methods in terms of power, although the overcalling of ROA due to its sliding window approach creates slightly higher rates of false discovery than a recently reported HMM model approach. While our simulations suggest that the wLOD method has >99.8% power to detect ROA longer than 1 Mb across SNV densities that are consistent with those frequently used in human population- and disease-genetic studies (Figure 3.8), they do not capture the diversity of historical events and sociogenetic processes that have influenced genomic autozygosity patterns in contemporary worldwide human populations. Thus, we next sought to evaluate how robust ROA inferences are among genotype datasets created via WGS, WES as well as commonly used microarrays using the TGP Phase 3 data. For the WGS, Omni2.5 and OmniExp datasets we applied the wLOD method at the window size and overlap fraction suggested by our simulation analyses given their average SNV density across populations: 125 SNPs (40%), 95 SNPs (25%), and 80 SNPs (18%), respectively. As the SNV density of the WES dataset closely resembles that of the WGS dataset in the genomic regions it covers, we used the same window size and overlap fraction settings in both the WES and WGS datasets. For all datasets, μ was set to 1.18×10^{-8} [373] and M was set to 7, a conservative value broadly reflecting the average of effective population size estimates for populations included in the TGP [362][390][392]. For the Omni2.5 and OmniExp datasets ε was set to 4.71×10^{-4}, the average rate of discordance across samples between genotypes in our Omni2.5 dataset and those obtained for 1,693 of the 2,436 individuals directly with the Illumina HumanOmni2.5 BeadChip [368], while in the WGS and WES datasets ε was instead set separately for each genotype as one minus its reported likelihood. This has the potential to improve the accuracy of ROA calls in next-generation sequencing (NGS) datasets by incorporating the uncertainty of each genotype call into the wLOD score calculation, an important potential source of erroneous ROA calls in the context of their often higher and more variable per-genotype error rates compared with microarray-derived datasets [369][406]. As such, autozygous windows comprised of SNVs with low quality genotypes have a greater chance of being false-positive

signals than those with higher quality genotypes, while low quality heterozygous genotypes—that in one possibility may be genotype calling errors—located in runs of higher quality homozygous genotypes have the potential to mask true autozygous signals.

Comparing ROA inferred in the WGS and Omni2.5 datasets, we find Omni2.5 ROA to be frequently longer than their corresponding WGS ROA and in most cases to completely encompass the WGS ROA (Figure 3.10A). The magnitude of their length discrepancies decreases with increasing ROA length, consistent with the expected effects of decreased SNV density on the accuracy of inferred ROA boundaries. In addition, while all Omni2.5 ROA are present in the set of WGS ROA, the reverse is not true (Figure 3.10B). Many short ROA (< 500 kb) inferred in the WGS dataset are not found in the Omni2.5 dataset, with the fraction of missing ROA decreasing with increasing distance from Africa, reflecting the effect of increasing LD [393][394] on our ability to infer shorter ROA with the sparser set of SNVs in the Omni2.5 dataset. Concordance between the WGS and Omni2.5 datasets for intermediate (500 kb to 1.5 Mb) and long (> 1.5 Mb) ROA is generally high, although in many populations the fraction of WGS ROA missing in the set of Omni2.5 ROA remains nontrivial. These fractions generally increase as a function of distance from Africa, likely reflecting the reduction in haplotype diversity with decreasing genetic diversity [392]–[396] decreasing our ability to distinguish autozygosity from homozygosity by chance, particularly over extended genomic regions when genotypes are only available for a fixed set of SNVs that were selected for their generally high level of polymorphism worldwide.

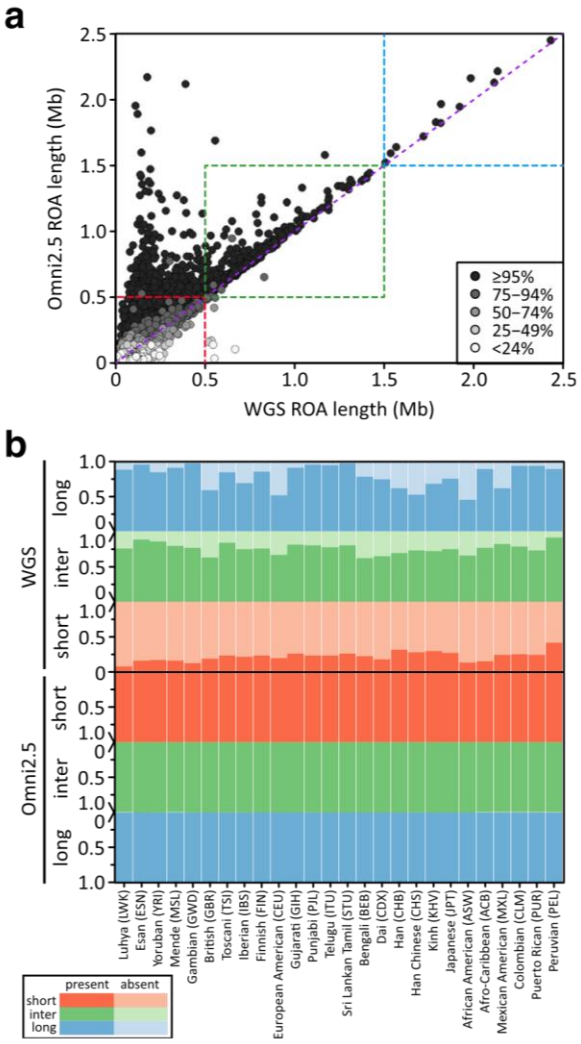


FIGURE 3.10-CONCORDANCE OF ROA INFERRED IN THE WGS AND OMNI2.5 DATASETS

A) A scatterplot compared the length of each WGS ROA with that of its corresponding Omni2.5 ROA in the European American (CEU) population. Each point is shaded according to the proportion of the WGS ROA that overlaps the Omni2.5 ROA. Dotted lines in red, green and blue demarcate the boundaries for short, intermediate and long class ROAs, respectively **B)** A series of stacked barplots compare the proportion of Omni2.5 ROA found in WGS ROA (top panel) and the proportion of WGS ROA found in Omni2.5 ROA (bottom panel) for each of the 26 populations. ROAs present in both datasets are colored in the darker shade, while ROAs absent from the dataset in question are in the lighter shade. Short ROAs (<500 kb) are in red, intermediate ROAs (500 kb to 1.5 Mb) in green, and long ROAs (>1.5 Mb) in blue. *Figure 5 in original publication* [360]

Similar patterns are observed when we compare ROA inferred in the Omni2.5 and OmniExp datasets, where almost all OmniExp ROA are present in the set of Omni2.5 ROA (Figure 3.11B) and encompass their generally shorter corresponding Omni2.5 ROA (Figure 3.11A). While many short ROA inferred in the Omni2.5 dataset are not found in the OmniExp dataset, both intermediate and long ROA are captured extremely consistently between the two datasets despite their different SNV densities. Likewise, when we compare ROA inferred in the WGS and WES datasets, almost all WES ROA are present in the set of WGS ROA (Figure 3.12B) and tend to encompass their generally shorter corresponding WGS ROA (Figure 3.12A). The numbers of long ROA inferred in the WGS dataset but not the WES dataset is similar to the pattern observed in Figure 3.10 for long ROAs found in the WGS dataset as compared to Omni2.5, which suggests that the non-

uniform and often sparse distribution of SNVs in the WES dataset does not impact long ROA inference more than would be expected following a general reduction in SNV density.

Overall, these findings are consistent with the higher density of SNVs in the WGS dataset and the presence of many more rare and low-frequency SNVs detected by NGS compared with microarray-based genotyping platforms—which are particularly informative about autozygosity under our likelihood model (Figure 3.2A) —greatly improving our ability to infer ROA. Nevertheless, many long ROA that are of interest in Mendelian and complex disease studies are well captured by the SNV sets included on the Omni2.5 and OmniExp BeadChips. However, the sparse and non-uniform genomic distribution of SNVs in the WES dataset creates difficulties when inferring short and intermediate ROA with the wLOD method, despite the presence of rare and low-frequency SNVs, while long ROA are instead captured almost as well as with genotyping microarrays. We therefore do not recommend using the wLOD method to infer ROA in WES datasets generated by future studies.

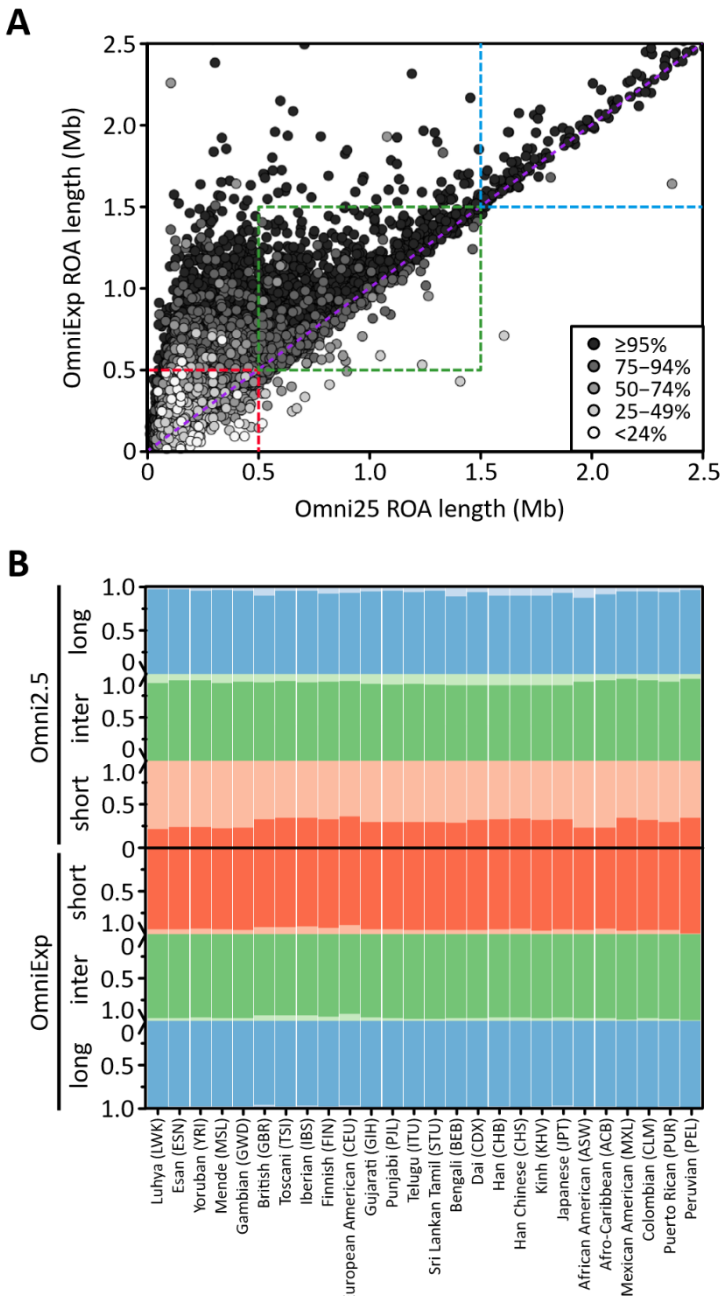


FIGURE 3.11-CONCORDANCE OF ROA INFERRED IN THE OMNI2.5 AND OMNIEXP DATASETS

A) A scatterplot compares the length of each Omni2.5 ROA with that of its corresponding OmniExp ROA in the European American (CEU) population. Each point is shaded according to the proportion of the Omni2.5 ROA that overlaps the OmniExp ROA. **B)** A series of stacked barplots compare the proportion of OmniExp ROA found in Omni2.5 ROA (top panel) and the proportion of Omni2.5 ROA in OmniExp ROA (bottom panel) for each of the 26 populations. Short (red), intermediate (green) and long (blue) ROAs are compared separately; ROAs present in both datasets are shown in the darker shade while ROAs absent from the respective dataset are in a lighter shade.

Figure S12 in original publication [360].

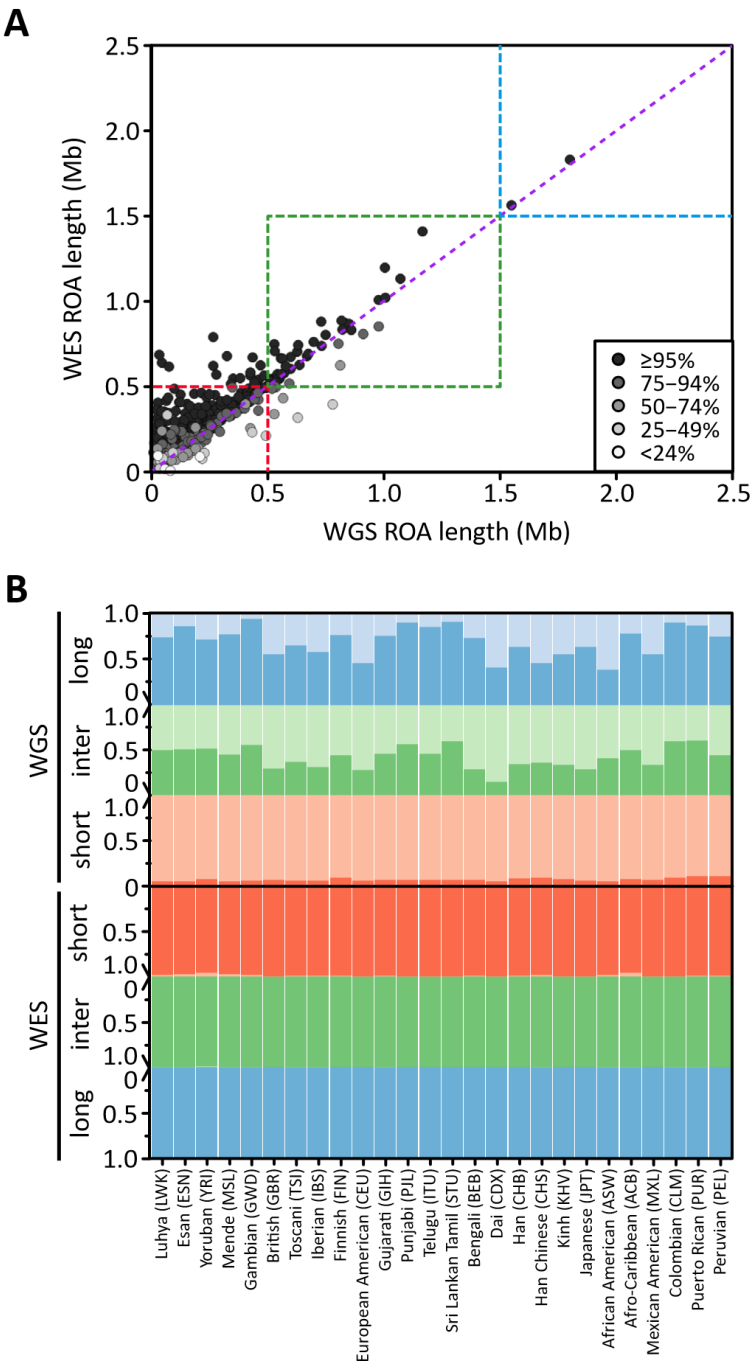


FIGURE 3.12-CONCORDANCE OF ROA INFERRED IN THE WGS AND WES DATASETS

A) A scatterplot compares the length of each WGS ROA with that of its corresponding WES ROA in the European American (CEU) population. Each point is shaded according to the proportion of the WGS ROA that overlaps the WES ROA. **B)** A series of stacked barplots compare the proportion of WES ROA found in WGS ROA (top panel) and the proportion of WGS ROA in WES ROA (bottom panel) for each of the 26 populations. Short (red), intermediate (green) and long (blue) ROAs are compared separately; ROAs present in both datasets are shown in the darker shade while ROAs absent from the respective dataset are in a lighter shade.
Figure S13 in original publication [360].

3.4.6 Classification of ROA

ROA of different lengths reflect homozygosity for haplotypes inherited IBD from common ancestors at different depths in an individual's genealogy: longer ROA most likely arise due to recent ancestors and shorter ROA due to more distant ancestors. We previously advocated the classification of ROA into *G* length-based classes using a Gaussian mixture model approach applied on their physical map lengths (in bp). This approach groups ROA based upon their supposed age [338]: A) short ROA that measure tens of kilobases and that are of the length at which baseline patterns in LD in a population produce autozygosity through the pairing of two copies of the same ancient haplotype, B) intermediate length ROA that measure hundreds of kilobases to several Mb and that are likely the result of background relatedness -recent but unknown kinship between parents due to limited effective population sizes─and C) long ROA that measure multiple megabases and are likely the result of recent parental relatedness (e.g. consanguinity). The choice of $G = 3$ was motivated by the observation that at $G > 3$, the additional classes were not discrete; that is, they were encompassed by one of the existing classes (Figure 3.13A and C).

That being said, the classification approach described above is limited by the imperfect correlation between physical and genetic map lengths (Figure 3.14). Genetic map lengths provide a more accurate representation of the relationship between ROA length and age which is not confounded by the non-uniform genomic distribution of recombination rates [404]–[406]. If we instead classify ROA based on their genetic map length (in cM) using a Gaussian mixture model we find that regardless of the number of classes considered, they are always discrete (Figure 3.13B and D). This would suggest that the original loss of discreteness when classifying based upon physical map length may reflect the confounding effects of physically long but genetically short (and vice versa) ROA on the overall length distribution. Nevertheless, regardless of whether physical or genetic map lengths are used the overall pattern of fit with increasing class number remains highly similar (Figure 3.14 A and B), where Bayesian Information Criterion (BIC) likelihoods plateau at around G = 5 with the WGS and Omni2.5 datasets and at around $G = 4$ classes with the OmniExp dataset (not shown). The smaller class number for the OmniExp dataset compared with the WGS and Omni2.5 datasets is consistent with the expectation that smaller ROA will be poorly captured by its sparser set of SNVs, ultimately leading to the loss of the shortest ROA class detected in the

WGS and Omni2.5 datasets. Note that for all populations the maximum BIC likelihood is reached at $G > 5$. Future studies investigating fine scale ROA patterns may wish to consider values of G at which BIC is maximized, however for illustrative purposes we consider $G = 5$ here since the increase in BIC at $G > 5$ is small.

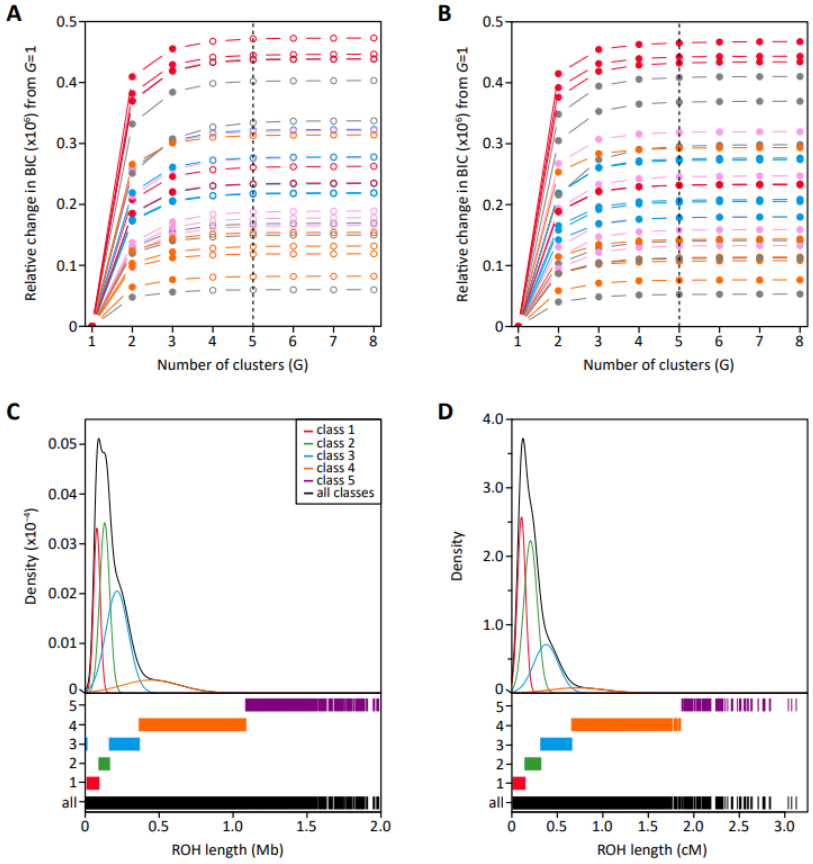


FIGURE 3.13-PROPERTIES OF GAUSSIAN MIXTURE MODEL CLASSIFICATION OF ROA

A and **B** illustrate the observed increases in BIC (relative to the BIC obtained at G=1) with increasing number of clusters for Gaussian mixture model classifications for **A)** physical map lengths and **B)** genetic map lengths. Lines are colored by the geographic affiliation of the population as in Figure 3.5 and Table 3.2. Unfilled circles denote values of G where ROA classes were not discrete (encompassed by another class) while filled circles instead represent G values for which ROA classes were discrete. The black dashed line at G=5 represents the point at which the relative change in BIC plateaued for both physical and genetic map lengths.

C and **D** provide example ROA classifications identified in the CEU population based upon **C**) physical map lengths and **D**) genetic map lengths using G=5. The top panel shows the Gaussian kernel density estimations of the ROA size distribution separately for each class and for all ROA (in black). The bottom panel shows the length of the ROAs and their inferred class assignments. Though all ROA lengths were used in these analyses, plots in **C** and **D** limit the range of the x-axis to ROAs < 2 Mb (**C**) and < 3.5 cM (**D**).

Figure S14 in original publication [360].

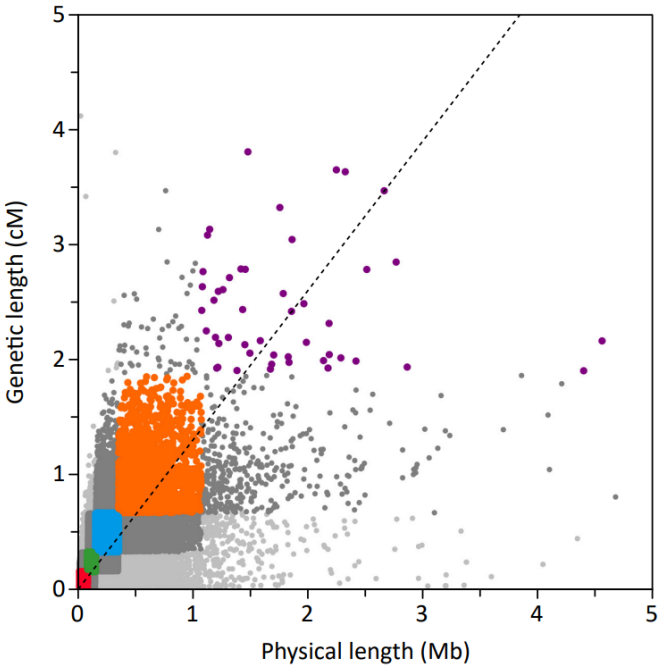


FIGURE 3.14-CORRELATION BETWEEN PHYSICAL AND GENETIC MAP LENGTHS OF ROA
A scatterplot comparing the physical and genetic map lengths of all ROA identified in individuals from the CEU population via the Omni2.5 dataset. Each point represents an ROA. If ROAs were placed in the same class, when classified using their physical and genetic map lengths, the point is colored as in Figure 3.13C (class 1 – red, class 2 – green, class 3 – blue, class 4 -orange and class 5 – purple). ROA whose class differed by one step are shown in dark grey, and ROA whose class differed by more than one step are shown in light grey. The black dashed line represents the commonly used conversion of 1.3 cM per Mb [372].
Figure S15 in original publication [360].

When considering a 5-class classification scheme, the longest class (G = 5) contains ROA that likely arise from recent parental relatedness and the penultimate longest class (G = 4) contains ROA that likely arise from recent population processes, while the shortest classes (G = 1–3) contain ROA arising through the pairing of two copies of much older haplotypes that have common ancestors at different times in the distant past. Sample size was observed to have a greater effect on ROA classification (Figure 3.15) than on wLOD score threshold, with the proportion of ROA whose classification differed from that assigned when all available individuals are used decreasing as a function of sample size.

Importantly, the proportion of misclassified ROA decreased with increasing ROA class, with those in the longest class (G = 5) infrequently misclassified (mean = 0.052 with SD = 0.029 across all 26 populations at a sample size of 25) while those in shorter classes were more frequently affected

(mean = 0.092 with SD = 0.046, mean = 0.091 with SD = 0.045, mean = 0.083 with SD = 0.045, and mean = 0.068 with SD = 0.042, for G = 4 to 1, respectively).

These observations indicate that sample size is an important factor when classifying ROA using a Gaussian mixture model, but in general samples sizes of at least 25 individuals should provide reasonably robust classification of ROA using this approach, particularly longer ROA that are of interest in genetic studies on Mendelian and complex diseases.

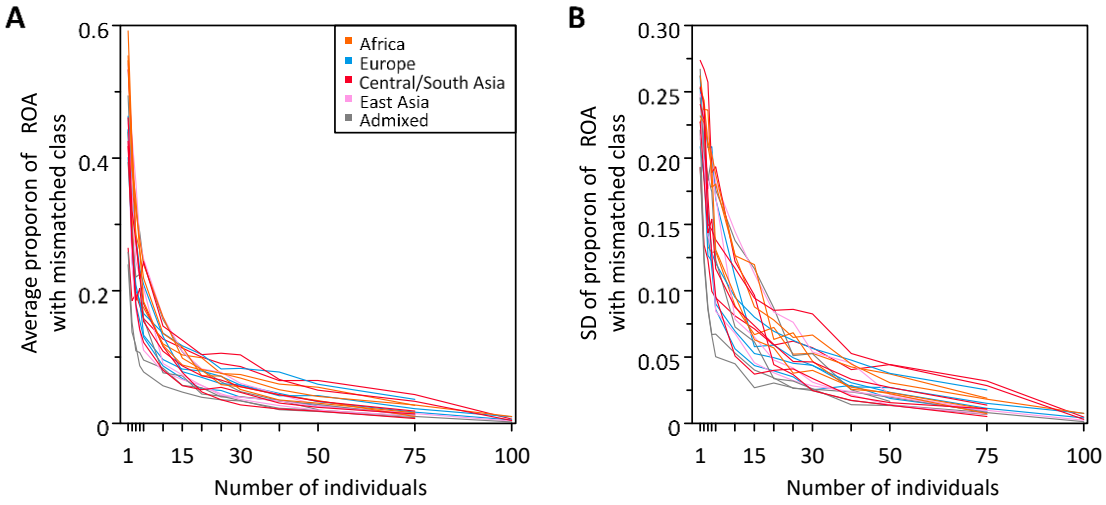


FIGURE 3.15-EFFECT OF SAMPLE SIZE ON ROA LENGTH CLASSIFICATION
A) The relationship between number of individuals and the average proportion of ROA whose classification at a given sample size differs from the obtained class **B)** The relationship between number of individuals and the standard deviation across 100 randomly sampled replicates. The Omni2.5 dataset was used for this analysis. Lines are colored by the geographic affiliation of the population: Africa -orange; Europe -blue; Central/South Asia -red; East Asia -pink; admixed -grey.
Figure S16 in original publication [360].

3.4.7 Geographic properties of the wLOD method

We have shown the wLOD method to be well powered to detect ROA in genetic datasets consistent with WGS and microarray-based genotyping, while our investigation of a Gaussian mixture model approach for ROA classification based upon their genetic map lengths indicates the presence of five ROA classes in Phase 3 of the TGP, a higher number than was used in our earlier study which used the Human Genome Diversity Panel (HGDP) and International HapMap Project (Phase 3) populations that used a microarray derived dataset and classified ROA based upon their physical

map lengths [338]. We thus next explored the population-genetic properties of the wLOD estimator and its inferred ROA.

3.4.8 Evidence of separate endogamic and consanguinity autozygosity signals in Asian Indians

As window size increased, a third mode appeared in the wLOD score distribution (dividing the right-hand autozygous mode in two (Figure 3.16A) for the following populations: East Asian Dai (CDX), Gujarati (GIH), Telugu (ITU), Punjabi (PJL) and Sri Lankan Tamil (STU); the latter 4 being Asian Indian populations. While a third mode also appeared in the wLOD score distribution of the 5th Asian Indian population present in this dataset, the Bengali (BEB), it was not as well defined as those of the other populations. As window size increased further, the area under both autozygous modes decreased until the left-hand autozygous mode disappeared followed sometime later by the right-hand autozygous mode. Notably, the distributions of all other populations in our dataset did not develop this third mode, and trimodality was not observed in the distribution of LOD scores for any of the 26 populations. The appearance of a trimodal distribution in these six populations potentially reflects the effects of two distinct cultural processes that occur in India and among the Dai: consanguinity [407][264] and endogamy [408][409]. In this scenario, the righthand autozygous mode represents ROA due to consanguinity that are enriched for alleles rare in the general population that segregate within inbred families, while the left-hand autozygous mode represents ROA due to endogamy that are enriched for alleles present at low frequency in the general population that segregate within specific endogamic groups. Compatible with this hypothesis, the three populations with the strongest trimodal pattern (STU, ITU, and DAI) have higher reported frequencies of consanguinity (38.2%, 30.8%, and 21.3% than those with weaker trimodal patterns (BEB, 5.0% ; GIH, 4.9%; PJL, 0.9%) [407][264]. For example, the consanguinity-associated mode of the ITU is much larger than the endogamy-associated mode, while the reverse is true for the GIH (Figure 3.16A), consistent with consanguinity being the primary force generating ROA in the ITU while endogamy is the dominant force in the GIH. To the best of our knowledge, none of the other populations included in Phase 3 of the TGP practise endogamy; consequently, we do not observe the emergence of a separate endogamy-associated autozygous mode in their wLOD score distributions.

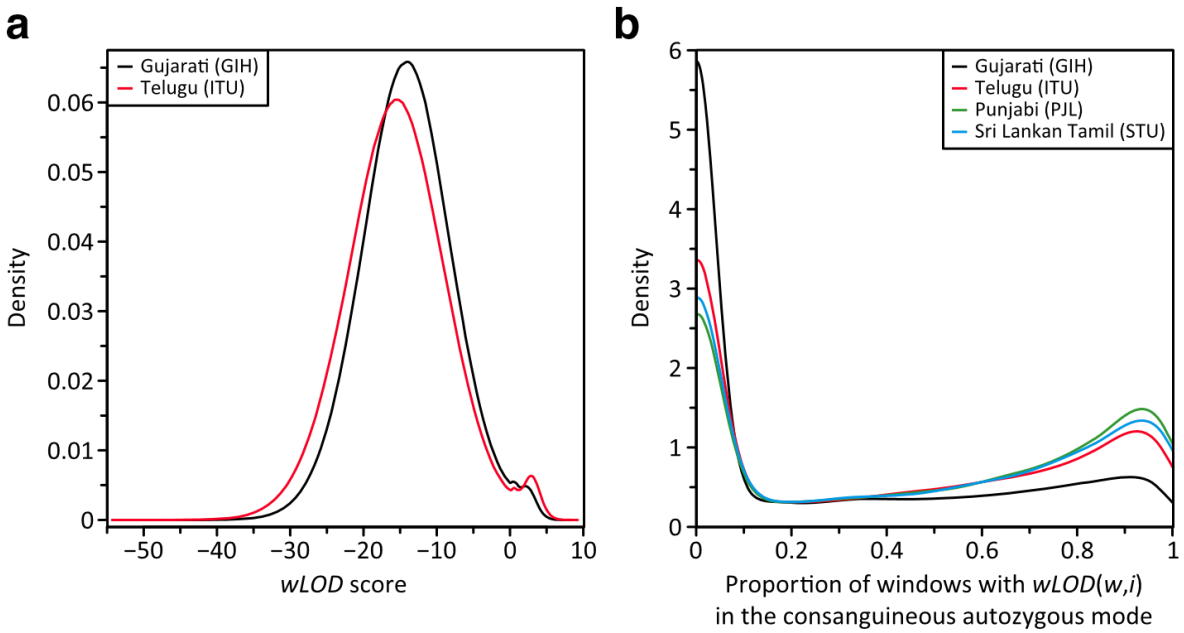


FIGURE 3.16-INFLUENCE OF CULTURAL PRACTICES ON THE DISTRIBUTION OF WLOD SCORES

A) Gaussian kernel density estimates of pooled wLOD scores from all individuals in the Asian Indian Gujarati (GIH) and Telugu (ITU) populations at window sizes 200 and 220 SNPs, respectively. It should be noted that similar patterns were also observed in the Asian Indian Punjabi (PJL), Sri Lankan Tamil (STU) and East Asian Dai (CDX) groups. Given they overlap significantly with the GIH and ITU density estimates they are excluded here for clarity. **B)** The proportion of windows comprising each inferred ROA present in the right-most (consanguinuity) mode is presented for 4 Asian Indian populations: Gujarati, Telugu, Punjabi and Sri Lankan Tamil. As in **A**, the East Asian Dai are excluded given that their Gaussian kernel density estimates are almost exclusively in the left- most mode (endogamy). *Figure 6 in original publication* [360].

If trimodal distributions are a reflection of the wLOD method being able to disentangle autozygosity signals arising from endogamy and consanguinity processes, we would expect inferred ROA to be delineated predominantly by windows from only one of the two autozygous modes. If on the other hand, the trimodal distribution is just an idiosyncrasy of the wLOD estimator, resultant ROA would presumably be delineated by a random mix of windows drawn from the two autozygous modes. We thus investigated how windows in the putative endogamy and consanguinity-associated modes cluster to form inferred ROA for each population exhibiting a clear trimodal wLOD score distribution. We first constructed ROA from windows with wLOD scores above the minimum between the non-autozygous and left-most autozygous modes in their wLOD score distribution [338]. Next, for each inferred ROA, we calculated the proportion of their

underlying autozygous windows that had wLOD scores within the right-most putative consanguinity-associated mode (i.e. above the minimum between the two autozygous modes).

Inferred ROA were found to frequently be delineated by windows drawn predominantly from one of the two autozygous modes (Figure 3.16B). A large well-defined peak is observed at low proportions, representing those ROA comprised of > 90% of windows drawn from the left-hand endogamy-associated mode. A more diffuse peak is observed at higher proportions, representing those ROA comprised of > 80% of windows drawn from the right-hand consanguinity-associated mode. The dispersed appearance of the peak representing putative consanguinity-associated ROA can be explained as a reflection of the fact that the two autozygous modes are not distinct. At the ends of ROA arising via consanguinity, the wLOD scores of windows will naturally decrease as they increasingly span non-autozygous regions and overall support for autozygosity declines, leading them to increasingly fall within the endogamyassociated mode. Consequently, we would expect ROA arising via consanguinity to contain a small proportion of windows in the endogamy-associated mode, with the proportion varying based upon the overall strength of the autozygous signal (i.e. ROA conferring generally higher wLOD scores will have lower proportions of windows in the endogamy-associated mode). Nevertheless, across populations, 68.9% (PJL) to 84.5% (CDX) of all ROA had >80% of their component windows drawn from a single autozygous mode.

Additional support for trimodality in the wLOD score distribution reflecting distinct autozygosity signals arising from endogamy and consanguinity processes is provided by a comparison of how the proportion of windows drawn from the consanguinity-associated mode changes with ROA length (Figure 3.17). Almost all ROA longer than 5 Mb are comprised predominantly of windows drawn from the consanguinity-associated mode (>90%), while proportions among ROA shorter than 5 Mb are much more variable. This pattern is consistent with the expectation that ROA arising via consanguinity will in general be much longer than those arising via endogamy.

Overall, the properties of ROA constructed from the trimodal wLOD score distributions present in the Asian Indian and East Asian Dai populations are compatible with the wLOD method being capable of disentangling autozygosity signals that arise from different cultural processes at sufficiently large window sizes. However, further work in well-defined populations that practise

both endogamy and consanguinity will be required to fully evaluate this apparent property of the wLOD method.

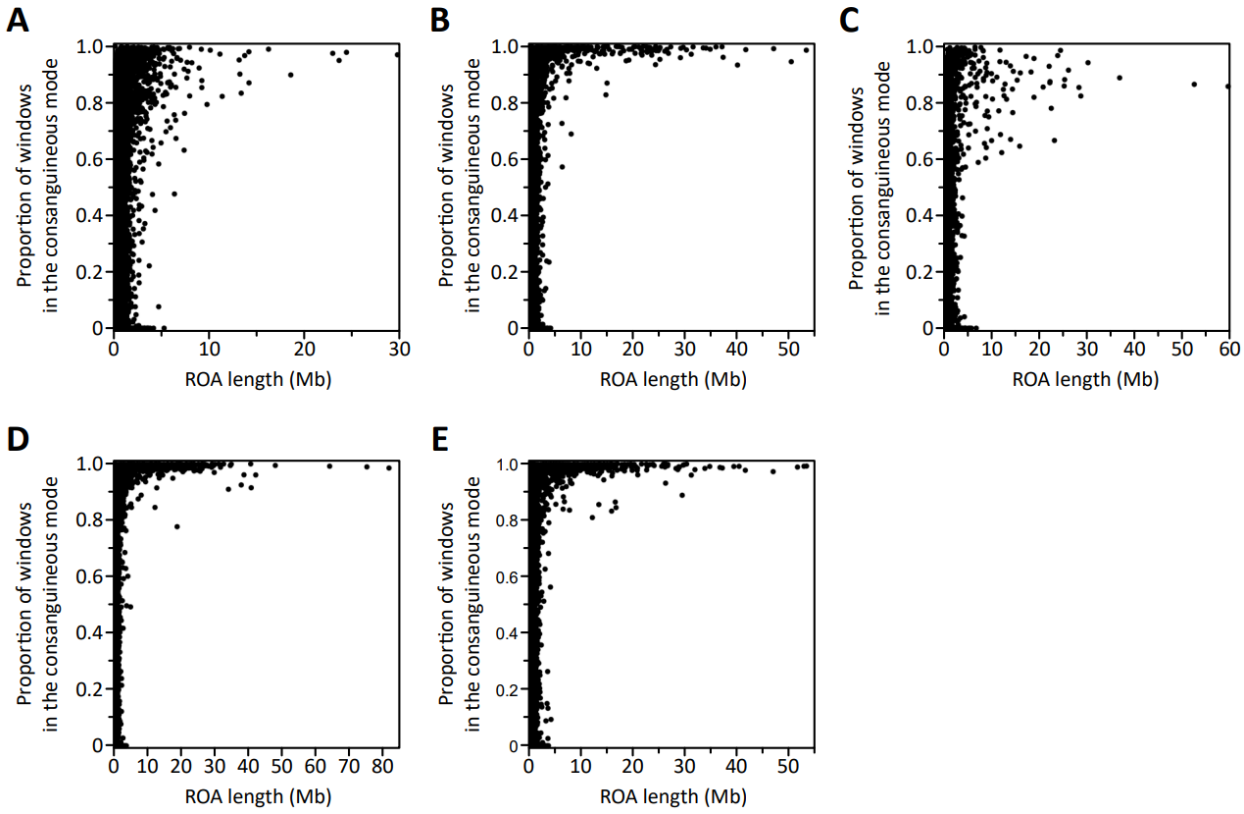


FIGURE 3.17-RELATIONSHIP BETWEEN ROA LENGTH AND PROPORTION OF WINDOWS DRAWN FROM THE CONSANGUINITY MODE

Four scatterplots illustrate the relationship between ROA length and the proportion of windows found in the putative consanguinity-associated autozygous mode for the following populations: **A)** Gujarati (Spearman's $\rho = 0.234$, $P < 10^{-16}$), **B)** Telugu ($\rho = 0.261$, $P < 10^{-16}$), **C)** East Asian Dai ($\rho = 0.185$, $P < 10^{-16}$), **D)** Punjabi ($\rho = 0.302$, $P < 10^{-16}$), and **E)** Sri Lankan Tamil ($\rho = 0.286$, $P < 10^{-16}$). *Figure S17 in original publication* [360].

3.4.9 Population patterns in ROA

To evaluate how genome-wide patterns in ROA inferred with the wLOD method and classified into 5 classes via a Gaussian mixture model applied to their genetic map lengths accord with those of earlier studies, we next performed the first high-resolution survey of ROA patterns in the TGP Phase 3 populations based upon ROA inferred in the WGS dataset as described above.

Consistent with previous studies [341][414][415], ROA of different lengths have different continental patterns among the 26 populations included in Phase 3 of the TGP, both with regards to their total lengths (Figure 3.18) in individual genomes as well as in their non-uniform distributions across the genome (Figure 3.19) that are correlated with spatially variable genomic properties such as recombination rate and signals of natural selection, reflecting the distinct forces generating ROA of different lengths. Total lengths and numbers of ROA in the shortest (G = 1–3) and to some extent intermediate (G = 4) classes increase with distance from Africa, rising in a stepwise fashion in successive continental groups (Figure 3.18), in agreement with the observed reduction in haplotype diversity with increasing distance from Africa [394][416]–[418]. Those of the longest class (G = 5) do not show a similar stepwise pattern, instead exhibiting higher and more variable values in populations where consanguinity in more frequent (Table 3.2) and inbreeding coefficient estimates are generally higher [415]. Notably, the East Asian Dai have remarkably high total lengths of short ROA (G = 1–3), potentially reflecting their small population size (~1.2 million) in Yunnan province, China [416], where the TGP samples were collected—and complex evolutionary history [421][422].

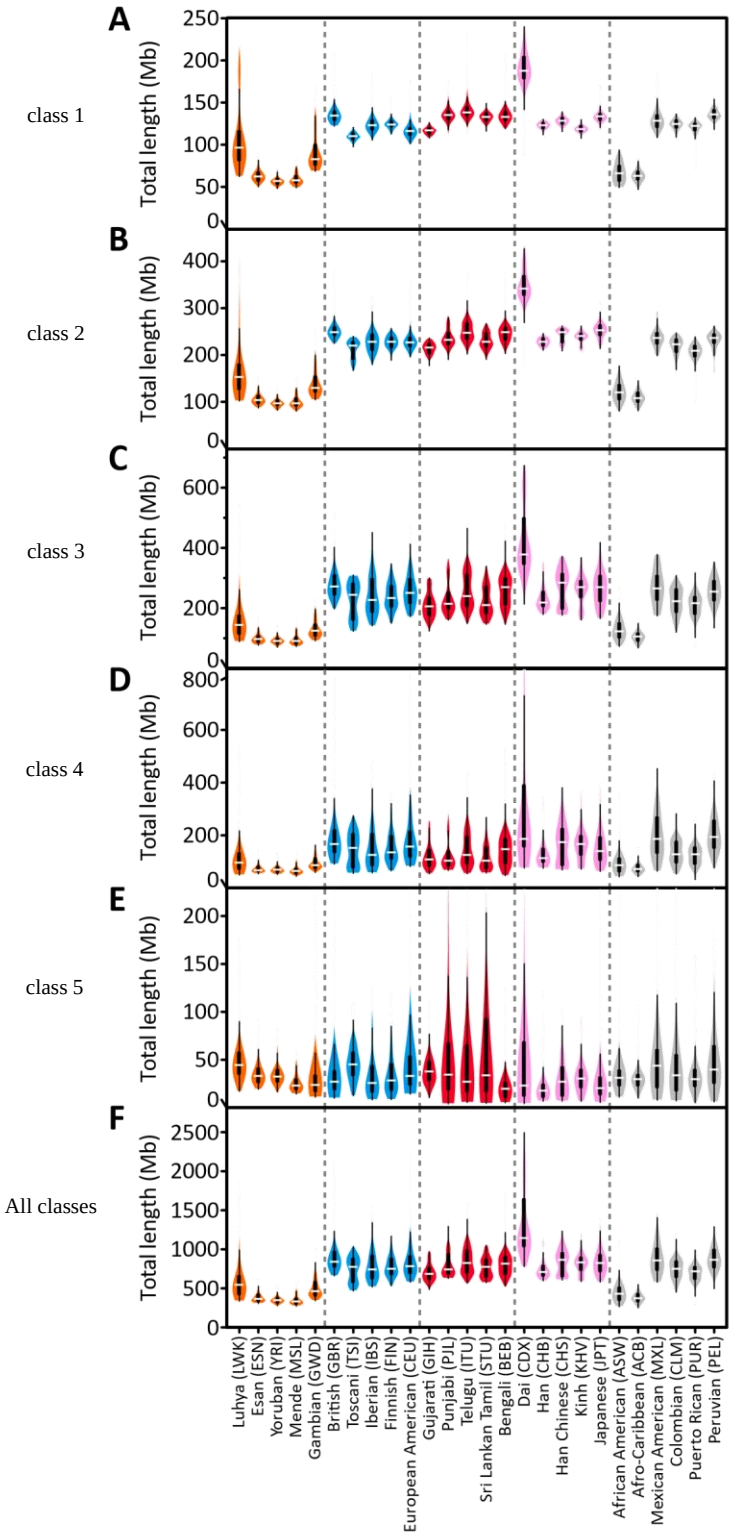


FIGURE 3.18-POPULATION SPECIFIC DISTRIBUTIONS OF TOTAL ROA LENGTH PER INDIVIDUAL

Each violin plot represents total ROA length across all individuals in each population for a given ROA class, or for all ROA classes combined. Populations are ordered from left to right by geographic region and within each region by increasing geographic distance from Addis Ababa. *Figure S18 in original publication* [360].

FIGURE 3.19-DISTRIBUTION OF WORLDWIDE ROA FREQUENCIES ACROSS CHROMOSOMES 1-22
For each autosomal chromosome, the average proportion of individuals in the WGS dataset who have an ROA overlapping SNVs in a non-overlapping 50 kb windows is shown for each of the 5 ROA length classes. Each row represents an ROA class, and each column represents a window. The intensity of a point increases with increasing average ROA frequency, as indicated by the color scale. Under the proportion bars for classes 1-5 two additional metrics are shown: SNV density of each window and an ideogram of chromosome banding. Below the ideogram, vertical black lines represent the average recombination rate in each window, with line heights proportional to average recombination rate.
Figure S19 in original publication [360].

3.4.10 Recombination and natural selection

The correlation between the genomic distribution of ROA and recombination rate decreases with increasing ROA class (not shown), consistent with the expectation that the patterns of genetically shorter ROA will be determined by recombination to a greater extent than longer ROA, which due to their more recent origins have had fewer opportunities for recombination events to systematically influence their patterns. Conversely, correlation between ROA and signatures of natural selection is strongest for class 2–3 ROA, to some extent intermediate class 4 ROA, and very weak for the shortest (G = 1) and longest (G = 5) ROA classes (not shown). These patterns are compatible with natural selection having primarily influenced genomic diversity patterns in the distant past, with autozygosity for the relics of haplotypes that arose during those events manifesting as class 2–4 ROA, dependent upon how long ago the event occurred.

The long-term effects of natural selection on patterns of ROA might be expected to be most evident in genomic regions encompassing genes implicated in one or more Mendelian diseases, where purifying selection acting on strongly deleterious alleles, which may occur more frequently in such genes due to their apparent importance for human health, would be expected to increase levels of autozygosity relative to genes much less frequently subjected to purifying selection. Using the union of two previously reported lists of genes associated with autosomal dominant (669) and recessive (1,130) diseases in the Online Mendelian Inheritance of Man (OMIM) database [419]–[421], we created a list containing genes not associated with autosomal dominant or recessive diseases (24,260; "non-OMIM" henceforth); genes associated with both autosomal dominant and recessive diseases were ignored. For each individual, we then calculated the fraction of the total lengths of all autosomal dominant, autosomal recessive, or non-OMIM transcribed regions that are overlapped by ROA based on their genomic positions in build hg19 of the UCSC reference genome assembly. Strikingly, regardless of the ROA length class considered, the fraction for OMIM dominant genes was almost always higher than that of non-OMIM genes ($p < 10^{-16}$ in all comparisons; Wilcoxon signed rank test), while the opposite was true for OMIM recessive genes ($p < 10^{-16}$ in all comparisons; Figure 3.20). Nevertheless, the pattern is strongest for intermediate length ROA classes (G = 2–4) and weakest for the shortest (G = 1) and longest (G = 5) classes. Together, these results suggest that deleterious alleles may occur less frequently in non-OMIM genes than in OMIM dominant genes. Deleterious alleles in OMIM dominant genes could be

removed from the population via purifying selection, and thus creates increased autozygosity at lengths consistent with population-level processes rather than inbreeding. One possible explanation for the decreased autozygosity around OMIM recessive genes compared with non-OMIM genes would be increased embryonic lethality and/or childhood mortality with individuals homozygous for deleterious recessive mutations in OMIM recessive genes, leading to reduced autozygosity in genomic regions encompassing them in the extant population.

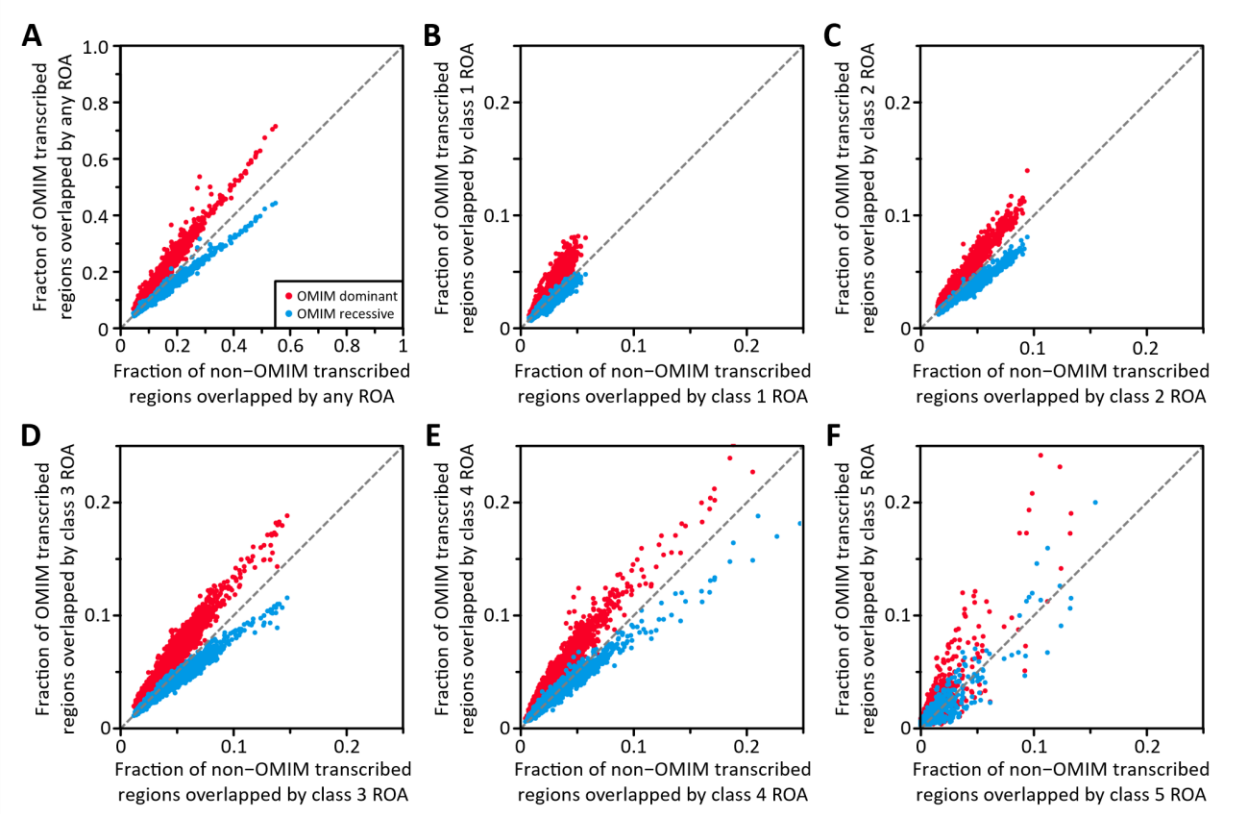


FIGURE 3.20-FRACTION OF OMIM VS NON-OMIM GENES WITHIN ROA
Six scatterplots compare the fraction of non-OMIM transcribed genes to the fraction of OMIM transcribed genes in each individual's genome. OMIM dominant genes are shown in red, and OMIM recessive genes are shown in blue. The identity line is drawn in grey. Fraction of OMIM vs non-OMIN genes is shown separately for **A)** all ROA, **B)** class 1 ROA, **C)** class 2 ROA, **D)** class 3 ROA, **E)** class 4 ROA, and **F)** class 5 ROA.
Figure S22 in original publication [360].

Genes that have been the target of positive selection might be expected to reside within genomic regions that are more frequently autozygous in the general population than those harboring genes that have not. Considering the fraction of each gene's transcribed region that is in a ROA in each individual's genome, we compared their median fraction across individuals in each population (Figure 3.21). While most genes have a median fraction of about zero, a number of genes that lie within genomic regions spanned by ROA in more than 90% of individuals in a population. Across populations, we observe 54 such instances with long class 5 ROA that represent seven distinct genomic regions (Appendix Table 6) and 159 with intermediate length class 4 ROA that represent 22 distinct regions (Appendix Table 7). The short ROA classes had 31 (in 9 distinct regions), 7 (in 5 distinct regions) and 480 (in 46 distinct regions) instances for ROA classes 1-3, respectively (Appendix Tables Appendix Table 8-10). While most genes in these regions fall within the non-OMIM group, two of the genes enriched for class 4 ROA (CFC1 and SMN1) and nine of the genes enriched for class 1 ROA (SLC25A20, NDUFAF3, LAMB2, GPX1, NPRL2, ACY1, MRPS16, LCAT, and COX4I2) are from the OMIM recessive group, while one gene enriched for class 1 ROA is from the OMIM dominant group (THAP1). Future investigation of genes that are unusually frequently overlapped by ROA in the general population may provide new insights into the role of recessive variation in human phenotypic diversity and common disease risk as well as the genes within which such variation acts.

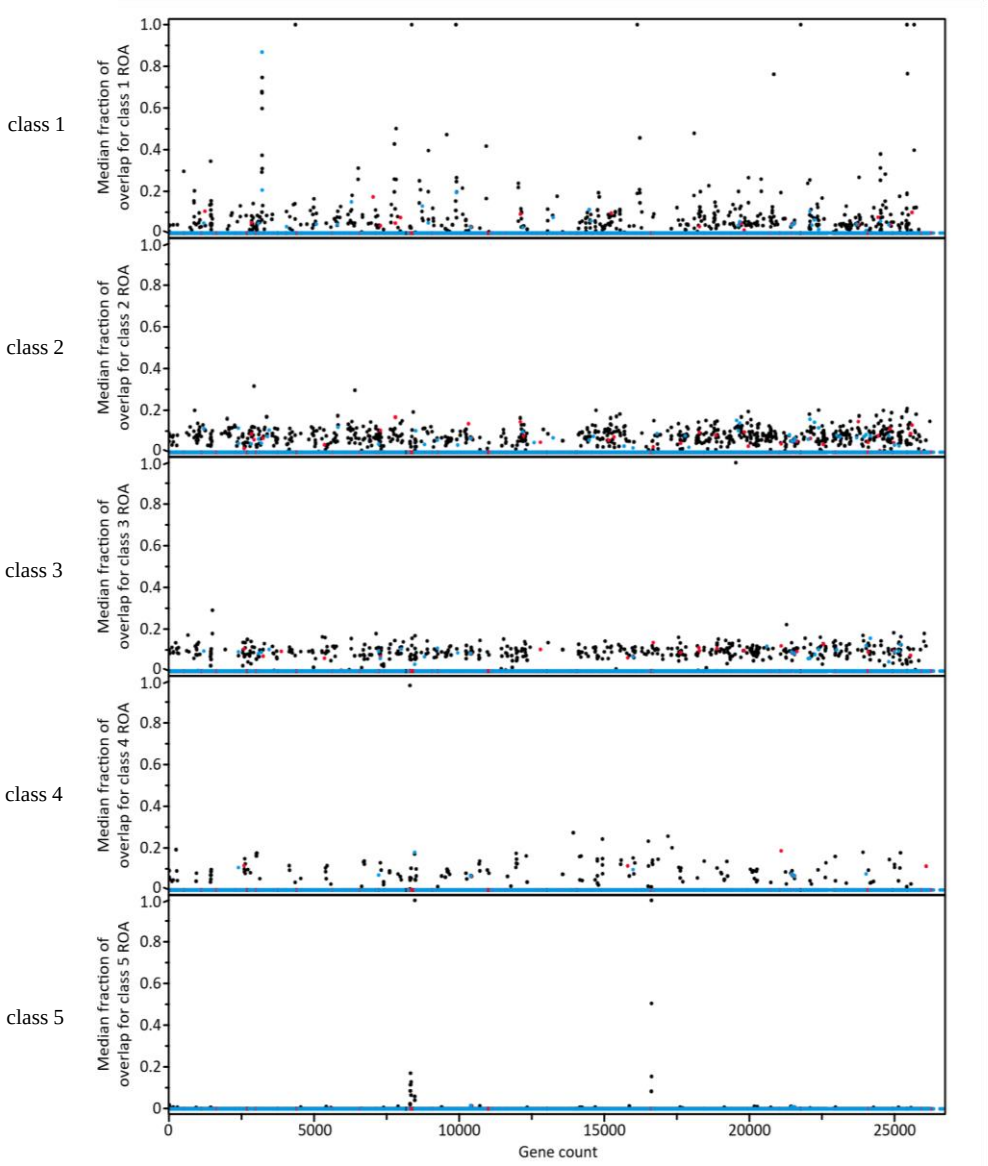


FIGURE 3.21-MEDIAN FRACTION OF EACH AUTOSOMAL GENE IN ROA IN EUROPEAN AMERICANS

Using the CEU population, and for each ROA class, Manhattan plots depict the median fraction of transcribed genes that overlap ROA. Genes are ordered from left to right by increasing position in the genome. OMIM recessive genes are shown in red, OMIM dominant genes are in blue, and non-OMIM genes are in black.

Figure S23 in original publication [360].

3.4.11 Genomic distribution

Genomic distributions of shorter ROA (*G* = 1–4) are similar among populations from the same geographic region and closely mirror the patterns of pairwise F_{ST} among populations, while those of the longest ROA class (G = 5) vary more widely among populations (data not shown). Overall, these patterns suggest that shorter ROA (*G* = 1–4) -for which neighboring populations have similar patterns, reflect autozygosity that arises through population processes on different evolutionary timescales, while longer ROA (*G* = 5), for which neighboring populations do not necessarily have similar patterns, reflect autozygosity that instead arises through more recent cultural processes such as inbreeding [338].

3.4.12 Autozygosity hotspots

The non-uniform genomic distribution of the different ROA classes and their variability among populations creates autozygosity hotspots that are in some instances shared among subsets of the populations. For example, there is a hotspot for class 4 ROA on the q-arm of chromosome 2 that is common to three of the five European populations and encompasses the human lactase gene (LCT; Figure 3.22) that was not detected in our original study of the HGDP and HapMap populations (which included 10 European groups) [338]. In this genomic region, we observe high frequencies of intermediate length class 4 ROA in the Northern European FIN and GBR populations as well as the European American (CEU) group, but not in the Southern European TSI and IBS populations or any other population in the dataset. The presence and absence of this hotspot broadly reflects worldwide patterns in lactase persistence frequency [426][427]. Lactase persistence is most frequent in Northwestern Europe [428][429] where it is caused primarily by a single mutation in LCT that rose to high frequency as a consequence of natural selection in response to the rise of milk consumption and pastoralism [427][430][431]. It decreases in frequency through Eastern and Southern Europe and Central/South Asia reaching near-zero frequencies in East Asia and the Americas [426][429][432]–[434], while it is present to varying degrees in admixed Mestizo [431]–[433] and African American [428][431] populations as a consequence of their recent European ancestry. Thus, we observe high levels of autozygosity around LCT in the GBR, FIN, and CEU populations and markedly lower but noticeable levels in the IBS, but no observable signal in the TSI or any of the Asian or admixed populations. While lactase persistence is present at moderately high frequency in sub-Saharan Africa it is caused by

several different mutations [423][434], and the African populations included in The TGP are located predominantly in historically non-milking areas of the continent [426]. Consequently, we do not observe a similar autozygosity signal in the African populations as we do in the Northern European populations.

Interestingly, we also observe a hotspot for the longest ROA class (G = 5) at the same location in the Northern European CEU and GBR populations ~770 kb downstream of the LCT gene (Figure 3.22) while a weaker spike in class 5 ROA frequency is seen in the FIN population. This hotspot encompasses four genes within its core region (chr2:135,375,000–135,775,000) that encode a transmembrane protein (TMEM163), an aminocarboxymuconate semialdehyde decarboxylase (ACMSD), cyclin T2 (CCNT2), and a mitogen-activated protein kinase kinase kinase (MAP3K19). The maximum normalized haplotype-based selection statistic nS_L score [387] observed in the CEU, GBR, and FIN populations within the core region is 4.980, 4.818, and 4.962, respectively. The latter represent a high nS_L statistic, which suggests that the observed ROA hotspot could reflect the outcome of recent positive selection. However, none of the genes within this hotspot are known to have functional consequences when mutated, leaving the cause of this ROA hotspot and its putative signals of positive selection enigmatic.

Overall, frequency patterns in this genomic region of the different ROA classes in the Northern European CEU, GBR, and FIN populations are consistent with positive selection having occurred at two different timepoints. The extended haplotypes created by historical positive selection acting on the single LCT mutation that arose in ancestral Northern Europeans have, over subsequent generations, decreased appreciably in length, but due to the marked reduction in haplotype diversity in the surrounding region commonly create intermediate length class 4 ROA through background population processes. Conversely, the presence of extended IBD haplotypes creating longer class 5 ROA in a genomic region ~770 kb away from LCT would be compatible with positive selection acting much more recently, in agreement with the atypically high nS_L scores observed within this region in these populations.

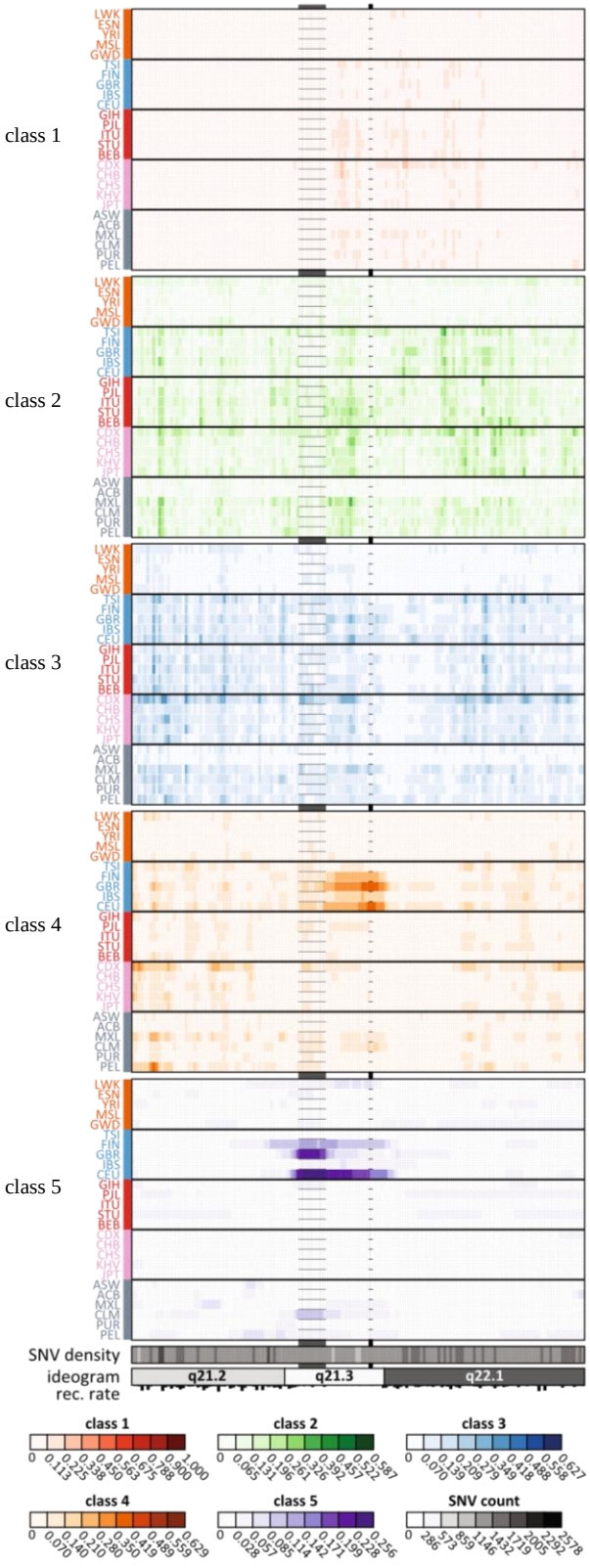


FIGURE 3.22-PER POPULATION ROA FREQUENCIES WITHIN AN ROA HOTSPOT ON CHROMOSOME 2

ROAs of various sizes overlapping SNVs on the q-arm of chromosome 2 (from 132,500,000 to 140,200,000 bp) are shown. Each row represents a population, ordered from top to bottom by geographic affiliation, and by increasing geographic distance from Addis Ababa.

Each column represents a non-overlapping 50 kb wide window. For each population, at a given window and ROA class, a point represents the population average of ROA frequency at that location, Increased color intensity represents increased average ROA frequency, as indicated by the color scale. The SNV density of each window and an ideogram of chromosome banding are shown at the bottom, with average recombination rate in each window represented by a vertical black line below the ideogram, where line heights proportional to average recombination rate. **1** and **2** mark the location of the class 5 ROA hotspot in the CEU/GBR and the LCT gene, respectively.

Figure 7 in original publication [360].

3.4.13 Statistical inference of enrichment of autozygosity signals between groups

A unique feature of the wLOD ROA detection approach is the availability of log-likelihoods of autozygosity for each window in each individual examined. It is therefore possible to directly compare the strength of autozygosity signals between two or more groups of individuals to identify those windows that have significantly greater evidence for shared autozygosity signals in one group compared with the others [358]. In one possibility, such an approach could be used to identify genomic regions that have stronger signals of autozygosity in affected versus unaffected individuals and thus may harbor disease associated mutations. Similarly, genomic regions with significantly stronger signals of autozygosity in one subset of a population compared to another other may reflect founder effects if there is limited gene flow between them or the presence of adaptive alleles in one subset but not the other that have risen to high frequency.

We demonstrate the principle of this approach using three of the five Central/South Asian groups included in Phase 3 of the TGP who represent subpopulations within the larger Indian population: BEB, GIH, ITU, PJL, and STU. Genetic diversity patterns in these five groups support the presence of two genetically distinguishable clusters within the GIH, ITU, and PJL (Figure 3.23). When instead compared pairwise, the larger of the two ITU clusters lies intermediate between the smaller ITU cluster and the larger of the two GIH, PJL, or STU clusters, while the largest of the PJL clusters overlaps significantly with the smaller GIH cluster (not shown). The GIH individuals were sampled in Houston, TX, while the BEB, ITU, PJL, and STU individuals were all sampled in the UK. Given the intermediate locations of the larger ITU and PJL clusters in the pairwise comparisons, they may potentially reflect admixed individuals within these sample sets. However, both clusters are tightly bunched arguing against this possibility given the normal dispersion of admixed individuals in such analyses owing to their continuum of admixture levels [435][436]. In another possibility, these distinct clusters might represent the unintentional sampling of distinct endogamic communities whose restrictive marital practices under the long-established Indian caste system has made them distinguishable genetically [437].

When searching for differential autozygosity signals across groups, an optimal window size can be defined as one in which the wLOD score distribution best discriminates between autozygous and non-autozygous windows. In this study, this is defined as the window size that maximizes the distance between the autozygous and non-autozygous modes—measured as the distance between

the modal score in each mode (Figure 3.3B and Figure 3.4). Using the WGS dataset and optimal window sizes of 450, 580, and 610 SNVs for the GIH, PJL, and ITU, respectively, we compared the wLOD scores of individuals present in each of their two clusters (Figure 3.23) and evaluated the significance of their observed differences with the permutation-based approach described in *Wang et al* [358]. One modification of the latter approach was the replacement of the two sample t-test with the Wilcoxon rank-sum test which is less sensitive to the presence of outliers but has similar power to detect a location shift [438]. Briefly, separately for each group, we first create a distribution of test statistics under the null hypothesis of no difference in wLOD scores between clusters using 1000 permutations of cluster labels, recording for each permutation the maximum observed test statistic across all windows genome-wide. Next, separately for each window, a genome-wide adjusted p-value for the significance of the observed differences in wLOD scores between clusters is then calculated as the proportion of the maximum genome-wide test statistics observed in the 1000 permutations that exceeded the test statistic obtained with the true labels for that window. Finally, for each cluster, genomic regions enriched for autozygosity signals in that cluster compared with the other were defined by joining together overlapping windows with a permutation p-value (P_{perm}) ≤ 0.05.

Intriguingly, while we would not a priori expect to observe significant differences in the strength of autozygosity signals between the two apparent clusters within the GIH, ITU, and PJL sample sets, we did identify one genomic region significantly enriched for autozygosity signals in cluster A compared with cluster B in both the ITU and PJL (Figure 3.24 and Table 3.3). In contrast, no regions were identified in the GIH (Figure 3.25). The genomic region in the ITU lies within the transcription elongation regulator 1 like (TCERG1L) gene, associated with regulation of plasma levels of the adipokine adiponectin [439], a modulator of glucose regulation and fatty acid oxidation implicated in obesity, diabetes, coronary artery disease and Crohn's disease risk [440]–[443]. The genomic region in the PJL encompasses the transmembrane phosphoinositide 3phosphatase and tensin homolog 2 (TPTE2) gene, a paralog of the phosphatase and tensin homolog (PTEN) tumor suppressor [444] implicated in hepatic carcinogenesis [445] and that has been found to harbor SNPs with significant allele frequency differences between males and females in European and African populations [446]. In the absence of more detailed information on these individuals, the underlying basis for these differential autozygosity signals remains

unknown. However, their identification highlights the potential of our approach to identify genomic regions with differential autozygosity signals between groups potentially reflective of variants that have experienced differential selection histories or that influence differences in their predisposition to disease. Moreover, these findings highlight the need for further investigations among well-defined endogamic groups from India to facilitate our understanding of the genomic consequences of the long-established caste system.

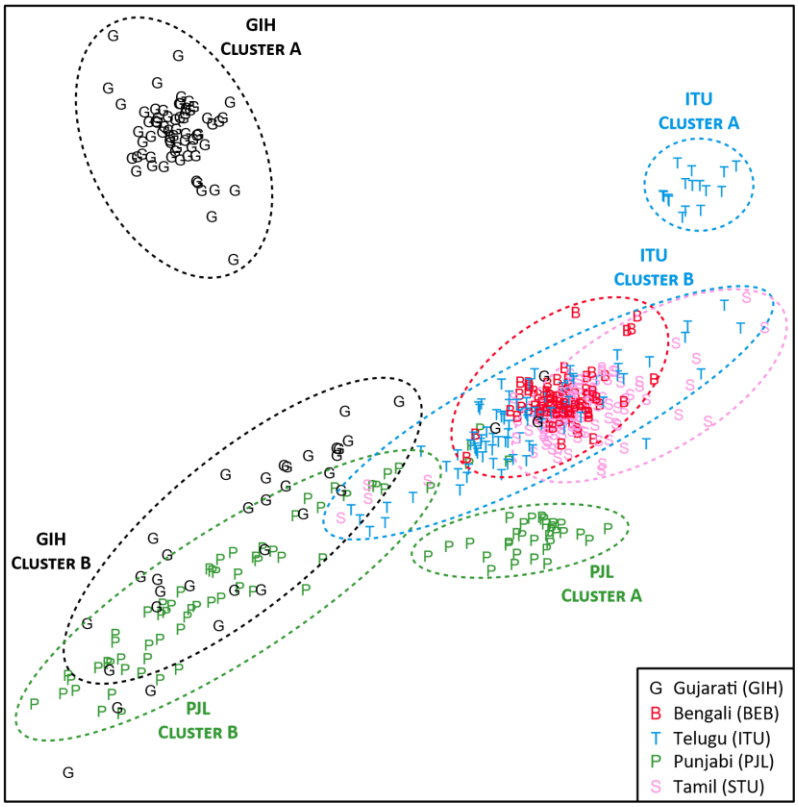


FIGURE 3.23-EVIDENCE OF GENETIC STRUCTURE IN 3 ASIAN-INDIAN GROUPS
A multidimensional scaling (MDS) representation of pairwise allele sharing dissimilarities (ASD) among individuals in the five Asian Indian groups. GIH, ITU, and PJL groups each resolved into two sub-clusters, highlighted by the dotted circles. The pairwise ASD matrix was constructed using asd (https://github.com/szpiech/asd) considering only the Asian Indian individuals and MDS applied to this matrix using cmdscale in R.
Figure S25 in original publication [360].

Population			Genomic region				Number of windows	MinimumP_{perm}	RefSeq genes
ID	Name	Cluster	Chr	Begin (bp)	End (bp)	Length (bp)			
ITU	Telugu	A	10	132,953,074	133,048,305	95,232	189	0.013	*TCERG1L*
PJL	Punjabi	A	13	20,001,572	20,181,691	180,120	28	0.044	*TPTE2*

TABLE 3.3-GENOMIC REGIONS ENRICHED FOR AUTOZYGOSITY SIGNALS IN THE ITU AND PJL SUBGROUPS

The wLOD algorithm uncovered two genomic regions enriched for autozygosity signals: in the Telugu (ITU) on chromosome 10 within the TCERG1L gene (transcription elongation regulator 1 like) and in the Punjabi (PJL) on chromosome 13 encompassing the TPTE2 gene (phosphoinositide 3- phosphatase and tensin homolog 2). Cluster A refers to the MDS plot in Figure 3.23 and P_{perm} to a permutation p-value previously described [359]. Genes were identified using NCBI Reference Sequence (RefSeq).

Table 3 in original publication [360].

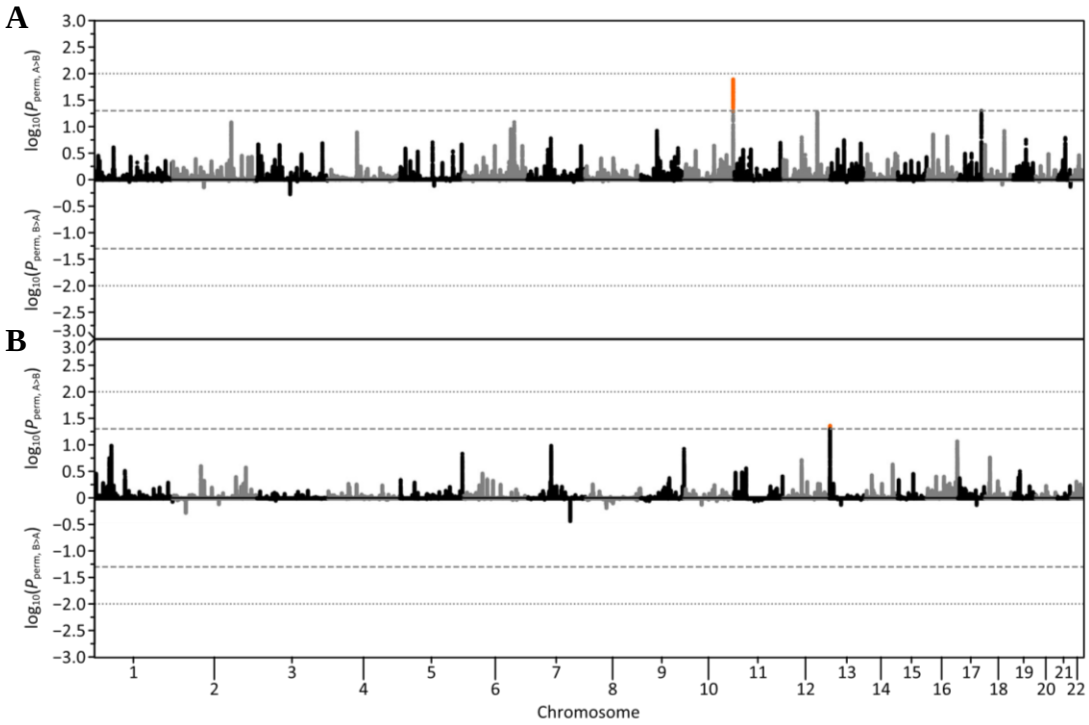


FIGURE 3.24-DISTRIBUTION OF DIFFERENTIAL ROA SIGNALS BETWEEN SUBGROUPS IN THE TELUGU AND PUNJABI

Manhattan plot showing for each window the $\log_{10}$(P-value) of pairwise comparisons of per-individual wLOD scores in the two clusters (Clusters A and B -Figure 3.23) present in the: **A)** ITU (580 SNV window) and **B)** PJL (610 SNV window). For each population, tests of comparison via genome-wide adjusted P-values show which of the two clusters is enriched for autozygosity signals (ie: higher wLOD scores). wLOD scores higher in Cluster A vs in B are shown above y = 0 (the top panel) and those that are higher in Cluster B vs in A are shown below y = 0 (bottom panel). P-values represent the proportion of permuted test-statistics that exceed the test statistics obtained with true labels [358]. Windows with $P > 0.05$ are shown in black and those with $P \leq 0.05$ are shown in orange. The horizontal grey dashed line denotes P = 0.05 and while the grey dotted line denotes P = 0.01.

Figure 8 in original publication [360].

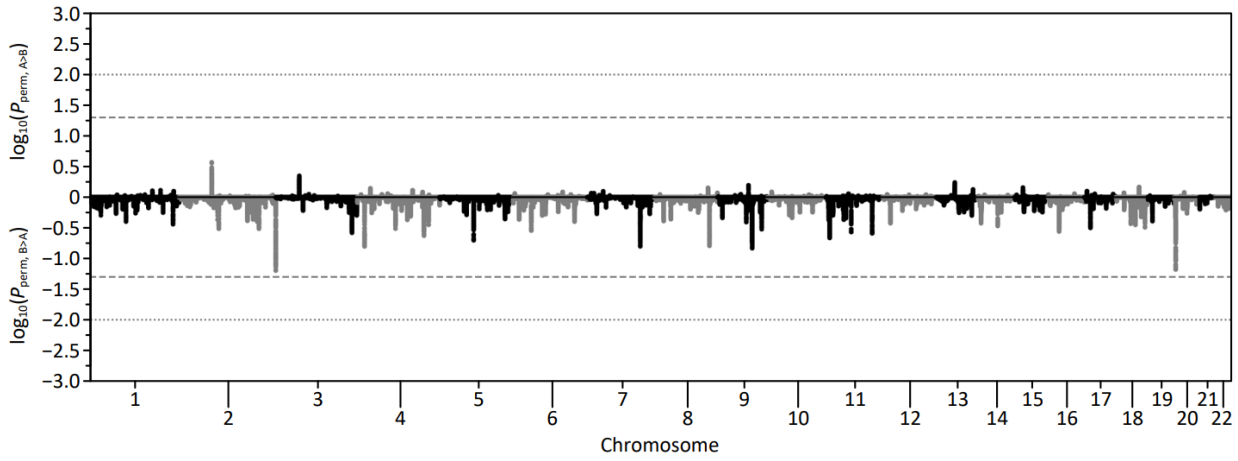


FIGURE 3.25-DISTRIBUTION OF DIFFERENTIAL ROA SIGNALS BETWEEN THE SUBGROUPS IN THE GUJARATI

Manhattan plot showing for each window the $\log_{10}(P)$ of pairwise comparisons of per-individual wLOD scores in the two Clusters present in the GIH (Figure 3.23). Format of this figure is identical to Figure 3.24.

Figure S26 in original publication [360].

3.5 Discussion

We have reported an improved likelihood-based estimator for the detection of ROA in genome-wide SNV genotsype data derived from either microarray platforms or WGS that accounts for autocorrelation among genotyped positions and variability in the confidence of individual genotype calls as well as the probabilities of unobserved mutation and recombination events. Fully accounting for LD among SNVs in a given window is important, because in genomic regions of high LD many pairs of individuals will share common haplotypes that are IBS but not IBD. Thus, including such spurious windows would add noise when looking for ROA for the purpose of autozygosity mapping. The incorporation of LD in our model reduces false-positive ROA detection, affording us the ability to identify smaller ROA segments with greater fidelity. An alternative approach to accounting for LD is to prune the dataset prior to its analysis. However, since MAF strongly influences pairwise LD estimates [447] leading to a floor effect beyond short distances when SNP pairs have large MAF differences, such an approach commonly requires those SNV with MAF less than 5% to be removed [376], which would significantly reduce the power of the wLOD method to detect ROA by removing those low-frequency and rare variants whose homozygosity is most indicative of autozygosity under its likelihood model (Figure 3.2A). Further,

such pruning cannot completely remove LD from the dataset being analyzed, with a pairwise r^2 threshold of 0.5 typically applied [376]. The incorporation of LD into the model therefore better controls for the autocorrelation of autozygosity signals among nearby SNV than is attainable with LD pruning, thereby improving the specificity of the ROA it detects particularly in regions of moderate to high LD.

Similarly, accounting for the probabilities of unobserved recombination and mutation events in the genomic interval spanned by the window becomes increasingly important as a function of inter-marker distance, particularly in situations where these probabilities become non-trivial such as in lower-density microarray-derived genotype datasets. By modeling these probabilities based on an assumed minimum number of generations since the MRCA of the apparent autozygous haplotypes, which we have set here based on the reported effective sizes of the populations included in the TGP [366][368][389], we minimize the number of false positive ROA that can be erroneously inferred when recombination and mutations events onto very similar haplotype backgrounds give the appearance of autozygosity when paired with a non-recombined haplotype. An alternative approach would be to set an arbitrary maximum inter-marker distance allowed when calling ROA; dividing into two inferred ROAs that spans an inter-marker interval greater than that maximum. However, this has the potential to erroneously break-up long ROA, potentially impacting downstream analyses that use ROA length as filtering criteria. By incorporating mutation and recombination weightings into the wLOD model we therefore take a more informed and less-biased approach to this issue, thereby improving the inference of longer ROA particularly in datasets containing sparser sets of SNVs.

We have shown the wLOD ROA inference method to be well-powered to infer ROA in genetic datasets consistent with those generated by WGS and microarray-based genotyping. We recommend using this method together with a model-based ROA classification approach [338] based on genetic map lengths to distinguish ROA arising from population-level LD patterns on different evolutionary timescales (classes G = 1–4) from those arising from more recent cultural processes such as inbreeding (class G = 5). Our findings suggest that our inference approach is robust for analyses of as few as 10 individuals. However, model-based classification requires at least 25 individuals to provide a robust classification solution. Moreover, to ensure allele frequency

and LD estimates used with the wLOD estimator are close to their true value in the population, at least 30 unrelated individuals should ideally be used in their estimation [379][448]. Intriguingly, our observation of trimodal wLOD score distributions for a subset of the 26 populations analyzed here, all known to practise both endogamy and consanguinity to varying degrees, suggests that this method may be able to distinguish autozygosity arising from different cultural processes that act on different time scales. Future work within well-defined endogamic and non-endogamic groups that practice consanguinity, as well as within simulated datasets exploring the breadth of possible isolation and inbreeding parameters observed in human populations, will be required to clarify this apparent property of the wLOD method and evaluate its potential human genetics applications.

Comparisons of the ROA inferred using the wLOD method on different microarray-derived and NGS datasets created from the TGP WGS data suggest that long and to some extent intermediate length ROA are captured consistently by WGS and microarray-derived datasets. However, inference of shorter ROA does vary substantially among the different datasets as a consequence of the decreasing resolution and sensitivity attainable as the genome-wide density of genotyped positions decreases. The latter observation is particularly well reflected by the notable lack of consistency between ROA inferred in the WES dataset as compared to those identified in the WGS dataset. Nevertheless, population-genetic analyses of genomic ROA patterns among the 26 populations included in the TGP on the basis of WGS data are consistent with our previous findings in the 64 worldwide populations included in the HGDP [449][450] and International HapMap Project [451] on the basis of ~600,000 microarray-derived SNP genotypes [338]. These observations would therefore suggest that ROA studies using microarray-derived genotype data have similar power to detect genomic ROA patterns, and in particular those of longer ROA that are of interest to the disease genetic community due to their enrichment of deleterious variation carried in homozygous form [452] as those using WGS data.

We have compared the wLOD method against a commonly used genotype counting method implemented in the software PLINK, as well as the recently reported HMM method of the BCFtools software package, under two demographic scenarios in which ROA of interest in population- and disease-genetic studies. In our genetic simulations, the PLINK approach performed surprisingly well, potentially reflecting their relatively short duration that limited the

opportunities for new mutations to arise on the IBD haplotypes that ultimately underlied ROA in the final generation. Indeed, only ~ 4.01% and ~ 14.36% of SNVs in our simulated datasets were de novo mutations not present in the founder individuals under scenarios 1 and 2, respectively, while just ~ 2.14% and ~ 2.91% of SNVs had MAF < 5%. Conversely, across the 26 populations in the TGP WGS data on average 56% of SNVs had MAF < 5%. Nevertheless, the wLOD method had greater power to detect ROA versus PLINK across all SNV densities considered here. This difference reflects the limited ability of the PLINK approach, which allows for only occasional missing or heterozygous genotypes when determining the status of a window to account for possible genotyping errors and mutations, to distinguish genomic regions that are homozygous-by-chance from those that are autozygous. In contrast, the wLOD method incorporates population allele frequency and LD estimates and an assumed genotyping error rate as well as accounts for the probabilities of unobserved mutations and recombination events when inferring the autozygosity status of a window, enabling more rigorous assessments of the possibility of genotyping errors and the loss of information caused by missing data. In addition, it provides a more precise measure of the probability that a given window is truly autozygous rather than simply homozygous by chance. Thus, the greater power of the wLOD method compared with PLINK reflects the greater number of false negative ROA expected under the naïve autozygosity model implemented in PLINK.

Comparisons of the wLOD method with the recently reported RoH function of BCFtools have shown it to have improved power to detect ROA, and smaller ROA in particular, across all SNV densities considered here, which are representative of WGS and microarray-based genotyping platforms. However, false discovery rates of the wLOD method are slightly higher than those of BCFtools, wholly reflecting a more permissive placement of ROA boundaries marginally outside of their true locations as a consequence of the sliding window approach employed. While the underlying likelihood models of the wLOD and BCFtools approaches are similar, there are two aspects of the wLOD method that explain its higher power. First, by summing over all SNVs within a given window, the wLOD method is better able to detect the autozygosity signals of ROA comprised of older (shorter) haplotypes whose constituent SNVs individually provide only weak to modest autozygosity support than the pointwise HMM employed by BCFtools. Second, the wLOD method adjusts each SNV’s log-likelihood by the probabilities that no unobserved

recombination and mutation events have occurred in the interval between it and the preceding SNV in the last *M* generations (Equation 2), where *M* is set based on the expected time since the most recent common ancestor in an individual's maternal and paternal lineages given the effective size of the population. BCFtools does not account for unobserved mutations in its inference model, and only allows for up to a single recombination event to have occurred within a given interval [377]. Thus, for longer ROA and those comprised of older haplotypes inherited IBD from an ancient ancestor, we would a priori expect BCFtools to have greater difficulty in making inferences as it will underestimate the number of recombination events that may have occurred as these haplotypes segregate in the general population. This may potentially underlie the noticeably erratic patterns observed with its power to detect ROA greater than 1.5 Mb in the higher SNV density simulated datasets (Figure 3.8).

Finally, the wLOD method distinguishes itself from BCFtools and PLINK through its ability to directly detect genomic regions enriched for autozygosity signals in one population or group compared with one or more others without requiring the inference of ROA first. We have applied this approach within the Gujarati (GIH), Punjabi (PJL), and Telugu (ITU) Asian Indian groups, comparing wLOD scores in two distinct clusters of individuals identified via multidimensional scaling of allele sharing dissimilarities (Figure 3.23). We identified two genomic regions enriched for autozygosity signals in one of the two clusters, one in the ITU and another in the PJL, that contain genes implicated in the regulation of metabolism and liver cancer risk, respectively (Table 3.3). If we instead set a more permissive threshold of $P_{perm} \leq 0.1$ when defining enriched regions, we identify an additional seven genomic regions marginally enriched for autozygosity in one cluster compared with the other (Table 3.4). One of the seven regions was identified on chromosome 2 in ITU cluster A and contains two genes: G6PC2, a pancreatic glucose-6-phosphatase implicated in the modulation of fasting plasma glucose levels [453] that is a major target of cell-mediated autoimmunity in diabetes [454], and the ATP-binding cassette transporter gene ABCB11, mutations in which cause autosomal recessive progressive familial intrahepatic cholestasis [455]. In addition, a region on chromosome 17 also identified in ITU cluster A contains seven genes that include USH1G, mutations in which cause autosomal recessive deafness in humans [456]. Finally, a region on chromosome 16 identified in PJL cluster A contains four genes including the mechanically-activated ion channel gene PIEZO1, mutations in which cause

autosomal recessive generalized lymphatic dysplasia [457] as well as autosomal dominant hemolytic anemia [458].

The presence of genes that cause autosomal recessive diseases in three of the seven marginally significant regions—a highly unlikely observation ($p < 0.008$ across 1000 random draws of genomic regions of equivalent size) —suggests the intriguing possibility that, if these clusters do indeed represent distinct endogamic communities, they may be the hallmark of cultural and natural selection processes related to the differential presence of deleterious genetic variants in these genes.

Population			Genomic region				Number of windows	Minimum P_{perm}	RefSeq Genes & miRNA
ID	Name	Group	Chr	Begin (bp)	End (bp)	Length (bp)			
GIH	Gujarati	2	2	242,977,775	243,178,150	200,376	42	0.064	*LOC728323*
GIH	Gujarati	2	20	2,799,801	2,893,954	94,154	32	0.067	*PCED1A, VPS16, PTPRA*
ITU	Telugu	1	2	169,745,523	169,865,984	120,462	39	0.084	*G6PC2, ABCB11*
ITU	Telugu	1	6	137,323,081	137,441,791	118,711	3	0.083	*IL20RA*
ITU	Telugu	1	12	100,036,343	100,203,990	167,648	43	0.055	*ANKS1B, FAM71C*
ITU	Telugu	1	17	72,781,311	72,933,576	152,266	32	0.051	*TMEM104, GRIN2C, FDXR, FADS6, USH1G, OTOP2, OTOP3*
PJL	Punjabi	1	16	88,737,170	88,812,876	75,707	12	0.087	*SNAI3, RNF166, CTU2, MIR4722, PIEZO1*

TABLE 3.4-GENOMIC REGIONS ENRICHED FOR AUTOZYGOSITY SIGNALS IN THE GIH, ITU AND PJL SUBGROUPS

The format of this table is identical to Table 3.3. Seven additional genomic regions are found in either Cluster 1 or 2 using a more permissive threshold of $P_{perm} \leq 0.1$

Table S8 in original publication [360].

3.6 Conclusions and Future work

This chapter has detailed the properties, validation and performance of wLOD in both simulated and real-world genetic data of varying SNV density. Overall, wLOD is well powered to detect ROA of different ancestral depths in both WGS and microarray datasets (however is not recommended for use in exome datasets) and has either comparable or improved performance over other publicly available tools. In addition, wLOD is uniquely able to facilitate direct comparisons of autozygosity signals. Future studies involving two or more distinct populations could thus gain insight into demographic histories or phenotypic distinctions across populations by calculating proportions of shared autozygosity signals within each group.

This study uses wLOD to explore the genomic content and distribution of autozygosity segments. Of the transcribed genomic areas, ROA segments were found more likely to overlap OMIM genes than non-OMIM genes. Most of the 26 populations contained at least one transcribed genomic region spanned by a ROA in the exact same location for nearly all individuals of a given population, which could be representative of past selection events. One such example is the LCT gene which was detected in an autozygosity hotspot in Northern European populations, with ROA frequency patterns consistent with positive selection at 2 different timepoints. Future work can therefore use wLOD to investigate which genes are disproportionately overlapped by ROA in the general population and further elucidate the role of recessive variation in human phenotypic diversity and common disease risk. Lastly, preliminary evidence suggests that the wLOD autozygosity score distinguishes between endogamic and consanguineous cultural processes in several Asian Indian groups in the TGP. However, additional comparative autozygosity analyses using well-defined endogamic Indian populations are needed to evaluate this property of wLOD and to shed light on the genomic architecture resulting from a long-established caste system.

To facilitate community adoption of the wLOD ROA inference method as well as classification based on genetic map length via a Gaussian mixture model, we have implemented these approaches in the software GARLIC (Genomic Autozygosity Regions Likelihood-based Inference and Classification) [401] that can be downloaded at: *https://github.com/szpiech/garlic*. As a guide, analysis of the 97 individuals in the CEU population on a Dell Precision T7600 workstation running RedHat Enterprise Linux (v.7.3) with multi-threading support enabled (16×2.60 GHz threads total) took ~ 2½ minutes for the OmniExp dataset, ~ 6½ minutes for the Omni2.5 dataset, and ~40 min for the WGS dataset, and occupied at most ~ 3 Gb, ~ 7 Gb, and ~ 20 Gb of RAM, respectively.

Chapter 4 Concluding remarks

The study of modern human populations has the potential to provide an unparalleled perspective on past evolutionary events, origins and demographic histories of modern human groups. Akin to the moulding of sedimentary rock by the environment, evolutionary forces can likewise mark a population's gene pool with distinctive patterns that persist across subsequent generations. In a sense, the evolutionary history of a species exists as a material archive in the population's genome.

Evolutionary processes tend to produce structurally recognizable patterns, meaning they can be mathematically modelled and contrasted against genomic patterns seen in real-world populations. Such models then allow inferences to be made on the possible scenarios that led to the presently observed genomic variation within or across populations. As an example, levels of genetic variation are determined by two opposing evolutionary forces: genetic drift and gene flow (migration). Across a geographical area, gene flow or the movement of genetic material between distinct populations will reduce genetic differences [459]. On the other hand, increased genetic drift (due to lowered population sizes) leads to a higher level of genetic variation across geographical space. In addition, environmental/sociocultural events have the potential to cause frequency shifts in either beneficial or harmful genetic variants within the gene pool which tends to increase the population structure across a geographic area [460][459].

Quantifying the spatial distribution of a given set of genetic markers or the overall magnitude of genomic diversity can inform on the size and distribution of ancestral groups, rates of population growth or decline, dispersal behaviours (migration events), mating patterns, measures of relatedness (levels of heterozygosity or autozygosity), divergence times between distinct human groups and areas of the genome influenced by selective forces [245][460]–[462][463].

This thesis examines present-day human variation and by doing so explores the legacy of past evolutionary events. Chapter 2 catalogs a series of complex anthropometric traits in Central African Pygmy populations whose unique phenotype is believed to have evolved, at least in part, as an adaptation to a tropical rainforest environment. Chapter 3 on the other hand describes a method to infer ancestral 'relatedness' signals in human populations which arise as a consequence of limited gene pools. In Chapter 2, the net effect of past evolutionary processes is the reduced

body size of Central African Pygmies studied via a comprehensive anthropometric analysis. Once the latter is paired with genetic data it should inform on both the evolutionary history of Central African populations and the molecular mechanisms involved in human scaling patterns and proportional growth of long bones. The net effect of past evolutionary events in Chapter 3 is studied via the distribution of homozygous haplotypes across populations. Autozygosity-detection tools like wLOD facilitate the exploration of the latter both within and across populations, which has the potential to inform on sociocultural processes of both ancestral and modern populations and more broadly fosters an improved understanding of human recessive variation and its role in complex traits and disease.

Appendix Tables for Chapter 2: Phenotypic profile of Central African Pygmies and non-Pygmies in Cameroon

	E1 = trait ~ Sex + ageCategories + PygmyStatus + standing height + (1 | Region)									Model selection					
	FIXED EFFECT						FIXED & RANDOM			AIC				AIC weights	
	For Pygmy status predictor														
Raw traits:	Estimate	Standard error	t value	LL	UL	Effect direction	R^2M	R^2C	ΔR^2	E*	E1	Δ	↓	E*	E1
Sitting height	0.62	0.37	1.71	-0.08	1.35		0.80	0.80	0.00	1470.4	1485.8	15.5	*	1.00	0.00
Leg length	-0.20	0.41	-0.50	-0.99	0.59		0.89	0.89	0.00	1533.7	1548.1	14.4	*	1.00	0.00
Tibia length	-0.18	0.26	-0.69	-0.69	0.31		0.82	0.82	0.00	1237.6	1256.6	18.9	*	1.00	0.00
Femur length	-0.04	0.35	-0.10	-0.68	0.67		0.75	0.75	0.00	1428.7	1444.5	15.8	*	1.00	0.00
Arm length	-0.09	0.60	-0.14	-1.23	1.15		0.65	0.66	0.00	1795.8	1805.3	9.6	*	0.99	0.01
Arm span	-0.05	1.09	-0.05	-2.12	2.16		0.78	0.78	0.00	2184.8	2187.4	2.5	*	0.78	0.22
Cubit length	0.30	0.36	0.81	-0.41	1.00		0.61	0.61	0.00	1465.6	1481.2	15.6	*	1.00	0.00
Humerus length	-0.46	0.33	-1.41	-1.08	0.19		0.55	0.59	0.04	1405.2	1408.2	3.0	*	0.82	0.18
Head circumference	-2.06	0.64	-3.21	-3.28	-0.74		0.22	0.25	0.03	1493.5	1500.4	6.9	*	0.97	0.03
Head length	-0.20	0.12	-1.71	-0.42	0.04		0.40	0.41	0.00	717.4	744.0	26.6	*	1.00	0.00
Head breadth	-0.59	0.10	-5.85	-0.78	-0.39		0.45	0.45	0.00	612.4	640.9	28.5	*	1.00	0.00
Shoulder breadth	0.17	0.44	0.40	-0.68	1.02		0.47	0.47	0.00	1586.0	1599.7	13.7	*	1.00	0.00
Hip breadth	-0.23	0.26	-0.90	-0.74	0.28		0.34	0.43	0.08	1250.2	1250.0	0.2	E1	0.48	0.52
Weight	-6.27	1.14	-5.50	-8.46	-4.01		0.60	0.65	0.05	2224.6	2208.9	15.7	E1	0.00	1.00
Waist circumference	-3.75	1.05	-3.56	-5.77	-1.65		0.25	0.31	0.06	2156.9	2150.1	6.8	E1	0.03	0.97
Tricep skin-fold	-3.16	0.60	-5.26	-4.32	-1.97		0.32	0.44	0.12	1786.7	1770.2	16.4	E1	0.00	1.00
Bicep skin-fold	-1.16	0.37	-3.14	-1.87	-0.43		0.12	0.21	0.10	1461.7	1461.6	0.1	E1	0.48	0.52
Calf skin-fold	-1.92	0.46	-4.16	-2.81	-1.00		0.25	0.41	0.16	1548.6	1528.6	20.0	E1	0.00	1.00
Suprailliac skin-fold	-0.99	0.43	-2.30	-1.80	-0.12		0.08	0.12	0.03	1542.9	1553.2	10.2	*	0.99	0.01
Wrist breadth	-1.61	0.51	-3.14	-2.60	-0.60		0.49	0.49	0.00	1691.9	1704.0	12.1	*	1.00	0.00
Calf circumference	-1.37	0.72	-1.91	-2.76	0.03		0.20	0.20	0.00	1910.1	1918.8	8.8	*	0.99	0.01
Bicep circumference	-3.26	0.43	-7.55	-4.08	-2.38		0.39	0.41	0.01	1578.0	1590.4	12.4	*	1.00	0.00
Ratio traits:															
Sitting height : s.height	0.00	0.00	1.72	0.00	0.01		0.25	0.25	0.00	-1839.0	-1773.1	65.9	*	1.00	0.00
Leg length : s.height	0.00	0.00	-0.60	-0.01	0.00		0.13	0.13	0.00	-1778.7	-1713.5	65.2	*	1.00	0.00
Tibia length: s.height	0.00	0.00	-0.29	-0.01	0.00		0.02	0.03	0.01	-1778.2	-1713.7	64.5	*	1.00	0.00
Arm length : s.height	0.00	0.00	-0.17	-0.01	0.01		0.04	0.05	0.01	-1509.5	-1449.3	60.2	*	1.00	0.00
Arm span : s.height	0.00	0.01	-0.05	-0.01	0.01		0.09	0.11	0.02	-1127.1	-1073.8	53.3	*	1.00	0.00
Cubit length : arm length	0.01	0.00	2.25	0.00	0.01		0.06	0.20	0.15	-1769.5	-1728.9	40.6	*	1.00	0.00
Humerus length : arm length	-0.01	0.00	-2.25	-0.01	0.00		0.06	0.20	0.15	-1769.5	-1728.9	40.6	*	1.00	0.00
Head length : s.height	0.00	0.00	-1.79	0.00	0.00		0.53	0.53	0.00	-2577.1	-2500.0	77.1	*	1.00	0.00
Shoulder breadth : hip breadth	0.02	0.02	1.10	-0.01	0.05		0.26	0.33	0.07	-603.9	-568.6	35.3	*	1.00	0.00
Shoulder breadth : s.height	0.00	0.00	0.44	0.00	0.01		0.07	0.07	0.00	-1713.3	-1649.1	64.2	*	1.00	0.00
WHtR	-0.02	0.01	-3.62	-0.04	-0.01		0.24	0.31	0.07	-1153.3	-1109.8	43.5	*	1.00	0.00
BMI	-2.55	0.43	-5.95	-3.38	-1.70		0.16	0.28	0.12	1591.4	1582.8	8.5	E1	0.01	0.99

Legend

Based on the 95% CI:

For PygmyStatus predictor:

Non-Pygmy > Pygmy for this trait

Pygmy > Non-Pygmy for this trait

Model selection

AIC is used to assess which model better represents collected data. E1 is contrasted to E*:

E* = trait ~ Sex + ageCategories + PygmyStatus + standing height

E1 = trait ~ Sex + ageCategories + PygmyStatus + standing height + **(1 | Region)**

↓- lowest AIC value (ie: best model according to the selected criterion)

The model with the larger AIC is defined as the Candidate model

Δ AIC - calculated *as the absolute difference* beween AIC obtained from E1 and E*

Δ AIC:

<3 Substantial evidence to support the candidate model

3 to 9 Candidate model has considerably less support

9+ Essentially no support for the candidate model

Appendix Table 1- Supplementary information for Table 2.6

A summary of hypothesis testing and model selection results using LMM E1 on anthropometric traits.

	E2						E1						Model selection			
	FIXED EFFECT						FIXED EFFECT						AIC			
Raw traits:	Estimate	Standard error	t value	LL	UL	Effect direction	Estimate	Standard error	t value	LL	UL	Effect direction	E2	**E1**	Δ	↓
Sitting height	-4.60	0.40	-11.56	-5.37	-3.82		0.62	0.37	1.71	-0.08	1.35		1744.5	1485.8	258.7	**E1**
Leg length	-8.56	0.56	-15.18	-9.66	-7.45		-0.20	0.41	-0.50	-0.99	0.59		1963.2	1548.1	415.0	**E1**
Tibia length	-4.01	0.29	-13.89	-4.57	-3.44		-0.18	0.26	-0.69	-0.69	0.31		1536.1	1256.6	279.6	**E1**
Femur length	-4.57	0.36	-12.69	-5.27	-3.86		-0.04	0.35	-0.10	-0.68	0.67		1674.3	1444.5	229.7	**E1**
Arm length	-5.44	0.53	-10.28	-6.48	-4.40		-0.09	0.60	-0.14	-1.23	1.15		1929.7	1805.3	124.4	**E1**
Arm span	-13.09	1.08	-12.11	-15.20	-10.97		-0.05	1.09	-0.05	-2.12	2.16		2390.9	2187.4	203.6	**E1**
Cubit length	-2.65	0.31	-8.45	-3.26	-2.03		0.30	0.36	0.81	-0.41	1.00		1588.5	1481.2	107.3	**E1**
Humerus length	-2.79	0.27	-10.40	-3.32	-2.27		-0.46	0.33	-1.41	-1.08	0.19		1492.0	1408.2	83.9	**E1**
Head circumference	-2.84	0.45	-6.28	-3.72	-1.95		-2.06	0.64	-3.21	-3.28	-0.74		1496.3	1500.4	4.1	E2
Head length	-0.57	0.09	-6.57	-0.74	-0.40		-0.20	0.12	-1.71	-0.42	0.04		752.7	744.0	8.7	**E1**
Head breadth	-0.74	0.07	-10.11	-0.88	-0.59		-0.59	0.10	-5.85	-0.78	-0.39		634.9	640.9	6.0	E2
Shoulder breadth	-2.21	0.34	-6.43	-2.88	-1.53		0.17	0.44	0.40	-0.68	1.02		1648.4	1599.7	48.7	**E1**
Hip breadth	-1.66	0.20	-8.19	-2.06	-1.26		-0.23	0.26	-0.90	-0.74	0.28		1297.4	1250.0	47.4	**E1**
Weight	-14.87	0.95	-15.57	-16.73	-13.00		-6.27	1.14	-5.50	-8.46	-4.01		2303.8	2208.9	94.9	**E1**
Waist circumference	-6.88	0.77	-8.99	-8.38	-5.38		-3.75	1.05	-3.56	-5.77	-1.65		2161.6	2150.1	11.5	**E1**
Tricep skin-fold	-3.20	0.42	-7.55	-4.03	-2.37		-3.16	0.60	-5.26	-4.32	-1.97		1763.2	1770.2	7.0	E2
Bicep skin-fold	-1.29	0.26	-4.93	-1.80	-0.78		-1.16	0.37	-3.14	-1.87	-0.43		1453.9	1461.6	7.7	E2
Calf skin-fold	-1.91	0.32	-5.93	-2.54	-1.28		-1.92	0.46	-4.16	-2.81	-1.00		1521.1	1528.6	7.5	E2
Suprailliac skin-fold	-1.39	0.30	-4.60	-1.99	-0.80		-0.99	0.43	-2.30	-1.80	-0.12		1547.3	1553.2	5.9	E2
Wrist breadth	-3.67	0.39	-9.48	-4.43	-2.91		-1.61	0.51	-3.14	-2.60	-0.60		1727.6	1704.0	23.6	**E1**
Calf circumference	-3.34	0.53	-6.30	-4.37	-2.30		-1.37	0.72	-1.91	-2.76	0.03		1927.8	1918.8	9.0	**E1**
Bicep circumference	-4.26	0.31	-13.52	-4.87	-3.64		-3.26	0.43	-7.55	-4.08	-2.38		1593.5	1590.4	3.1	**E1**
Ratio traits:																
Sitting height : s.height	0.01	0.00	7.26	0.01	0.02		0.00	0.00	1.72	0.00	0.01		-1763.8	-1773.1	9.2	**E1**
Leg length : s.height	-0.01	0.00	-4.41	-0.01	0.00		0.00	0.00	-0.60	-0.01	0.00		-1716.7	-1713.5	3.3	E2
Tibia length: s.height	0.00	0.00	0.48	0.00	0.00		0.00	0.00	-0.29	-0.01	0.00		-1730.8	-1713.7	17.1	E2
Arm length : s.height	0.00	0.00	0.79	0.00	0.01		0.00	0.00	-0.17	-0.01	0.01		-1465.3	-1449.3	16.0	E2
Arm span : s.height	0.01	0.00	1.71	0.00	0.02		0.00	0.01	-0.05	-0.01	0.01		-1086.3	-1073.8	12.6	E2
Cubit length : arm length	0.01	0.00	4.48	0.00	0.01		0.01	0.00	2.25	0.00	0.01		-1744.9	-1728.9	16.0	E2
Humerus length : arm length	-0.01	0.00	-4.48	-0.01	0.00		-0.01	0.00	-2.25	-0.01	0.00		-1744.9	-1728.9	16.0	E2
Head length : s.height	0.01	0.00	8.79	0.00	0.01		0.00	0.00	-1.79	0.00	0.00		-2367.6	-2500.0	132.4	**E1**
Shoulder breadth : hip breadth	0.00	0.01	0.37	-0.02	0.03		0.02	0.02	1.10	-0.01	0.05		-581.6	-568.6	12.9	E2
Shoulder breadth : s.height	0.00	0.00	2.04	0.00	0.01		0.00	0.00	0.44	0.00	0.01		-1664.6	-1649.1	15.5	E2
WHtR	0.00	0.00	-0.89	-0.01	0.01		-0.02	0.01	-3.62	-0.04	-0.01		-1108.7	-1109.8	1.2	**E1**
BMI	-2.56	0.31	-8.30	-3.16	-1.95		-2.55	0.43	-5.95	-3.38	-1.70		1575.2	1582.8	7.7	E2

Legend

Based on the 95% CI:

Non-Pygmy > Pygmy for this trait

Pygmy > Non-Pygmy for this trait

Results are concordant between LMM E1 and E2

Model selection

AIC is used to assess which model better represents collected data. E1 is contrasted to E2:

E2 = trait ~ Sex + ageCategories + PygmyStatus +(1 | Region)

E1 = trait ~ Sex + ageCategories + PygmyStatus + **standHeight** + (1 | Region)

↓ - lowest AIC value (ie: best model according to the selected criterion)

The model with the larger AIC is defined as the Candidate model

Δ AIC - calculated *as the absolute difference* beween AIC obtained from E1 and E2

Δ AIC:

<3	Substantial evidence to support the candidate model
3 to 9	Candidate model has considerably less support
9+	Essentially no support for the candidate model

APPENDIX TABLE 2-SUPPLEMENTARY INFORMATION FOR TABLE 2.7

A summary of hypothesis testing and model selection results using two hypothesis test procedures: LMM E1 and E2 conducted on anthropometric trait data.

Note: The regression coefficient and 95% CI data presented for E1 is identical to Appendix Table 1 and Table 2.6. Though R^2 statistics for E1 are not presented here, they can be found in the aforementioned tables.

E3 = trait ~ Sex + ageCategories + PygmyStatus + standing height + (PygmyStatus * standing height) + (1 | Region)

Raw traits:	FIXED EFFECT: For PygmyStatus Estimate	Standard error	t value	LL	UL	Effect direction	For PygmyStatus*standHeight Estimate	Standard error	t value	LL	UL	Effect direction	FIXED & RANDOM R^2M	R^2C	ΔR^2	Model Selection AIC E1	E3	Δ	↓
Sitting height	5.83	5.11	1.14	-4.22	15.70		-0.03	0.03	-1.02	-0.10	0.03		0.80	0.80	0.00	1485.8	1491.8	6.0	E1
Leg length	8.22	5.66	1.45	-2.81	19.25		-0.05	0.04	-1.49	-0.12	0.02		0.89	0.89	0.00	1548.1	1552.7	4.6	E1
Tibia length	1.98	3.58	0.55	-4.93	9.04		-0.01	0.02	-0.61	-0.06	0.03		0.82	0.82	0.00	1256.6	1263.9	7.4	E1
Femur length	6.32	4.82	1.31	-3.23	15.58		-0.04	0.03	-1.32	-0.10	0.02		0.75	0.75	0.00	1444.5	1449.9	5.4	E1
Arm length	14.28	8.36	1.71	-2.26	30.40		-0.09	0.05	-1.72	-0.19	0.01		0.66	0.66	0.01	1805.3	1808.4	3.1	E1
Arm span	25.76	15.13	1.70	-4.07	54.99		-0.16	0.10	-1.71	-0.35	0.03		0.78	0.79	0.01	2187.4	2189.3	1.9	E1
Cubit length	8.28	5.07	1.63	-1.59	18.14		-0.05	0.03	-1.58	-0.11	0.01		0.61	0.61	0.00	1481.2	1485.7	4.5	E1
Humerus length	6.32	4.51	1.40	-2.52	15.07		-0.04	0.03	-1.51	-0.10	0.01		0.55	0.60	0.05	1408.2	1413.1	5.0	E1
Head circumference	4.64	8.64	0.54	-12.24	21.45		-0.04	0.05	-0.78	-0.15	0.06		0.22	0.25	0.03	1500.4	1505.8	5.4	E1
Head length	1.57	1.64	0.96	-1.69	4.72		-0.01	0.01	-1.08	-0.03	0.01		0.40	0.41	0.00	744.0	752.1	8.1	E1
Head breadth	-1.08	1.40	-0.77	-3.81	1.66		0.00	0.01	0.35	-0.01	0.02		0.45	0.45	0.00	640.9	650.4	9.5	E1
Shoulder breadth	-3.01	6.11	-0.49	-14.91	8.89		0.02	0.04	0.52	-0.06	0.10		0.47	0.47	0.00	1599.7	1606.1	6.4	E1
Hip breadth	-6.02	3.60	-1.67	-13.07	0.98		0.04	0.02	1.61	-0.01	0.08		0.35	0.43	0.08	1250.0	1255.1	5.1	E1
Weight	7.12	15.84	0.45	-23.93	37.87		-0.09	0.10	-0.85	-0.28	0.11		0.60	0.65	0.05	2208.9	2212.9	4.0	E1
Waist circumference	-12.21	14.69	-0.83	-40.98	16.33		0.05	0.09	0.58	-0.13	0.24		0.25	0.31	0.06	2150.1	2154.7	4.6	E1
Tricep skin-fold	-19.86	8.30	-2.39	-36.10	-3.71		0.11	0.05	2.02	0.00	0.21	+	0.33	0.44	0.11	1770.2	1772.2	2.0	E1
Bicep skin-fold	-8.78	5.13	-1.71	-18.83	1.18		0.05	0.03	1.49	-0.01	0.11		0.12	0.21	0.09	1461.6	1466.4	4.8	E1
Calf skin-fold	-3.44	6.41	-0.54	-15.97	9.01		0.01	0.04	0.24	-0.07	0.09		0.25	0.40	0.15	1528.6	1535.1	6.5	E1
Suprailliac skin-fold	-2.20	5.95	-0.37	-13.90	9.33		0.01	0.04	0.20	-0.07	0.08		0.08	0.12	0.03	1553.2	1559.9	6.7	E1
Wrist breadth	-7.67	7.18	-1.07	-21.68	6.28		0.04	0.05	0.85	-0.05	0.13		0.49	0.49	0.00	1704.0	1709.6	5.6	E1
Calf circumference	-12.00	10.02	-1.20	-31.50	7.51		0.07	0.06	1.06	-0.06	0.19		0.20	0.20	0.00	1918.8	1923.4	4.5	E1
Bicep circumference	-16.54	5.97	-2.77	-28.35	-5.03		0.08	0.04	2.23	0.01	0.16	+	0.40	0.41	0.01	1590.4	1592.1	1.8	E1
Ratio traits:																			
Sitting height : s.height	0.06	0.03	1.74	-0.01	0.12		0.00	0.00	-1.62	0.00	0.00		0.25	0.26	0.01	-1773.1	-1758.6	14.5	E1
Leg length : s.height	0.04	0.04	1.03	-0.03	0.11		0.00	0.00	-1.08	0.00	0.00		0.13	0.13	0.00	-1713.5	-1697.7	15.8	E1
Tibia length: s.height	-0.03	0.04	-0.77	-0.10	0.04		0.00	0.00	0.75	0.00	0.00		0.02	0.04	0.01	-1713.7	-1697.3	16.4	E1
Arm length : s.height	0.09	0.05	1.66	-0.02	0.19		0.00	0.00	-1.67	0.00	0.00		0.05	0.06	0.01	-1449.3	-1435.9	13.3	E1
Arm span : s.height	0.16	0.10	1.65	-0.03	0.34		0.00	0.00	-1.66	0.00	0.00		0.10	0.12	0.02	-1073.8	-1061.5	12.2	E1
Cubit length : arm length	0.00	0.03	0.13	-0.06	0.07		0.00	0.00	0.04	0.00	0.00		0.06	0.20	0.15	-1728.9	-1711.9	17.0	E1
Humerus length : arm length	0.00	0.03	-0.13	-0.07	0.06		0.00	0.00	-0.04	0.00	0.00		0.06	0.20	0.15	-1728.9	-1711.9	17.0	E1
Head length : s.height	0.02	0.01	2.04	0.00	0.04		0.00	0.00	-2.18	0.00	0.00	-	0.53	0.54	0.01	-2500.0	-2485.3	14.7	E1
Shoulder breadth : hip breadth	0.17	0.21	0.84	-0.23	0.58		0.00	0.00	-0.76	0.00	0.00		0.26	0.33	0.06	-568.6	-555.8	12.8	E1
Shoulder breadth : s.height	-0.02	0.04	-0.51	-0.10	0.06		0.00	0.00	0.55	0.00	0.00		0.07	0.07	0.00	-1649.1	-1632.7	16.4	E1
WHtR	-0.05	0.09	-0.49	-0.22	0.13		0.00	0.00	0.24	0.00	0.00		0.16	0.28	0.12	-1109.8	-1094.8	15.0	E1
BMI	-4.15	5.98	-0.69	-15.87	7.46		0.01	0.04	0.27	-0.06	0.08		0.24	0.30	0.07	1582.8	1589.5	6.6	E1

Legend

Based on the 95% CI:

For Pygmy status predictor:
- Non-Pygmy > Pygmy for this trait
- Pygmy > Non-Pygmy for this trait

For PygmyStatus*standing height predictor:
- \+ LL and UL are both positive
- \- LL and UL are both negative

Model selection

AIC is used to assess which model better represents collected data. E3 is contrasted to E1:

E1 = trait ~ Sex + ageCategories + PygmyStatus + standing height+ (1 | Region)

E3 = trait ~ Sex + ageCategories + PygmyStatus + standing height + (**PygmyStatus*standing height)** + (1 | Region)

↓ - lowest AIC value (ie: best model according to the selected criterion)

The model with the larger AIC is defined as the Candidate model

Δ AIC - calculated *as the absolute difference* beween AIC obtained from E3 and E1

Δ AIC:
- <3 Substantial evidence to support the candidate model
- 3 to 9 Candidate model has considerably less support
- 9+ Essentially no support for the candidate model

APPENDIX TABLE 3-SUPPLEMENTARY INFORMATION FOR TABLE 2.8

A summary of hypothesis testing and model selection results using LMM E3 conducted on anthropometric trait data.

	E1 = trait ~ Sex + ageCategories + PygmyStatus + BMI + (1 | Region)									Model selection					
	FIXED EFFECT						FIXED & RANDOM			AIC				AIC weights	
	For Pygmy status predictor														
	Estimate	Standard error	t value	LL	UL	Effect direction	R^2M	R^2C	ΔR^2	E*	E1	Δ	↓	E*	E1
Total cholesterol	-6.85	4.47	-1.53	-15.47	1.95		0.11	0.11	0.00	3065.9	3052.3	13.6	E1	0.00	1.00
HDL cholesterol	-7.11	2.15	-3.30	-11.31	-2.91		0.08	0.08	0.00	2615.5	2609.1	6.3	E1	0.04	0.96
non-HDL cholesterol	1.16	4.16	0.28	-6.69	9.60		0.10	0.11	0.01	2132.8	2119.5	13.3	E1	0.00	1.00
LDL cholesterol	-1.06	3.90	-0.27	-8.52	6.75		0.08	0.14	0.06	2024.3	2007.9	16.4	E1	0.00	1.00
LDL : HDL cholesterol	0.12	0.13	0.97	-0.12	0.38		0.03	0.06	0.03	524.6	545.2	20.6	*	1.00	0.00
Total : HDL cholesterol	0.18	0.16	1.18	-0.12	0.49		0.04	0.04	0.00	650.1	670.1	20.0	*	1.00	0.00
SBP	3.17	1.86	1.71	-0.45	6.79		0.12	0.12	0.00	2641.8	2637.1	4.8	E1	0.08	0.92
DBP	-0.35	1.26	-0.28	-2.92	2.00		0.08	0.09	0.01	2388.8	2387.6	1.2	E1	0.36	0.64
Pulse	-3.14	1.72	-1.82	-6.63	0.13		0.06	0.08	0.03	2592.0	2585.5	6.5	E1	0.04	0.96
Pulse pressure	3.23	1.19	2.72	0.96	5.61		0.08	0.13	0.04	2355.4	2349.8	5.6	E1	0.06	0.94
MAP	0.74	1.37	0.54	-1.93	3.41		0.10	0.10	0.00	2446.9	2445.2	1.7	E1	0.30	0.70
Triglycerides	6.78	6.91	0.98	-7.06	20.02		0.02	0.06	0.04	3339.1	3316.9	22.2	E1	0.00	1.00
Glucose	6.19	2.72	2.28	0.80	11.43		0.02	0.11	0.09	2770.0	2749.6	20.4	E1	0.00	1.00

Legend

Based on the 95% CI:

For PygmyStatus predictor:

Non-Pygmy > Pygmy for this trait

Pygmy > Non-Pygmy for this trait

Model selection

AIC is used to assess which model better represents collected data. E1 is contrasted to E*:

E* = trait ~ Sex + ageCategories + PygmyStatus + BMI

E1 = trait ~ Sex + ageCategories + PygmyStatus + BMI + **(1 | Region)**

↓- lowest AIC value (ie: best model according to the selected criterion)

The model with the larger AIC is defined as the Candidate model

Δ AIC - calculated *as the absolute difference* beween AIC obtained from E1 and E*

Δ AIC:

<3 Substantial evidence to support the candidate model

3 to 9 Candidate model has considerably less support

9+ Essentially no support for the candidate model

APPENDIX TABLE 4-SUPPLEMENTARY INFORMATION FOR TABLE 2.10

A summary of hypothesis testing and model selection results using LMM E1 on anthropometric traits.

Appendix Tables for Chapter 3: A weighted likelihood inference method to detect autozygosity patterns in genotypic data)

SNV subset	Scenario 1		Scenario 2	
	Window size	% overlap	Window size	% overlap
18,000	60	7	70	5
50,000	70	18	80	19
80,000	80	28	90	22
125,000	80	29	100	26
750,0000	130	37	120	32

APPENDIX TABLE 5-OPTIMAL WINDOW SIZE AND OVERLAP PROPORTION IN SIMULATED DATASETS USING WLOD
Table S1 in original publication [360].

Genomic region			Genes	Populations
Chr	Begin (bp)	End (bp)		
8	7,118,140	7,143,880	*LINC00965*	ASW, LWK
8	7,627,105	7,628,835	*FAM90A10P*	LWK, TSI
15	28,818,925	28,834,386	*HERC2P11*	CEU
21	9,825,831	9,826,263	*MIR3648, MIR3687*	CEU, JPT, LWK, TSI, GBR, CDX, GWD
21	9,907,188	9968594	*TEKT4P2*	CDX
21	11,020,841	11,098,937	*BAGE, BAGE2, BAGE3, BAGE4, BAGE5*	TSI, GBR, IBS, CLM, PUR
21	14,982,497	15,013,906	*POTED*	LWK, CDX

APPENDIX TABLE 6-TRANSCRIBED GENOMIC REGIONS SPANNED BY CLASS 5 ROA IN >90% OF AT LEAST ONE POPULATION
Seven transcribed regions (listed under Genes column) found across 3 chromosomes are spanned by class 5 ROAs in more than 90% of individuals for at least one of the listed populations.
Table S3 in original publication [360].

Genomic region			Genes	Populations
Chr	Begin (bp)	End (bp)		
1	12,998,504	13,002,353	*PRAMEF9*	TSI
1	206,317,458	206,332,104	*CTSE*	JPT, CHS
2	111,132,685	111,142,113	*LINC01106*	TSI
2	131,220,388	131,357,148	*POTEI*, ***CFC1***	TSI, CLM, PUR
4	9,212,382	9,370,796	*USP17L5, USP17L10, USP17L11, USP17L12, USP17L13, USP17L15, USP17L17, USP17L18, USP17L19, USP17L20, USP17L21, USP17L22, USP17L24, USP17252, USP17L26, USP17L27, USP17L28, USP17L29, USP17L30, USP17L6P, USP17L9P*	TSI
4	41,937,136	41,988,484	*TMEM33, DCAF4L1*	CDX
4	70,146,216	70,160,768	*UGT2B28*	ASW, LWK, YRI, ACB, ESN, GWD
5	69,140,495	74,867,509	*LOC653080, SERF1A, SERF1B,* ***SMN1****, SMN2, SMA4, GTF2H2B, SMA5, LOC441081, GUSBP9, NAIP, GTF2H2, LOC647859*	CDX
7	74,572,383	74,867,509	*NCF1C, CASTOR2*	CDX
8	7,118,140	7,143,880	*LINC00965*	PJL
8	7,812,534	7,946,611	*FAM66E, USP17L8, MIR548I3, FOXD4L5*	CDX, BEB, ITU, PJL
10	17,794,259	18,134,122	*TMEM236, MIR511*	CDX
10	48,155,942	48,158,691	*CTSLP2*	LWK, GWD
10	51,253,907	51,371,344	*PARGP1*	MXL, CDX, ITU
14	101,391,157	101,454,566	*SNORD113-1, SNORD113-2, SNORD113-4, SNORD113-5, SNORD113-6, SNORD113-7, SNORD113-9, SNORD114-1, SNORD114-2, SNORD114-3, SNORD114-4, SNORD114-5, SNORD114-6, SNORD114-7, SNORD114-8, SNORD114-9, SNORD114-10, SNORD114-11, SNORD114-12, SNORD114-13, SNORD114-14, SNORD114-15, SNORD114-16, SNORD114-17, SNORD114-18, SNORD114-19, SNORD114-20, SNORD114-21, SNORD114-22, SNORD114-23, SNORD114-24, SNORD114-25, SNORD114-26, SNORD114-27*	JPT, CDX
14	106,383,837	106,438,358	*FAM30A, ADAM6*	CEU, CHB, CIH, JPT, MXL, TSI, FIN, GBR, IBS, CLM, PUR, BEB, ESN, GWD, MSL
15	30,653,442	30,782,516	*CHRFAM7A, GOLGA8R, LOC100288203*	CDX
15	102,462,344	102,463,262	*OR4F4*	LWK, GWD
16	21,413,454	21,445,776	*NPIPB3, LOC100190986*	PUR
19	110,678	202,209	*OR4F17, LINC01002*	CDX
19	55,280,873	55,360,024	*ZNF818P, LOC101928804, KIR2DL1, KIR3DL1, KIR2DS4*	TSI, GBR, CDX, PEL, BEB, ITU, PJL STU
21	17,911,408	17,912,231	*MIR99A, MIRLET7C*	CDX

OMIM autosomal recessive genes are in **bold**.

APPENDIX TABLE 7-TRANSCRIBED GENOMIC REGIONS SPANNED BY CLASS 4 ROA IN >90% OF AT LEAST ONE POPULATION

Twenty-two transcribed regions (listed under Genes column) found across 12 chromosomes are spanned by intermediate-length class 4 ROAs in more than 90% of individuals for at least one of the listed populations.

Table S4 in original publication [360].

Genomic region			Genes	Populations
Chr	Begin (bp)	End (bp)		
1	248,756,130	248,790,429	*OR2T10, OR2T11*	ASW, LWK, YRI, ACB, GWD, MSL
2	114,256,660	114,258,727	*FOXD4L1*	CDX
4	70,146,216	70,160,768	*UGT2B28*	CEU, MXL, TSI, GBR, IBS, BEB, ITU, PJL, PEL
9	46,116,942	46,168,270	*LOC105376064, FAM27E2*	TSI
11	61,567,096	61,582,712	*FADS1, MIR1908*	ITU, STU
13	19,582,398	19,586,774	*LINC00442*	PEL
16	2,688,982	2,696,130	*FLJ42627*	CDX
16	21,413,454	21,445,776	*NPIPB3, LOC100190986*	CDX
19	43,715,942	43,752,798	*LOC284344*	KHV

APPENDIX TABLE 8-TRANSCRIBED GENOMIC REGIONS SPANNED BY CLASS 3 ROA IN >90% OF AT LEAST ONE POPULATION

Nine transcribed regions (listed under Genes column) found across 8 chromosomes are spanned by short-length class 3 ROAs in more than 90% of individuals for at least one of the listed populations.
Table S5 in original publication [360].

Genomic region			Genes	Populations
Chr	Begin (bp)	End (bp)		
8	41,284,173	41,284,246	*SNORD65B*	CHB
9	44,384,584	44,391,314	*LOC101927827*	MXL, BEB, ITU
11	62,334,482	62,334,543	*MIR6747*	CDX
15	25,440,067	25,440,148	*SNORD115-14*	CLM
20	62,159,775	62,168,723	*PTK6*	CHB

APPENDIX TABLE 9-TRANSCRIBED GENOMIC REGIONS SPANNED BY CLASS 2 ROA IN >90% OF AT LEAST ONE POPULATION

Five transcribed regions (listed under Genes column) found across 5 chromosomes are spanned by short-length class 2 ROAs in more than 90% of individuals for at least one of the listed populations.
Table S6 in original publication [360].

Genomic region			Genes	Populations
Chr	Begin (bp)	End (bp)		
1	16,860,385	16,866,530	*LINC01783, FAM231B*	CDX
1	50,883,222	50,889,119	*DMRTA2*	JPT, CHS, CDX
1	120,839,004	120,855,681	*FAM72B*	MXL, CDX
1	142,697,420	142,713,605	*ANKRD20A12P*	CDX
2	91,824,708	91,970,153	*LOC654342, GGT8P*	MXL, CDX
3	48,658,274	52,029,958	*TMEM89, MIR6824, MIR4793,* ***SLC25A20****, ARIH2OS, P4HTM, WDR6, DALRD3, MIR435,* ***NDUFAF3****, MIR191, IMPDH2, QARS, MIR6890,* ***LAMB2****, LAMB2P1, CCDC71, KLHDC8B, C3orf84, C3orf62, MIR4271,* ***GPX1****, TCTA, APEH, UBA7, MIR5193, CAMKV, RBM5,* ***NPRL2****, CYB561D2, TMEM115, MIR4787, IQCF2, IQCF3, IQCF4, IQCF5, RRP9, PARP3,* ***ACY1****, RPL29*	CHB, GIH, JPT, MXL, TSI, GBR, CHS, PUR, CDX, PEL, KHV, BEB, ITU, PJL, STU
6	27,114,860	27,925,960	*HIST1H2AH, MIR3143, LOC100131289, HIST1H2BL, HIST1H2AI, HIST1H3H, HIST1H2AJ, HIST1H2BM, HIST1H4J, HIST1H4K, HIST1H2AK, HIST1H2BN, HIST1H2AL, HIST1H1B, HIST1H4L, HIST1H3J, HIST1H2AM, HIST1H2BO, OR2B2, OR2B6*	CEU, JPT, MXL, TSI, IBS, CHS, CDX, PEL
6	86,386,724	86,387,377	*SNHG5, SNORD50A, SNORD50B*	CDX
7	27,168,125	27,170,399	*HOXA4*	ITU
8	10,524,487	10,524,580	*MIR4286*	CHB
8	42,691,816	42,751,418	*THAP1, MIR4469*	TSI, JPT, FIN, IBS, ITU
9	43,608,372	43,630,730	*FAM74A7, SPATA31A6*	CEU, TSI, YRI, FIN, GBR, IBS, CDX, ACB, ITU, STU
9	44,384,584	44,391,314	*LOC101927827*	CDX
9	66,494,268	66,555,601	*PTGER4P2-CDK2AP2P2, LOC728673*	TSI, IBS, GWD, STU
9	99,671,356	99,704,572	*LOC441454, NUTM2G*	CEU, GBR, IBS, CLM, PUR
10	74,870,132	75,193,319	*NUDT13, SNORA11F, ECD, DNAJC9,* ***MRPS16****, MSS51*	MXL, TSI, FIN, GBR, IBS, PUR, PEL, STU
10	75,404,638	75,415,863	*SYNPO2L*	TSI
11	38,670,407	38,676,799	*LINC01493*	MXL, TSI
11	46,624,855	46,638,777	*HARBI1*	PEL
11	50,003,008	50,004,071	*OR4C12*	MXL, CDX
11	51,411,377	51,516,211	*OR4CA5, OR4C46*	MXL
11	55,110,676	55,433,572	*OR4A16, OR4C11, OR4P4, OR4S2, OR4C6*	ASW, CEU, GIH, LWK, MXL, TSI, YRI, GBR, CLM, PUR, CDX, ACB, BEB, GWD, ITU, MSL, PJL, STU
11	55,606,227	57,429,337	*OR5D16, OR5I1, OR10AG1, OR7E5P, OR5F1, OR5AS1, OR8I2, OR8H2, OR8H3, OR8J3, OR5J2, OR5T1, OR8K3, OR8K1, OR8U8, OR8U1, OR5AL1, OR5R1, OR5M8, OR5M11, OR5M10, OR5M1, TIMM10, MIR130A, CLP1*	MXL, CDX, PEL, KHV
11	89,819,117	89,820,299	*UBTFL1*	CDX
12	34,175,215	34,181,236	*ALG10*	JPT, CHS, CDX, KHV
12	111,374,405	111,395,622	*LINC01405, LOC10536980*	CDX, PEL
12	123,745,516	123,756,863	*CDK2AP1*	JPT
14	59,930,239	59,972,124	*GPR135, L3HYPDH, JKAMP*	CHB, CDX
14	67,908,571	67,908,647	*MIR5694*	CDX
15	24,686,273	24,693,114	*PWRN3*	ASW, CEU, JPT, LWK, TSI, FIN, IBS, CLM, PUR, ACB, BEB, GWD, ITU, MSL, PJL
15	42,491,767	42,491,864	*MIR627*	IBS
15	72,668,453	72,879,654	*TMEM202, MIR630*	GIH, CDX, BEB, PJL, STU
16	14,397,823	14,403,228	*MIR193B, MIR365A*	CDX
16	32,888,796	32,896,463	*SLC6A10P*	BEB
16	33,961,051	33,962,503	*LINC00273*	CEU, GIH, TSI, IBS, CLM, PUR, CDX, BEB, ITU, PJL, STU
16	34,403,801	34,404,762	*UBE2MP1, FRG2DP*	MXL, TSI, GWD, STU
16	46,723,557	46,920,386	*ORC6, SNORD148*	CDX, PEL, KHV
16	47,999,599	48,005,598	*LINC02134*	IBS, CDX, BEB
16	67,906,925	68,267,402	*EDC4, NRN1L, PSMB10,* ***LCAT****, DPEP3, ESRP2, MIR6773*	CHS, CDX, CHB
17	28,951,335	28,953,825	*SH3GL1P2*	TSI, GBR, IBS
19	28,281,400	28,284,848	*LINC00662*	CHB, CDX
20	25,593,572	26,188,914	*NANP, LINC01733, NCOR1P1, MIR663A*	CDX
20	29,637,583	30,619,984	*MLLT10P1, DEFB121, 1D1, MIR3193,* ***COX4I2****, ABALON, CCM2L*	CEU, CHB, GIH, MXL, IBS, CHS, CDX, PEL, KHV, BEB, ITU, PJL, STU
20	34,020,826	34,288,902	*GDF5OS, MIR1289-1, RBM12, ROMO1*	JPT, CDX, PEL, IBS, CLM
22	42,086,546	42,094,140	*C22orf46*	JPT
22	43,608,679	43,609,667	*LOC105373051*	CDX

OMIM autosomal dominant genes are underlined.
OMIM autosomal recessive genes are in **bold**.

APPENDIX TABLE 10-TRANSCRIBED GENOMIC REGIONS SPANNED BY CLASS 1 ROA IN >90% OF AT LEAST ONE POPULATION

Forty-six transcribed regions (listed under Genes column) found across 18 chromosomes are spanned by short-length class 1 ROAs in more than 90% of individuals for at least one of the listed populations.
Table S7 in original publication [360].

Citations -Chapter 1

[1] J. Huxley, *Evolution : the modern synthesis*. Harper, 1942.

[2] S. Casson, *The discovery of man*. London, 1940.

[3] J. H. Rowe, "The Renaissance Foundations of Anthropology," *Am. Anthropol.*, vol. 67, no. 1, pp. 1–20, Feb. 1965, doi: 10.1525/aa.1965.67.1.02a00010.

[4] S. Quirke, "Race in Ancient Egypt," *University College London*, 2003. https://www.ucl.ac.uk/museums-static/digitalegypt/social/race.html (accessed Apr. 28, 2021).

[5] J. Monge-nájera, "Quantitative biological data in Pliny's Naturalis Historia," 2008.

[6] N. G. Jablonski, "Skin color and race," *Am J Phys Anthr.*, vol. 175, pp. 437–447, 2020, doi: 10.1002/ajpa.24200.

[7] W. Hodos, "Evolution and the Scala Naturae," in *Encyclopedia of Neuroscience*, Springer Berlin Heidelberg, 2008, pp. 1212–1215.

[8] M. E. Svensson, "How Linnaeus classified humans: why red, white, yellow and black people were assigned particular temperaments," *Ann. Hist. Philos. Biol.*, vol. 17, pp. 303–315, 2012.

[9] APA Dictionary of Psychology, "Humoral theory." https://dictionary.apa.org/humoral-theory (accessed Jun. 02, 2021).

[10] Britannica, "Johann Friedrich Blumenbach, German anthropologist." https://www.britannica.com/biography/Johann-Friedrich-Blumenbach (accessed Aug. 22, 2021).

[11] R. J. Richards, "The beautiful skulls of Schiller and the Georgian girl: Quantitative and aesthetic scaling of the races, 1770-1850," *Johann Friedrich Blumenbach Race Nat. Hist. 1750-1850*, pp. 142–176, 2018, doi: 10.4324/9781315184777.

[12] J. Blumenbach, *The anthropological treatises of Johann Friedrich Blumenbach : Blumenbach, Johann Friedrich, 1752-1840 : Free Download, Borrow, and Streaming : Internet Archive*. London, 1865.

[13] W. Bodmer, "Genetic characterization of human populations: From abo to a genetic map of the british people," *Genetics*, vol. 199, no. 2, pp. 267–279, Feb. 2015, doi: 10.1534/genetics.114.173062.

[14] M. E. Steiper, "DNA Markers of Human Variation," in *Human Evolutionary Biology*,

Cambridge University Press, 2010, pp. 238–64.

[15] K. Landsteiner, “Agglutination phenomena of normal human blood,” *Wien Klin Wochenschr*, vol. 113, no. 20–21, pp. 768–9, Oct. 1901.

[16] L. Hirschfeld and H. Hirschfeld, “Serological differences between the blood of different races; the result of research on the Macadonian front,” *Lancet*, vol. 194, no. 5016, pp. 675–679, Oct. 1919, doi: 10.1016/S0140-6736(01)48686-7.

[17] S. Wooding, “Phenylthiocarbamide: A 75-Year Adventure in Genetics and Natural Selection.”

[18] E. Matsunaga, “The dimorphism in human normal cerumen,” *Ann. Hum. Genet.*, vol. 25, no. 4, pp. 273–286, May 1962, doi: 10.1111/J.1469-1809.1962.TB01766.X.

[19] Sahi T, “The inheritance of selective adult-type lactose malabsorption,” *Scand J Gastroenterol Suppl.*, vol. 30, 1974.

[20] A. L. Fox, “The Relationship Between Chemical Constitution and Taste,” *Proc. Natl. Acad. Sci.*, vol. 18, no. 1, pp. 115–120, 1932, doi: 10.1073/pnas.18.1.115.

[21] B. W. Cavalli-Sforza LL, “Human Evolution,” in *The genetics of human populations*, San Francisco: WH Freeman and Company, 1971, pp. 695–717.

[22] R. Meier, “Classic Markers of Human Variation,” in *Human Evolutionary Biology*, M. MP, Ed. Cambridge University Press, 2010, pp. 214–237.

[23] P. E. Carson, C. L. Flanagan, C. E. Ickes, and A. S. Alving, “Enzymatic deficiency in primaquine-sensitive erythrocytes,” *Science (80-.).*, vol. 124, no. 3220, pp. 484–485, Sep. 1956, doi: 10.1126/science.124.3220.484-a.

[24] L. S. Greene, “G6PD Deficiency as Protection Against falciparum Malaria: An Epidemiologic Critique of Population and Experimental Studies,” 1993.

[25] “G6PD glucose-6-phosphate dehydrogenase [Homo sapiens (human)],” *NCBI - Gene*, 2021. https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=2539 (accessed Aug. 30, 2021).

[26] “HLA-A major histocompatibility complex, class I, A [Homo sapiens (human)] ,” *NCBI - Gene*, Aug. 28, 2021. https://www.ncbi.nlm.nih.gov/gene/3105 (accessed Aug. 30, 2021).

[27] S. Buhler and A. Sanchez-Mazas, “HLA DNA Sequence Variation among Human Populations: Molecular Signatures of Demographic and Selective Events,” *PLoS One*, vol. 6, no. 2, p. 14643, 2011, doi: 10.1371/journal.pone.0014643.

[28] E. Thorsby, “A short history of HLA,” *Tissue Antigens*, vol. 74, no. 2. pp. 101–116, Aug. 2009, doi: 10.1111/j.1399-0039.2009.01291.x.

[29] A. Sanchez-Mazas *et al.*, “Immunogenetics as a tool in anthropological studies,” *Immunology*, vol. 133, no. 2. pp. 143–164, Jun. 2011, doi: 10.1111/j.1365-2567.2011.03438.x.

[30] L. L. Cavalli-Sforza and A. W. F. Edwards, “Phylogenetic Analysis Models and Estimation Procedures,” 1967.

[31] W. C. Boyd, “Use of blood groups in human classification,” *Science (80-.).*, vol. 112, no. 2903, pp. 187–196, 1950, doi: 10.1126/science.112.2903.187.

[32] A. E. Mourant, *The Distribution of the Human Blood Groups*. Oxford: Blackwell Publishing, 1954.

[33] O. Franc *et al.*, “Principal Component Analysis under Population Genetic Models of Range Expansion and Admixture,” 2010, doi: 10.1093/molbev/msq010.

[34] K. B. Mullis and F. A. Faloona, “Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction,” *Methods Enzymol.*, vol. 155, no. C, pp. 335–350, Jan. 1987, doi: 10.1016/0076-6879(87)55023-6.

[35] D. Gresham, M. J. Dunham, and D. Botstein, “Comparing whole genomes using DNA microarrays,” *Nat. Rev. Genet. 2008 94*, vol. 9, no. 4, pp. 291–302, Apr. 2008, doi: 10.1038/nrg2335.

[36] M. Jakobsson *et al.*, “Genotype, haplotype and copy-number variation in worldwide human populations,” *Nature*, vol. 451, no. 7181, pp. 998–1003, Feb. 2008, doi: 10.1038/nature06742.

[37] J. O. Korbel *et al.*, “Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome Europe PMC Funders Group,” *Science (80-.).*, vol. 318, no. 5849, pp. 420–426, 2007, doi: 10.1126/science.1149504.

[38] D. V Babushok and H. H. Kazazian, “Progress in Understanding the Biology of the Human Mutagen LINE-1,” *Hum. Mutat.*, vol. 28, no. 6, 2007, doi: 10.1002/humu.20486.

[39] M. Kato *et al.*, “Population-genetic nature of copy number variations in the human genome,” *Hum. Mol. Genet.*, vol. 19, no. 5, pp. 761–773, 2010, doi: 10.1093/hmg/ddp541.

[40] D. G. Wang *et al.*, “Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome,” 1998.

[41] C. RL, S. M, and W. AC, “Mitochondrial DNA and human evolution,” *Nature*, vol. 325, no. 6099, pp. 31–36, 1987, doi: 10.1038/325031A0.

[42] J. L. Weber and P. E. May, “Abundant Class of Human DNA Polymorphisms Which Can Be Typed Using the Polymerase Chain Reaction,” 1989.

[43] N. Kinney *et al.*, “Abundance of ethnically biased microsatellites in human gene regions,” *PLoS One*, vol. 14, no. 12, p. e0225216, Dec. 2019, doi: 10.1371/JOURNAL.PONE.0225216.

[44] N. E. Morton, “Sequential Tests for the Detection of Linkage,” *Am. J. Hum. Genet.*, vol. 7, no. 3, pp. 277–318, 1955.

[45] T. H. Morgan, “What are ‘factors’ in Mendelian explanations?,” *Am. Breeders Assoc. Reports*, vol. 5, pp. 365–368, 1909.

[46] Hardy G.H, “Mendelian Proportions in a Mixed Population,” *Science (80-.).*, vol. 28, no. 706, pp. 49–50, Jul. 1908, doi: 10.1126/science.28.706.49.

[47] R.A Fisher, “The Correlation between Relatives on the Supposition of Mendelian Inheritance,” *Trans. R. Soc. Edinburgh*, vol. 52, no. 2, p. pp 399-433, 1918.

[48] R. B. Robbins, “Some applications of mathematics to breeding problems III,” 1917.

[49] Bateson M.A, Saunders E.R, and Punett R.C, “Experimental studies in the physiology of heredity,” *Reports to Evol. Comm. R. Soc.*, 1904.

[50] Lobo I and Shaw K, “Discovery and types of genetic linkage,” *Nature Education*, 2008. https://www.nature.com/scitable/topicpage/discovery-and-types-of-genetic-linkage-500/ (accessed Aug. 28, 2021).

[51] K. W. Broman and J. L. Weber, “Long homozygous chromosomal segments in reference families from the centre d’Etude du polymorphisme humain.,” *Am. J. Hum. Genet.*, vol. 65, no. 6, pp. 1493–500, Dec. 1999, doi: 10.1086/302661.

[52] N. A. Rosenberg *et al.*, “Genetic structure of human populations,” *Science (80-.).*, vol. 298, no. 5602, pp. 2381–2385, Dec. 2002, doi: 10.1126/science.1078311.

[53] C. HM *et al.*, “A human genome diversity cell line panel,” *Science*, vol. 296, no. 5566, pp. 261b – 262, Apr. 2002, doi: 10.1126/SCIENCE.296.5566.261B.

[54] L. L. Cavalli-Sforza, “The Human Genome Diversity Project: past, present and future,” *Nat. Rev. Genet.*, vol. 6, no. 4, pp. 333–340, Apr. 2005, doi: 10.1038/nrg1596.

[55] “Genome-Wide Association Studies Fact Sheet,” *National HUman Genome Research*

Institute. https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet (accessed Aug. 28, 2021).

[56] S. Ikegawa, "A Short History of the Genome-Wide Association Study: Where We Were and Where We Are Going," *Genomics Inform.*, vol. 10, no. 4, p. 220, 2012, doi: 10.5808/GI.2012.10.4.220.

[57] GWAS Catalog Blog, "Trait annotation in the GWAS catalog," Aug. 2021. https://ebispot.github.io/gwas-blog/background-trait-update (accessed Aug. 28, 2021).

[58] T. I. H. Consortium*, "The International HapMap Project," *Nature*, vol. 426, 2003.

[59] Donnelly Peter, "Progress and challenges in genome-wide association studies in humans.: EBSCOhost," *Nature*, vol. 456, 2008.

[60] T. I. H. 3 Consortium, "Integrating common and rare genetic variation in diverse human populations," *Nature*, vol. 467, no. 7311, pp. 52–58, Sep. 2010, doi: 10.1038/nature09298.

[61] "1000 Genomes Phase 3 release | IGSR data collection." https://www.internationalgenome.org/data-portal/data-collection/phase-3 (accessed Aug. 22, 2021).

[62] S. C. Darwin and A. Wallace, "On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection," *Zool. J. Linn. Soc. London*, pp. 11–24, 1858, doi: https://doi.org/10.1111/j.1096-3642.1858.tb02500.x.

[63] P. Sloan, "Evolutionary Thought Before Darwin," *Stanford Encyclopedia of Philosophy*, 2019. https://plato.stanford.edu/entries/evolution-before-darwin/#SpecPermChanAnti (accessed Aug. 29, 2021).

[64] T.-S. C. Jobling Mark, Hollox Edward, Hurles Matthew, Kivisild Toomas, *Human Evolutionary Genetics*, 2nd ed., no. 4. Garland Science, 2014.

[65] C. Dandine-Roulland, R. Laurent, I. Dall'Ara, B. Toupance, and R. Chaix, "Genomic evidence for MHC disassortative mating in humans," *Proceedings. Biol. Sci.*, vol. 286, no. 1899, p. 20182664, 2019, doi: 10.1098/rspb.2018.2664.

[66] R. Sebro, G. M. Peloso, J. Dupuis, and N. J. Risch, "Structured mating: Patterns and implications," *PLoS Genet.*, vol. 13, no. 4, 2017, doi: 10.1371/journal.pgen.1006655.

[67] N. Risch *et al.*, "Ancestry-related assortative mating in Latino populations," *Genome Biol.*, vol. 10, no. 11, p. 132, 2009, doi: 10.1186/gb-2009-10-11-r132.

[68] Y.-M. Hur, "Assortative Mating for Personality Traits, Educational Level, Religious

Affiliation, Height, Weight, and Body Mass Index in Parents of a Korean Twin Sample," *Twin Res.*, vol. 6, no. 6, pp. 467–470, 2021, doi: 10.1375/twin.6.6.467.

[69] S. Wright, "On the roles of directed and random changes in gene frequency in the genetics of populations," *Int. J. Org. Evol.*, vol. 2, no. 4, 1948.

Citations – Chapters 2 & 3

[1] S. Ulijaszek and J. Komlos, “From a History of Anthropometry to Anthropometric History,” in *Human Variation: From the Laboratory to the Field*, CRC Press, 2010, pp. 183–197.

[2] M. H. Stone, “The Cubit: A History and Measurement Commentary,” *J. Anthropol.*, vol. 2014, pp. 1–11, 2014, doi: 10.1155/2014/489757.

[3] N. Utkualp and I. Ercan, “Anthropometric Measurements Usage in Medical Sciences,” 2015, doi: 10.1155/2015/404261.

[4] L. Finlay, “Craniometry and Cephalometry: A History Prior to the Advent of Radiography,” *Angle Orthod.*, vol. 50, no. 4, pp. 312–321, 1980.

[5] D. J. P. Ferreira M Piedade, Dabral de Mello Duarte, “The Grammar of Movement: A Step Towards a Corporeal Architecture,” *Nexus Netw. J.*, vol. 13, no. 1, pp. 131–149, 2011, doi: 10.1007/s00004-011-0058-4.

[6] A. Albrizio, “Biometry and Anthropometry: from Galton to Constitutional Medicine,” 2007.

[7] “Craniometry | Encyclopedia.com.” https://www.encyclopedia.com/medicine/divisions-diagnostics-and-procedures/medicine/craniometry (accessed Dec. 14, 2020).

[8] P. Roy, *The Cambridge history of science: Eighteenth-century science*, vol. 4. Cambridge University Press, 2003.

[9] “Anthropometry - Definition, History and Applications | Biology Dictionary.” https://biologydictionary.net/anthropometry/ (accessed Dec. 14, 2020).

[10] J. M. Hecht, “Vacher de Lapouge and the rise of Nazi science,” *J. Hist. Ideas*, vol. 61, no. 2, pp. 285–303, 2000, doi: 10.2307/3654029.

[11] M.-T. C. Ulijaszek S, Ed., *Anthropometry: the individual and the population*. Cambridge University Press, 1994.

[12] “Historical Anthropometrics.” https://eh.net/encyclopedia/historical-anthropometrics/ (accessed Dec. 14, 2020).

[13] J. C. Kenna, “Sir Francis Galton’s Contribution to Anthropology,” *J. R. Anthropol. Inst. Gt. Britain Irel.*, vol. 94, no. 2, p. 80, Jul. 1964, doi: 10.2307/2844375.

[14] J. Aldrich, *The Origins of Modern Statistics*, vol. 1. Oxford University Press, 2017.

[15] M. R. Dietrich and R. A. Skipper, Jr, “R. A. Fisher and the Foundations of Statistical Biology,” *Dartmouth Fac. Open Access Artic.*, vol. 33, pp. 147–160, 2013.

[16] B. P. McEvoy and P. M. Visscher, “Genetics of human height,” *Econ. Hum. Biol.*, vol. 7, no. 3, pp. 294–306, Dec. 2009, doi: 10.1016/j.ehb.2009.09.005.

[17] W. King, “The Reputed Fossil Man of the Neanderthal,” *Q. J. Sci.*, vol. 1, no. 96, 1864.

[18] M. Cartmill, “Human Uniqueness and Theoretical Content in Paleoanthropology,” *Int. J. Primatol.*, vol. 11, no. 3, 1990.

[19] J. R. R. Drell, “Neanderthals: A History of Interpretation,” *Oxford J. Archaeol.*, vol. 19, no. 1, pp. 1–24, Feb.

2000, doi: 10.1111/1468-0092.00096.

[20] F. Boas, “Changes in the Bodily Form of Descendants of Immigrants,” *Am. Anthropol.*, vol. 14, no. 3, pp. 530–562, Jul. 1912, doi: 10.1525/aa.1912.14.3.02a00080.

[21] J. Komlos, “Anthropometric history: an overview of a quarter century of research,” *Anthropol. Anzeiger*, vol. 67, no. 4, pp. 341–356, Mar. 2010, doi: 10.1127/0003-5548/2009/0027.

[22] F. A. Milan, *The Human Biology of Circumpolar Populations*. Cambridge University Press, 1980.

[23] P. T. Baker, *The Biology of High Altitude Peoples*. Cambridge: Cambridge University Press, 1978.

[24] G. W. Lasker, “Human biological adaptability,” *Science (80-.).*, vol. 166, no. 3912, pp. 1480–1486, 1969, doi: 10.1126/science.166.3912.1480.

[25] “Anthropometry - Definition, History and Applications | Biology Dictionary.” https://biologydictionary.net/anthropometry/ (accessed Apr. 04, 2021).

[26] S. L. Lindstedt and W. A. Calder, “Body Size, Physiological Time, and Longevity of Homeothermic Animals,” *Q. Rev. Biol.*, vol. 56, no. 1, pp. 1–16, Mar. 1981, doi: 10.1086/412080.

[27] R. W, “The application of zoological laws to man,” *Rep Brit Assoc Adv Sci*, vol. 1, pp. 832–847., 1908.

[28] P. T. Katzmarzyk and W. R. Leonard, “Climatic influences on human body size and proportions: Ecological adaptations and secular trends,” *Am. J. Phys. Anthropol.*, vol. 106, no. 4, pp. 483–503, Aug. 1998, doi: 10.1002/(SICI)1096-8644(199808)106:4<483::AID-AJPA4>3.0.CO;2-K.

[29] C. B. Ruff, “Morphological adaptation to climate in modern and fossil hominids,” *Am. J. Phys. Anthropol.*, vol. 37, no. 19 S, pp. 65–107, Jan. 1994, doi: 10.1002/ajpa.1330370605.

[30] G. Stulp and L. Barrett, “Evolutionary perspectives on human height variation,” *Biol. Rev.*, vol. 91, no. 1, pp. 206–234, 2016, doi: 10.1111/brv.12165.

[31] C. C. Roseman and B. M. Auerbach, “Ecogeography, genetics, and the evolution of human body form,” *J. Hum. Evol.*, vol. 78, pp. 80–90, Jan. 2015, doi: 10.1016/j.jhevol.2014.07.006.

[32] K. R. R. Savell, B. M. Auerbach, and C. C. Roseman, “Constraint, natural selection, and the evolution of human body form,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, no. 34, pp. 9492–9497, Aug. 2016, doi: 10.1073/pnas.1603632113.

[33] C. B. Ruff, “Climatic adaptation and hominid evolution: The thermoregulatory imperative,” *Evol. Anthropol. Issues, News, Rev.*, vol. 2, no. 2, pp. 53–60, Jun. 2005, doi: 10.1002/evan.1360020207.

[34] C. Bergmann, “Ueber die Verhaltnisse der Warmeokonomie der Thierezuihrer Grosse,” *Gottinger Stud.*, vol. 1, pp. 595–708, 1847.

[35] A. JA, “The influence of physical conditions on the genesis of species,” *Radic. Rev*, vol. 1, pp. 108–140, 1877.

[36] E. Mayr, “Geographical Character Gradients and Climactic Adaptation,” *Evolution (N. Y).*, vol. 10, no. 1, pp. 105–110, Mar. 1956, doi: 10.1111/j.1558-5646.1956.tb02836.x.

[37] S. Meiri, “Bergmann’s Rule - what’s in a name?,” *Global Ecology and Biogeography*, vol. 20, no. 1. John Wiley & Sons, Ltd, pp. 203–207, Jan. 01, 2011, doi: 10.1111/j.1466-8238.2010.00577.x.

[38] B. Rensch, “Some Problems of Geographical Variation and Species Formation,” *Proc. Linn. Soc. London*, vol. 150, no. 4, pp. 275–285, Dec. 1938, doi: 10.1111/j.1095-8312.1938.tb00182k.x.

[39] C. Ruff, “Variation in Human Body Size and Shape,” *Annu. Rev. Anthropol.*, vol. 31, no. 1, pp. 211–232, Oct. 2002, doi: 10.1146/annurev.anthro.31.040402.085407.

[40] E. Schreider, “Ecological Rules, Body Heat Regulation, and Human Evolution,” *Evolution (N. Y).*, vol. 18, no. 1, pp. 1–9, Mar. 1964, doi: 10.1111/j.1558-5646.1964.tb01562.x.

[41] M. J. Tilkens, C. Wall-Scheffler, T. D. Weaver, and K. Steudel-Numbers, “The effects of body proportions on thermoregulation: an experimental assessment of Allen’s rule,” *J. Hum. Evol.*, vol. 53, no. 3, pp. 286–291, Sep. 2007, doi: 10.1016/j.jhevol.2007.04.005.

[42] P. T. Katzmarzyk and W. R. Leonard, “Climatic influences on human body size and proportions: Ecological adaptations and secular trends,” *Am. J. Phys. Anthropol.*, vol. 106, no. 4, pp. 483–503, Aug. 1998, doi: 10.1002/(SICI)1096-8644(199808)106:4<483::AID-AJPA4>3.0.CO;2-K.

[43] D. F. Roberts, “Body weight, race and climate,” *Am. J. Phys. Anthropol.*, vol. 11, no. 4, pp. 533–558, Dec. 1953, doi: 10.1002/ajpa.1330110404.

[44] J. R. Bindon and P. T. Baker, “Bergmann’s Rule and the Thrifty Genotype,” Wiley-Liss, Inc, 1997.

[45] R. I. Murrill, “Some Aspects of Human Racial Adaptation,” *Trans. Kansas Acad. Sci.*, vol. 63, no. 3, pp. 194–214, 1960.

[46] L. Betti, S. J. Lycett, N. von Cramon-Taubadel, and O. M. Pearson, “Are human hands and feet affected by climate? A test of Allen’s rule,” *Am. J. Phys. Anthropol.*, vol. 158, no. 1, pp. 132–140, Sep. 2015, doi: 10.1002/ajpa.22774.

[47] T. W. Holliday, “Postcranial evidence of cold adaptation in European Neandertals,” *Am. J. Phys. Anthropol.*, no. 104, pp. 245–258, 1997.

[48] T. W. Holliday and C. E. Hilton, “Body proportions of circumpolar peoples as evidenced from skeletal data: Ipiutak and Tigara (Point Hope) versus Kodiak Island Inuit,” *Am. J. Phys. Anthropol.*, vol. 142, no. 2, pp. 287–302, Jun. 2010, doi: 10.1002/ajpa.21226.

[49] B. Bogin and M. I. Varela-Silva, “Leg length, body proportion, and health: A review with a note on beauty,” *Int. J. Environ. Res. Public Health*, vol. 7, no. 3, pp. 1047–1075, 2010, doi: 10.3390/ijerph7031047.

[50] M. A. Frelat and P. Mittereocker, “Postnatal ontogeny of tibia and femur form in two human populations: A multivariate morphometric analysis,” *Am. J. Hum. Biol.*, vol. 23, no. 6, pp. 796–804, Nov. 2011, doi: 10.1002/ajhb.21217.

[51] Umme Salma Durbar, “Racial Variations in Different Skulls,” *J. Pharm. Sci. Res.*, vol. 6, no. 11, pp. 370–72, 2014, doi: 10.1111/j.1469-185x.1976.tb01063.x.

[52] B. M. Auerbach and A. D. Sylvester, “Allometry and apparent paradoxes in human limb proportions: Implications for scaling factors,” *Am. J. Phys. Anthropol.*, vol. 144, no. 3, pp. 382–391, Mar. 2011, doi: 10.1002/ajpa.21418.

[53] B. M. Auerbach, “Human skeletal variation in the New World during the Holocene: Effects of climate and subsistence across geography and time - Part I,” 2008.

[54] T. W. Holliday, “Brachial and crural indices of European Late Upper Paleolithic and Mesolithic humans,” *J. Hum. Evol.*, vol. 36, no. 5, pp. 549–566, May 1999, doi: 10.1006/jhev.1998.0289.

[55] M. R. Feldesman and R. L. Fountain, "'Race' specificity and the femur/stature ratio," *Am. J. Phys. Anthropol.*, vol. 100, no. 2, pp. 207–224, 1996, doi: 10.1002/(SICI)1096-8644(199606)100:2<207::AID-AJPA4>3.0.CO;2-U.

[56] H. A. M. Daanen and W. D. Van Marken Lichtenbelt, "Human whole body cold adaptation.," *Temp. (Austin, Tex.)*, vol. 3, no. 1, pp. 104–18, 2016, doi: 10.1080/23328940.2015.1135688.

[57] M. A. Serrat, D. King, and C. O. Lovejoy, "Temperature regulates limb length in homeotherms by directly modulating cartilage growth," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 49, pp. 19348–19353, 2008, doi: 10.1073/pnas.0803319105.

[58] M. A. E. R. Ashoub, "Effect of two extreme temperatures on growth and tail-length of mice [37]," *Nature*, vol. 181, no. 4604. Nature Publishing Group, p. 284, 1958, doi: 10.1038/181284a0.

[59] M. E. Weaver and D. L. Ingram, "Morphological Changes in Swine Associated with Environmental Temperature," *Ecology*, vol. 50, no. 4, pp. 710–713, Jul. 1969, doi: 10.2307/1936264.

[60] D. C. Geary, "Sexual selection and human life history," *Adv Child Dev Behav*, vol. 30, pp. 41–101, 2002.

[61] H. G. Beyer, "The influence of exercise on growth.," *J. Exp. Med.*, vol. 1, no. 3, pp. 546–558, Jul. 1896, doi: 10.1084/jem.1.3.546.

[62] M. A. Serrat, R. M. Williams, and C. E. Farnum, "Exercise mitigates the stunting effect of cold temperature on limb elongation in mice by increasing solute delivery to the growth plate," *J. Appl. Physiol.*, vol. 109, no. 6, pp. 1869–1879, Dec. 2010, doi: 10.1152/japplphysiol.01022.2010.

[63] H. Brodin, "Longitudinal bone growth, the nutrition of the epiphyseal cartilages and the local blood supply; an experimental study in the rabbit," *Acta Orthop Scand Suppl.*, vol. 20, pp. 1–92, 1955.

[64] B. Bogin, P. Smith, A. B. Orden, M. I. V. Silva, and J. Loucky, "Rapid change in height and body proportions of Maya American children," *Am. J. Hum. Biol.*, vol. 14, no. 6, pp. 753–761, Nov. 2002, doi: 10.1002/ajhb.10092.

[65] S. V. Subramanian, E. Özaltin, and J. E. Finlay, "Height of Nations: A Socioeconomic Analysis of Cohort Differences and Patterns among Women in 54 Low- to Middle-Income Countries," *PLoS One*, vol. 6, no. 4, p. e18962, Apr. 2011, doi: 10.1371/journal.pone.0018962.

[66] T. J. Cole, "Secular trends in growth," in *Proceedings of the Nutrition Society*, 2000, vol. 59, no. 2, pp. 317–324, doi: 10.1017/S0029665100000355.

[67] N. C. D. Risk and F. Collaboration, "A century of trends in adult human height," *Elife*, vol. 5, pp. 1–29, 2016, doi: 10.7554/eLife.13410.

[68] D. F. Roberts and D. R. Bainbridge, "Nilotic physique," *Am. J. Phys. Anthropol.*, vol. 21, no. 3, pp. 341–370, Sep. 1963, doi: 10.1002/ajpa.1330210309.

[69] P. Grasgruber *et al.*, "The mountains of giants: an anthropometric survey of male youths in Bosnia and Herzegovina," *R. Soc. Open Sci.*, vol. 4, no. 4, p. 161054, Apr. 2017, doi: 10.1098/rsos.161054.

[70] J. Hiernaux, "The People of Africa," in *The People of Africa*, Wiedenfeld and Nicolson, 1974, p. 115.

[71] W. H. Dietz, B. Marino, N. R. Peacock, and R. C. Bailey, "Nutritional status of Efe pygmies and Lese horticulturists," *Am. J. Phys. Anthropol.*, vol. 78, no. 4, pp. 509–518, 1989, doi: 10.1002/ajpa.1330780406.

[72] B. Bogin, C. Scheffler, and M. Hermanussen, “Global effects of income and income inequality on adult height and sexual dimorphism in height,” *Am. J. Hum. Biol.*, vol. 29, no. 2, p. e22980, Mar. 2017, doi: 10.1002/ajhb.22980.

[73] G. Davey Smith, R. Greenwood, D. Gunnell, P. Sweetnam, J. Yarnell, and P. Elwood, “Leg length, insulin resistance, and coronary heart disease risk: The Caerphilly study,” *J. Epidemiol. Community Health*, vol. 55, no. 12, pp. 867–872, 2001, doi: 10.1136/jech.55.12.867.

[74] L. Zuccolo *et al.*, “Height and prostate cancer risk: A large nested case-control study (ProtecT) and meta-analysis,” *Cancer Epidemiol. Biomarkers Prev.*, vol. 17, no. 9, pp. 2325–2336, Sep. 2008, doi: 10.1158/1055-9965.EPI-08-0342.

[75] M. Zakaria Okasha *et al.*, “Height, Leg Length, and Cancer Risk: A Systematic Review Fats and oils technology View project Does milk intake promote prostate cancer initiation or progression via effects on insulin-like growth factors (IGFs)? View project Height, Leg Length, and Cancer Risk: A Systematic Review,” vol. 23, no. 2, 2001, doi: 10.1093/oxfordjournals.epirev.a000809.

[76] G. David Batty *et al.*, “Height, wealth, and health: An overview with new data from three longitudinal studies,” *Econ. Hum. Biol.*, vol. 7, pp. 137–152, 2009, doi: 10.1016/j.ehb.2009.06.004.

[77] D. A. Lawlor, M. Taylor, G. Davey Smith, D. Gunnell, and S. Ebrahim, “Associations of components of adult height with coronary heart disease in postmenopausal women: The British women’s heart and health study,” *Heart*, vol. 90, no. 7, pp. 745–749, 2004, doi: 10.1136/hrt.2003.019950.

[78] J. E. Ferrie, C. Langenberg, M. J. Shipley, and M. G. Marmot, “Birth weight, components of height and coronary heart disease: Evidence from the Whitehall II study,” *Int. J. Epidemiol.*, vol. 35, no. 6, pp. 1532–1542, Dec. 2006, doi: 10.1093/ije/dyl184.

[79] N. T. Mueller *et al.*, “Relative leg length is associated with type 2 diabetes differently according to pubertal timing: The Brazilian longitudinal study of adult health,” *Am. J. Hum. Biol.*, vol. 27, no. 2, pp. 219–225, 2015, doi: 10.1002/ajhb.22641.

[80] C. Langenberg, R. Hardy, D. Kuh, and M. E. J. Wadsworth, “Influence of height, leg and trunk length on pulse pressure, systolic and diastolic blood pressure,” *J. Hypertens.*, vol. 21, no. 3, pp. 537–543, Mar. 2003, doi: 10.1097/00004872-200303000-00019.

[81] A. Jelenkovic *et al.*, “Genetic and environmental influences on height from infancy to early adulthood: An individual-based pooled analysis of 45 twin cohorts,” *Sci. Rep.*, vol. 6, no. 1, p. 28496, Sep. 2016, doi: 10.1038/srep28496.

[82] P. M. Visscher, “Sizing up human height variation,” *Nat. Genet. News Views*, vol. 40, no. 5, pp. 489–490, 2008, doi: 10.1038/ng0508-489.

[83] S. Stinson, “Growth Variation: Biological and Cultural Factors,” in *Human Biology*, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2012, pp. 587–635.

[84] M. W. Warren, “Prenatal long bone growth and proportionality in fetal abortus and stillborns,” *Am. J. Phys. Anthropol. (Suppl*, vol. 26, p. 226, Oct. 1998.

[85] K. Silventoinen *et al.*, “Heritability of Adult Body Height: A Comparative Study of Twin Cohorts in Eight

Countries," *Twin Research*, vol. 6, no. 5. Twin Res, pp. 399–408, Oct. 2003, doi: 10.1375/136905203770326402.

[86] R.A Fisher, "The Correlation between Relatives on the Supposition of Mendelian Inheritance," *Trans. R. Soc. Edinburgh*, vol. 52, no. 2, p. pp 399-433, 1918.

[87] S. Benonisdottir *et al.*, "Epigenetic and genetic components of height regulation," *Nat. Commun.*, vol. 7, no. 1, p. 13490, Dec. 2016, doi: 10.1038/ncomms13490.

[88] A. R. et al Wood, "Defining the role of common variation in the genomic and biological architecture of adult human height," *Nat. Genet.*, vol. 46, no. 11, pp. 1173–1186, 2015, doi: 10.1038/ng.3097.

[89] E. Marouli, "Rare and low-frequency coding variants alter human adult height," *Nature*, vol. 542, no. 7640, p. 186, 2017.

[90] J. Yang *et al.*, "Common SNPs explain a large proportion of the heritability for human height," *Nat. Genet.*, vol. 42, no. 7, pp. 565–569, 2010, doi: 10.1038/ng.608.

[91] J. Yang *et al.*, "Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index," *Nat. Genet.*, vol. 47, no. 10, pp. 1114–1120, 2015, doi: 10.1038/ng.3390.

[92] P. Simeone and S. Alberti, "Epigenetic heredity of human height.," *Physiol. Rep.*, vol. 2, no. 6, Jun. 2014, doi: 10.14814/phy2.12047.

[93] E. Marouli *et al.*, "Rare and low-frequency coding variants alter human adult height," *Nature*, vol. 542, no. 7640, pp. 186–190, Feb. 2017, doi: 10.1038/nature21039.

[94] M. Cavalli-Sforza, LL, Hewlett, B, Bahuchet, S, Verdu, P, Lupo, K, Ndanga, J, Kiahtipes, C, Moise, R, Froment, A, Sato, Hiroaki, Furniss, S, Lewis, J, Rupp, S, Takeuchi, K, Ichikawa, *Hunter-Gatherers of the Congo Basin: Cultural Histories and Biology of African Pygmies*. Transaction Publishers, New Brunswick, New Jersey, 2014.

[95] D. V Joiris, "The Framework of Central African Hunter-Gatherers and Neighbouring Societies," *African Study Monogr. Suppl. Issue*, vol. Suppl.28, no. 28, pp. 57–79, 2003.

[96] G. H. Perry and N. J. Dominy, "Evolution of the human pygmy phenotype," *Trends in Ecology and Evolution*, vol. 24, no. 4. pp. 218–225, Apr. 2009, doi: 10.1016/j.tree.2008.11.008.

[97] R. C. Bailey, G. Head, M. Jenike, B. Owen, R. Rechtman, and E. Zechenter, "Hunting and Gathering in Tropical Rain Forest: Is It Possible?," *Am. Anthropol.*, vol. 91, no. 1, pp. 59–82, Mar. 1989, doi: 10.1525/aa.1989.91.1.02a00040.

[98] J. M. Diamond, "Why are pygmies small?," *Nature*, vol. 354, no. 6349. Nature Publishing Group, pp. 111–112, 1991, doi: 10.1038/354111a0.

[99] A. E. Minetti, F. Saibene, L. P. Ardigò, G. Atchou, F. Schena, and G. Ferretti, "Pygmy locomotion," *Eur. J. Appl. Physiol. Occup. Physiol.*, vol. 68, no. 4, pp. 285–290, Jul. 1994, doi: 10.1007/BF00571444.

[100] L. Cavalli-Sforza, Ed., *African Pygmies*. Orlando Florida: Academic Press Inc, 1986.

[101] N. E. Ghesquiere JL, "Implications of small body size to strength and endurance of Pygmy and village children in the Ituri Forest," in *4th World conference on human ecology*, 1993.

[102] A. B. Migliano, L. Vinicius, and M. M. Lahr, "Life history trade-offs explain the evolution of human

pygmies," *Proc. Natl. Acad. Sci.*, vol. 104, no. 51, pp. 20216–20219, Dec. 2007, doi: 10.1073/pnas.0708024105.

[103] L. Quintana-Murci *et al.*, "Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 5, pp. 1596–1601, 2008, doi: 10.1073/pnas.0711467105.

[104] J. L. Newman, *The peopling of Africa : a geographic interpretation*. 1995.

[105] G. Berniell-Lee *et al.*, "Genetic and Demographic Implications of the Bantu Expansion: Insights from Human Paternal Lineages," *Mol. Biol. Evol.*, vol. 26, no. 7, pp. 1581–1589, Jul. 2009, doi: 10.1093/molbev/msp069.

[106] J. Diamond and P. Bellwood, "Farmers and their languages: The first expansions," *Science*, vol. 300, no. 5619. pp. 597–603, Apr. 25, 2003, doi: 10.1126/science.1078208.

[107] K. Bostoen, "The Bantu Expansion," in *Oxford Research Encyclopedia of African History*, Oxford University Press, 2018.

[108] D. J. Rowold, D. Perez-Benedico, O. Stojkovic, R. Garcia-Bertrand, and R. J. Herrera, "On the Bantu expansion," *Gene*, vol. 593, no. 1, pp. 48–57, 2016, doi: 10.1016/j.gene.2016.07.044.

[109] J. Vansina, "New linguistic evidence and 'The Bantu Expansion,'" *J. Afr. Hist.*, vol. 36, no. 2, pp. 173–195, 1995, doi: 10.1017/S0021853700034101.

[110] E. Patin *et al.*, "The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists.," *Nat. Commun.*, vol. 5, p. 3163, 2014, doi: 10.1038/ncomms4163.

[111] J. Knight, "Relocated to the roadside: Preliminary observations on the Forest Peoples of Gabon," *Afr. Study Monogr.*, vol. Suppl. 28, no. November, pp. 81–121, 2003.

[112] S. Bahuchet, "Changing Language, Remaining Pygmy," *Hum. Biol.*, vol. 84, no. 1, pp. 11–43, Mar. 2012, doi: 10.3378/027.084.0101.

[113] P. Verdu *et al.*, "Origins and Genetic Diversity of Pygmy Hunter-Gatherers from Western Central Africa," *Curr. Biol.*, vol. 19, no. 4, pp. 312–318, Feb. 2009, doi: 10.1016/j.cub.2008.12.049.

[114] M. Robillard and S. Bahuchet, "Les Pygmées et les autres : terminologie, catégorisation et politique," *http://journals.openedition.org/africanistes*, no. 82–1/2, pp. 15–51, Jun. 2012, doi: 10.4000/AFRICANISTES.4253.

[115] A. Carstairs-McCarthy and M. Ruhlen, "The Origin of Language: Tracing the Evolution of the Mother Tongue," *Language (Baltim).*, vol. 73, no. 3, p. 611, 1997, doi: 10.2307/415896.

[116] C. de Filippo, K. Bostoen, M. Stoneking, and B. Pakendorf, "Bringing together linguistic and genetic evidence to test the Bantu expansion," *Proc. R. Soc. B Biol. Sci.*, vol. 279, no. 1741, pp. 3256–3263, 2012, doi: 10.1098/rspb.2012.0318.

[117] F. V. R. Rozzi, Y. Koudou, A. Froment, Y. Le Bouc, and J. Botton, "Growth pattern from birth to adulthood in African pygmies of known age," *Nat. Commun.*, vol. 6, Jul. 2015, doi: 10.1038/ncomms8672.

[118] C. Batini *et al.*, "Insights into the Demographic History of African Pygmies from Complete Mitochondrial Genomes," *Mol. Biol. Evol.*, vol. 28, no. 2, pp. 1099–1110, Feb. 2011, doi: 10.1093/molbev/msq294.

[119] E. Patin *et al.*, "Inferring the Demographic History of African Farmers and Pygmy Hunter–Gatherers Using a

Multilocus Resequencing Data Set," *PLoS Genet.*, vol. 5, no. 4, p. e1000448, Apr. 2009, doi: 10.1371/journal.pgen.1000448.

[120] J. P. Jarvis *et al.*, "Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies," *PLoS Genet.*, vol. 8, no. 4, 2012, doi: 10.1371/journal.pgen.1002641.

[121] S. Hattori, "Current Issues Facing the Forest People in South-Eastern Cameroon: The dynamics of Baka life and their ethnic relationship with farmers," *Afr. Study Monogr.*, Feb. 2014.

[122] J. Vincent, M, Jans, C, Ghesquiere, "The newborn Pygmy and his mother," *Am. J. Physial Anthropol.*, vol. 20, pp. 237–47, 1962.

[123] D. Lemogoum *et al.*, "Effects of hunter-gatherer subsistence mode on arterial distensibility in cameroonian pygmies," *Hypertension*, vol. 60, no. 1, pp. 123–128, 2012, doi: 10.1161/HYPERTENSIONAHA.111.187757.

[124] B. T. Shea and R. C. Bailey, "Allometry and adaptation of body proportions and stature in African pygmies," *Am. J. Phys. Anthropol.*, vol. 100, no. 3, pp. 311–340, 1996, doi: 10.1002/(SICI)1096-8644(199607)100:3<311::AID-AJPA2>3.0.CO;2-X.

[125] J. Hiernaux, "Long-term biological effects of human migration from the Africa savana to the equitorial forest: a case study of human adaptation to a hot and wet climate," in *Population Structure and Human variation*, G. . Harrison, Ed. Cambridge University Press, 1977, pp. 203–204.

[126] N. W. Blackstone, "Allometry and Relative Growth: Pattern and Process in Evolutionary Studies," *Syst. Zool.*, vol. 36, no. 1, p. 76, Mar. 1987, doi: 10.2307/2413310.

[127] A. Shingleton, "Allometry: The Study of Biological Scaling," *Nat. Educ. Knowl.*, vol. 3, no. 10, p. 2, 2010.

[128] S. L. Lindstedt, "Allometry: Body Size Constraints in Animal Design," in *Drinking Water and Health, Volume 8: Pharmacokinetics in Risk Assessment*, National Research Council (US) Subcommittee on Pharmacokinetics in Risk Assessment, Ed. Washington (DC): National Academies Press (US), 1987.

[129] J. S. Huxley, *Problems of Relative Growth*. Dover Publishing Inc, 1933.

[130] P. Travaglino, C. Meazza, S. Pagani, G. Biddeci, and M. Bozzola, "Secular trends in growth of African pygmies and Bantu," *Hormones*, vol. 10, no. 2, pp. 144–148, 2011, doi: 10.14310/horm.2002.1304.

[131] I. Hagino, K. Hayashi, K. Kawamura, H. Sato, and T. Yamauchi, "Adolescent growth spurt and growth pattern factors related to the short stature of Pygmy hunter-gatherers of Southeast Cameroon," *Ann. Hum. Biol.*, vol. 40, no. 1, pp. 9–14, Jan. 2013, doi: 10.3109/03014460.2012.720711.

[132] F. Ramirez Rozzi, "Diversity in tooth eruption and life history in humans: illustration from a Pygmy population," *Sci. Rep.*, vol. 6, no. 1, p. 27405, 2016, doi: 10.1038/srep27405.

[133] A. Romero, F. V. Ramirez-Rozzi, and A. Pérez-Pérez, "Dental size variability in Central African Pygmy hunter-gatherers and Bantu-speaking farmers," *Am. J. Phys. Anthropol.*, vol. 166, no. 3, pp. 671–681, Jul. 2018, doi: 10.1002/ajpa.23458.

[134] B. T. Shea and A. M. Gomez, "Tooth scaling and evolutionary dwarfism: An investigation of allometry in human pygmies," *Am. J. Phys. Anthropol.*, vol. 77, no. 1, pp. 117–132, Sep. 1988, doi: 10.1002/ajpa.1330770117.

[135] H. Kesteloot, N. Ndam, S. Sasaki, M. Kowo, and V. Seghers, “A survey of blood pressure distribution in Pygmy and Bantu populations in Cameroon,” *Hypertension*, vol. 27, no. 1, pp. 108–113, 1996, doi: 10.1161/01.HYP.27.1.108.

[136] E. S. Eaton SB, “Hunter-gatherers and human health,” in *The Cambridge Encyclopedia of Hunters and Gatherers*, 1st ed., R. D. Richard B. Lee, Richard Heywood Daly, Ed. Cambridge, UK: Cambridge University Press, 1999, pp. 449–455.

[137] W. Ngatchou *et al.*, “Arterial stiffness and cardiometabolic phenotype of Cameroonian Pygmies and Bantus,” *J. Hypertens.*, p. 1, 2017, doi: 10.1097/HJH.0000000000001577.

[138] W. Ngatchou *et al.*, “Arterial stiffness and cardiometabolic phenotype of Cameroonian Pygmies and Bantus,” *J. Hypertens.*, vol. 36, no. 3, pp. 520–527, 2018, doi: 10.1097/HJH.0000000000001577.

[139] G. . Mann, O. . Roels, D. . Price, and J. . Merrill, “Cardiovascular Disease in African Pygmies: A suvey of the health status, serum lipids and diet of Pygmies in Congo,” *J. chron. Dis.*, vol. 15, pp. 341–371, 1961.

[140] L. D. *et al.*, “Association of urinary sodium excretion with blood pressure and risk factors associated with hypertension among Cameroonian pygmies and bantus: A cross-sectional study,” *BMC Cardiovasc. Disord.*, vol. 18, no. 1, pp. 1–10, 2018, doi: 10.1186/s12872-018-0787-3.

[141] H. S. Kawamura K, Yamauchi T, Hayashi K, “Blood Pressure of Baka Pygmies living in Southeastern Cameroon,” *Afr. Study Monogr.*, vol. 47, no. March, pp. 35–44, 2014.

[142] G. V. Mann, O. A. Roels, D. L. Price, and J. M. Merrill, “Cardiovascular disease in African Pygmies. A survey of the health status, serum lipids and diet of Pygmies in Congo,” *J. Chronic Dis.*, vol. 15, no. 4, pp. 341–371, Apr. 1962, doi: 10.1016/0021-9681(62)90082-6.

[143] A. Froment, “Human biology and the health of African rainforest inhabitants,” pp. 117–164, 2014.

[144] C. Cobbaert, P. Mulder, J. Lindemans, and H. Kesteloot, “Serum LP(a) levels in African Aboriginal Pygmies and Bantus, compared with Caucasian and Asian population samples,” *J. Clin. Epidemiol.*, vol. 50, no. 9, pp. 1045–1053, 1997, doi: 10.1016/S0895-4356(97)00129-7.

[145] E. Dounias and A. Froment, “When forest-based hunter-gatherers become sedentary: Consequences for diet and health,” *Unasylva*, vol. 57, no. 224, pp. 26–33, 2006.

[146] T. J. Merimee, D. L. Rimoin, and L. L. Cavalli-Sforza, “Metabolic Studies in the African Pygmy,” *J. Clin. Invest.*, vol. 51, pp. 395–401, 1972.

[147] T. J. Merimee, J. Zapf, and E. R. Froesch, “Dwarfism in the Pygmy,” *N. Engl. J. Med.*, vol. 305, no. 17, pp. 965–968, Oct. 1981, doi: 10.1056/NEJM198110223051701.

[148] D. L. Rimoin, T. J. Merimee, D. Rabinowitz, L. L. Cavalli-Sforza, and V. A. McKusick, “Peripheral Subresponsiveness to Human Growth Hormone in the African Pygmies,” *N. Engl. J. Med.*, vol. 281, no. 25, pp. 1383–1388, Dec. 1969, doi: 10.1056/nejm196912182812502.

[149] M. E. Geffner, N. Bersch, R. C. Bailey, and D. W. Golde, “Insulin-like growth factor I resistance in immortalized T cell lines from African Efe Pygmies.,” *J. Clin. Endocrinol. Metab.*, vol. 80, no. 12, pp. 3732–3738, Dec. 1995, doi: 10.1210/jcem.80.12.8530626.

[150] Y. Hattori *et al.*, “Decreased insulin-like growth factor I receptor expression and function in immortalized

African Pygmy T cells," *J. Clin. Endocrinol. Metab.*, vol. 81, no. 6, pp. 2257–2263, 1996, doi: 10.1210/jc.81.6.2257.

[151] T. J. Merimee, B. S. Hewlett, W. Wood, A. M. Bowcock, and L. L. Cavalli-Sforza, "The growth hormone receptor gene in the African pygmy.," *Trans. Assoc. Am. Physicians*, vol. 102, pp. 163–169, 1989.

[152] M. Bozzola *et al.*, "The shortness of Pygmies is associated with severe under-expression of the growth hormone receptor," *Mol. Genet. Metab.*, vol. 98, no. 3, pp. 310–313, Nov. 2009, doi: 10.1016/j.ymgme.2009.05.009.

[153] N. S. A. Becker *et al.*, "The role of GHR and IGF1 genes in the genetic determination of African pygmies' short stature," *Eur. J. Hum. Genet.*, vol. 21, no. 6, pp. 653–658, 2013, doi: 10.1038/ejhg.2012.223.

[154] P. Hsieh *et al.*, "Whole-genome sequence analyses of Western Central African Pygmy hunter-gatherers reveal a complex demographic history and identify candidate genes under positive natural selection," 2016, doi: 10.1101/gr.192971.115.

[155] J. E. Tierney, P. B. deMenocal, and P. D. Zander, "A climatic context for the out-of-Africa migration," *Geology*, vol. 45, no. 11, pp. 1023–1026, Nov. 2017, doi: 10.1130/G39457.1.

[156] P. Verdu *et al.*, "Sociocultural behavior, sex-biased admixture, and effective population sizes in central African pygmies and non-pygmies," *Mol. Biol. Evol.*, vol. 30, no. 4, pp. 918–937, 2013, doi: 10.1093/molbev/mss328.

[157] G. Destro-Bisol *et al.*, "Variation of female and male lineages in sub-Saharan populations: The importance of sociocultural factors," *Mol. Biol. Evol.*, vol. 21, no. 9, pp. 1673–1682, Sep. 2004, doi: 10.1093/molbev/msh186.

[158] N. A. Rosenberg *et al.*, "Genetic structure of human populations," *Science (80-.).*, vol. 298, no. 5602, pp. 2381–2385, Dec. 2002, doi: 10.1126/science.1078311.

[159] R. Rozzi and F. V Sardi, "Diversity among African Pygmies," *Divers. among African Pygmies. PLoS ONE*, vol. 5, no. 10, p. 13620, 2010, doi: 10.1371/journal.pone.0013620.

[160] E. Patin and L. Quintana-Murci, "The demographic and adaptive history of central African hunter-gatherers and farmers," *Current Opinion in Genetics and Development*, vol. 53. Elsevier Ltd, pp. 90–97, Dec. 01, 2018, doi: 10.1016/j.gde.2018.07.008.

[161] T. J. Pemberton *et al.*, "A genome scan for genes underlying adult body size differences between Central African hunter-gatherers and farmers," *Hum. Genet.*, vol. 137, no. 6–7, pp. 487–509, Jul. 2018, doi: 10.1007/s00439-018-1902-3.

[162] E. Patin *et al.*, "The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists," *Nat. Commun.*, vol. 5, 2014, doi: 10.1038/ncomms4163.

[163] G. H. Perry *et al.*, "Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers," *Proc. Natl. Acad. Sci.*, vol. 111, no. 35, pp. E3596–E3603, 2014, doi: 10.1073/pnas.1402875111.

[164] E. Patin *et al.*, "Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America," *Science (80-.).*, vol. 356, no. 6337, pp. 543–546, May 2017, doi: 10.1126/science.aal1988.

[165] S. Nell *et al.*, "Recent Acquisition of Helicobacter pylori by Baka Pygmies," *PLoS Genet.*, vol. 9, no. 9, Sep.

2013, doi: 10.1371/journal.pgen.1003775.

[166] J. K. Pickrell *et al.*, "Signals of recent positive selection in a worldwide sample of human populations," *Genome Res.*, vol. 19, no. 5, pp. 826–837, May 2009, doi: 10.1101/gr.087577.108.

[167] A. Bowcock and V. Sartorelli, "Polymorphism and mapping of the IGF1 gene, and absence of association with stature among African Pygmies," *Hum. Genet.*, vol. 85, no. 3, pp. 349–354, Aug. 1990, doi: 10.1007/BF00206760.

[168] J. Hollinger, I, Pennings PS, Hermisson, "Polygenic adaptation: From sweeps to subtle frequency shifts," *POS Genet.*, vol. 15, no. 3, 2019, doi: 10.1371/journal.pgen.1008035.

[169] T. G. Lohman, A. F. Roche, and R. Martorell, *Anthropometric Standardization Reference Manual*. Champaign, IL: Human Kinetics Books, 1991.

[170] L. Cavalli-Sforza, "Demographic data," in *African Pygmies*, Cavalli-Sforza LL (Stanford University), Ed. Academic Press Inc, 1986, pp. 34–35.

[171] F. V. Ramirez Rozzi, "Reproduction in the Baka pygmies and drop in their fertility with the arrival of alcohol," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 115, no. 27, pp. E6126–E6134, Jul. 2018, doi: 10.1073/pnas.1719637115.

[172] X. A. Harrison *et al.*, "A brief introduction to mixed effects modelling and multi-model inference in ecology," *PeerJ*, vol. 2018, no. 5, 2018, doi: 10.7717/peerj.4794.

[173] S. Nakagawa and H. Schielzeth, "A general and simple method for obtaining R^2 from generalized linear mixed-effects models," *Methods Ecol. Evol.*, vol. 4, no. 2, pp. 133–142, Feb. 2013, doi: 10.1111/j.2041-210x.2012.00261.x.

[174] "CRAN - Package MuMIn." https://cran.r-project.org/web/packages/MuMIn/index.html (accessed Apr. 03, 2021).

[175] N. D. Phillips, "Ch 15: Comparing regression models with anova()," *APS Observer*, 2017. https://bookdown.org/ndphillips/YaRrr/comparing-regression-models-with-anova.html (accessed May 04, 2020).

[176] H. K. Burnham K, Anderson D, "AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons," *Behav Ecol Sociobiol*, vol. 65, pp. 23–35, 2011.

[177] Hallquist Michael, "Model comparison and evidence," 2017. https://psu-psychology.github.io/psy-597-SEM/09_model_comparison/model_comparison.html#akaike-information-criterion-aic (accessed Apr. 03, 2021).

[178] B. OMeara, "Model Selection Using the Akaike Information Criterion (AIC)." http://brianomeara.info/aic.html (accessed Apr. 03, 2021).

[179] S. I. Vrieze, "Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC)," *Psychol. Methods*, vol. 17, no. 2, pp. 228–243, Jun. 2012, doi: 10.1037/a0027127.

[180] D. Posada and T. R. Buckley, "Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests," *Syst. Biol*, vol. 53, no. 5, pp. 793–808, 2004, doi: 10.1080/10635150490522304.

[181] L. J. Brandon, R. M. Mullis, S. S. Jonnalagadda, and M. H. Hughes, “Relationships and CHD risks of BMI, lipoproteins, lipids, and blood pressure in African-American men and women,” *Prev. Med. (Baltim).*, vol. 40, no. 3, pp. 349–354, Mar. 2005, doi: 10.1016/j.ypmed.2004.06.013.

[182] M. Guardiola *et al.*, “Body mass index correlates with atherogenic lipoprotein profile even in nonobese, normoglycemic, and normolipidemic healthy men,” *J. Clin. Lipidol.*, vol. 9, no. 6, pp. 824-831.e1, Nov. 2015, doi: 10.1016/j.jacl.2015.08.001.

[183] J. Kwagyan, C. E. Tabe, S. Xu, A. R. Maqbool, V. R. Gordeuk, and O. S. Randall, “The impact of body mass index on pulse pressure in obesity,” *J. Hypertens.*, vol. 23, no. 3, pp. 619–624, 2005.

[184] H. J. Yoo, M. Kim, S.-H. Lee, and J. H. Lee, “Elevated Lipoprotein-Associated Phospholipase A2 Independently Affects Age-Related Increases in Systolic Blood Pressure: A Nested Case-Control Study in a Prospective Korean Cohort,” *Int. J. Hypertens.*, vol. 2020, 2020, doi: 10.1155/2020/5693271.

[185] S. Vaid, L. Hanks, R. Griffin, and A. P. Ashraf, “Body mass index and glycemic control influence lipoproteins in children with type 1 diabetes,” *J. Clin. Lipidol.*, vol. 10, no. 5, pp. 1240–1247, Sep. 2016, doi: 10.1016/j.jacl.2016.07.010.

[186] P. J. Brennan, J. M. Simpson, R. B. Blacket, and C. A. McGilchrist, “The Effects of Body Weight on Serum Cholesterol, Serum Triglycerides, Serum Urate and Systolic Blood Pressure*,” *Aust. N. Z. J. Med.*, vol. 10, no. 1, pp. 15–20, Feb. 1980, doi: 10.1111/j.1445-5994.1980.tb03412.x.

[187] B. W. Sume, “Estimation of body height from percutaneous length of tibia in Debre Markos University students, North West Ethiopia,” *Egypt. J. Forensic Sci.*, vol. 9, no. 1, pp. 1–8, Dec. 2019, doi: 10.1186/s41935-019-0157-z.

[188] A. K. et al. Prasad, “Stature Estimation from Head Length and Breadth by Regression Analysis in Madhya Pradesh Population,” *Int. J. Anat. Radiol. Surg.*, vol. 08, no. 3, pp. AO22–AO24, Jul. 2019, doi: 10.7860/ijars/2019/41606:2491.

[189] A. Özaslan, M. Y. Işcan, I. Özaslan, H. Tuğcu, and S. Koç, “Estimation of stature from body parts,” *Forensic Sci. Int.*, vol. 132, no. 1, pp. 40–45, Mar. 2003, doi: 10.1016/S0379-0738(02)00425-5.

[190] C. W. Dupertuis and J. A. Hadden, “On the reconstruction of stature from long bones,” *Am. J. Phys. Anthropol.*, vol. 9, no. 1, pp. 15–54, 1951, doi: 10.1002/ajpa.1330090104.

[191] J. Beddoe, “On the Stature of the Older Races of England, as Estimated from the Long Bones,” 1888.

[192] “High cholesterol - Diagnosis and treatment - Mayo Clinic.” https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/diagnosis-treatment/drc-20350806 (accessed Apr. 06, 2021).

[193] Nhlbi, “High Blood Cholesterol What you need to know.”

[194] “Cholesterol: Types, Tests, Treatments, Prevention.” https://my.clevelandclinic.org/health/articles/11920-cholesterol-numbers-what-do-they-mean (accessed Apr. 06, 2021).

[195] E. Hernandez-Vila, “A Review of the JNC 8 Blood Pressure Guideline,” *Texas Hear. Inst. J.*, vol. 42, no. 3, pp. 226–228, 2015, doi: 10.14503/THIJ-15-5067.

[196] A. Van De Wiel, “The Effect of Alcohol on Postprandial and Fasting Triglycerides,” *Int. J. Vasc. Med.*, vol. 2012, 2012, doi: 10.1155/2012/862504.

[197] "Triglycerides: Why do they matter? - Mayo Clinic." https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/in-depth/triglycerides/art-20048186 (accessed Apr. 06, 2021).

[198] N. Rifai *et al.*, "Nonfasting Sample for the Determination of Routine Lipid Profile: Is It an Idea Whose Time Has Come?," 2015.

[199] Centers for Disease Control and Prevention, "Diabetes Tests ." https://www.cdc.gov/diabetes/basics/getting-tested.html (accessed Apr. 06, 2021).

[200] S. Bahuchet, "Changing language, remaining pygmy.," *Hum. Biol.*, vol. 84, no. 1, pp. 11–43, Feb. 2012, doi: 10.3378/027.084.0101.

[201] M. F. Steele and T. C. Chenier, "Arm-span, height, and age in black and white women," *Ann. Hum. Biol.*, vol. 17, no. 6, pp. 533–541, 1990, doi: 10.1080/03014469000001312.

[202] A. Pietak, S. Ma, C. W. Beck, and M. D. Stringer, "Fundamental ratios and logarithmic periodicity in human limb bones," *J. Anat.*, vol. 222, no. 5, pp. 526–537, May 2013, doi: 10.1111/joa.12041.

[203] J. W. Littler, "On the adaptability of man's hand. With reference to the equiangular curve," *Hand*, vol. 5, no. 3, pp. 187–191, Oct. 1973, doi: 10.1016/0072-968X(73)90027-2.

[204] A. Roselló-Díez and A. L. Joyner, "Regulation of long bone growth in vertebrates; It is time to catch up," *Endocrine Reviews*, vol. 36, no. 6. Endocrine Society, pp. 646–680, 2015, doi: 10.1210/er.2015-1048.

[205] E. Otto *et al.*, "Crosstalk of brain and bone—clinical observations and their molecular bases," *International Journal of Molecular Sciences*, vol. 21, no. 14. MDPI AG, pp. 1–40, Jul. 02, 2020, doi: 10.3390/ijms21144946.

[206] I. Mosialou *et al.*, "MC4R-dependent suppression of appetite by bone-derived lipocalin 2," *Nature*, vol. 543, no. 7645, pp. 385–390, Mar. 2017, doi: 10.1038/nature21697.

[207] "High HDL levels: Recommendations, balance, and tips." https://www.medicalnewstoday.com/articles/319275#finding-the-right-balance (accessed Apr. 17, 2021).

[208] "HDL Cholesterol - Understand the Test & Your Results." https://labtestsonline.org/tests/hdl-cholesterol (accessed Apr. 17, 2021).

[209] C. M. Madsen, A. Varbo, and B. G. Nordestgaard, "Extreme high high-density lipoprotein cholesterol is paradoxically associated with high mortality inmen and women: Two prospective cohort studies," *Eur. Heart J.*, vol. 38, no. 32, pp. 2478–2486, Aug. 2017, doi: 10.1093/eurheartj/ehx163.

[210] S. K. Karathanasis, L. A. Freeman, S. M. Gordon, and A. T. Remaley, "The Changing Face of HDL and the Best Way to Measure It," 2016, doi: 10.1373/clinchem.2016.257725.

[211] D. T. Ko *et al.*, "High-Density Lipoprotein Cholesterol and Cause-Specific Mortality in Individuals Without Previous Cardiovascular Conditions: The CANHEART Study," *J. Am. Coll. Cardiol.*, vol. 68, no. 19, pp. 2073–2083, Nov. 2016, doi: 10.1016/j.jacc.2016.08.038.

[212] E. Bruckert, "Epidemiology of low HDL-cholesterol: results of studies and surveys," doi: 10.1093/eurheartj/sul036.

[213] D. Lemogoum *et al.*, "Effects of hunter-gatherer subsistence mode on arterial distensibility in cameroonian pygmies," *Hypertension*, vol. 60, no. 1, pp. 123–128, Jul. 2012, doi:

10.1161/HYPERTENSIONAHA.111.187757.

[214] J. J. Noubiap *et al.*, "Prevalence of dyslipidaemia among adults in Africa: a systematic review and meta-analysis," *Lancet Glob. Heal.*, vol. 6, no. 9, pp. e998–e1007, Sep. 2018, doi: 10.1016/S2214-109X(18)30275-4.

[215] P. Ge *et al.*, "The High Prevalence of Low HDL-Cholesterol Levels and Dyslipidemia in Rural Populations in Northwestern China," 2015, doi: 10.1371/journal.pone.0144104.

[216] A. A. Alsheikh-Ali *et al.*, "Cardiovascular Risk Factor Burden in Africa and the Middle East: The Africa Middle East Cardiovascular Epidemiological (ACE) Study," *PLoS One*, vol. 9, no. 8, p. e102830, Aug. 2014, doi: 10.1371/journal.pone.0102830.

[217] M. O. Rourke and E. D. Frohlich, "Pulse Pressure - Is This a Clinically Useful Risk Factor?," *Hypertension*, vol. 34, pp. 372–374, 1999, doi: 10.1016/j.jacc.2015.11.022.

[218] F. Q. Nuttall, "Body mass index: Obesity, BMI, and health: A critical review," *Nutr. Today*, vol. 50, no. 3, pp. 117–128, 2015, doi: 10.1097/NT.0000000000000092.

[219] M. Ashwell, L. Mayhew, J. Richardson, and B. Rickayzen, "Waist-to-height ratio is more predictive of years of life lost than body mass index," *PLoS One*, vol. 9, no. 9, 2014, doi: 10.1371/journal.pone.0103483.

[220] C. M. Y. Lee, R. R. Huxley, R. P. Wildman, and M. Woodward, "Indices of abdominal obesity are better discriminators of cardiovascular risk factors than BMI: a meta-analysis," *J. Clin. Epidemiol.*, vol. 61, no. 7, pp. 646–653, 2008, doi: 10.1016/j.jclinepi.2007.08.012.

[221] L. M. Browning, S. D. Hsieh, and M. Ashwell, "A systematic review of waist-to-height ratio as a screening tool for the prediction of cardiovascular disease and diabetes: 05 could be a suitable global boundary value," *Nutr. Res. Rev.*, vol. 23, no. 2, pp. 247–269, 2010, doi: 10.1017/S0954422410000144.

[222] A. M. Nevill, M. J. Duncan, I. M. Lahart, and G. R. Sandercock, "Scaling waist girth for differences in body size reveals a new improved index associated with cardiometabolic risk," *Scand. J. Med. Sci. Sport.*, vol. 27, no. 11, pp. 1470–1476, 2017, doi: 10.1111/sms.12780.

[223] M. G. Swainson, A. M. Batterham, C. Tsakirides, Z. H. Rutherford, and K. Hind, "Prediction of whole-body fat percentage and visceral adipose tissue mass from five anthropometric variables," *PLoS One*, vol. 12, no. 5, pp. 1–12, 2017, doi: 10.1371/journal.pone.0177175.

[224] M. Ashwell and S. Gibson, "A proposal for a primary screening tool: 'Keep your waist circumference to less than half your height,'" *BMC Med.*, vol. 12, no. 1, pp. 1–6, 2014, doi: 10.1186/s12916-014-0207-1.

[225] S. Andersen, K. Fleischer Rex, P. Noahsen, H. C. Florian Sørensen, G. Mulvad, and P. Laurberg, "Raised BMI cut-off for overweight in Greenland Inuit - A review," *Int. J. Circumpolar Health*, vol. 72, no. SUPPL.1, 2013, doi: 10.3402/ijch.v72i0.21086.

[226] S. A. Lear, P. T. James, G. T. Ko, and S. Kumanyika, "Appropriateness of waist circumference and waist-to-hip ratio cutoffs for different ethnic groups," *Eur. J. Clin. Nutr.*, vol. 64, pp. 42–61, 2010, doi: 10.1038/ejcn.2009.70.

[227] S. M. Pearson, "Landscape Ecology and Population Dynamics," in *Encyclopedia of Biodiversity: Second Edition*, Elsevier Inc., 2013, pp. 488–502.

[228] E. Trinkaus, “An abundance of developmental anomalies and abnormalities in Pleistocene people.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 115, no. 47, pp. 11941–11946, Nov. 2018, doi: 10.1073/pnas.1814989115.

[229] P. D. Keightley, M. J. Lercher, and A. Eyre-Walker, “Evidence for Widespread Degradation of Gene Control Regions in Hominid Genomes,” doi: 10.1371/journal.pbio.0030042.

[230] A. H. Bittles and M. L. Black, “Consanguinity, human evolution, and complex diseases.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107 Suppl, no. Suppl 1, pp. 1779–86, Jan. 2010, doi: 10.1073/pnas.0906079106.

[231] M. Teschler-Nicola, C. Czerny, M. Oliva, D. Schamall, and M. Schultz, “Pathological Alterations and Traumas in the Human Skeletal Remains from Mladeč,” in *Early Modern Humans at the Moravian Gate*, Vienna: Springer Vienna, pp. 473–489.

[232] M. C. Dean *et al.*, “Longstanding dental pathology in Neandertals from El Sidrón (Asturias, Spain) with a probable familial basis,” *J. Hum. Evol.*, vol. 64, no. 6, pp. 678–686, Jun. 2013, doi: 10.1016/J.JHEVOL.2013.03.004.

[233] L. Ríos *et al.*, “Possible Further Evidence of Low Genetic Diversity in the El Sidrón (Asturias, Spain) Neandertal Group: Congenital Clefts of the Atlas.,” *PLoS One*, vol. 10, no. 9, p. e0136550, 2015, doi: 10.1371/journal.pone.0136550.

[234] X.-J. Wu, S. Xing, and E. Trinkaus, “An Enlarged Parietal Foramen in the Late Archaic Xujiayao 11 Neurocranium from Northern China, and Rare Anomalies among Pleistocene Homo,” *PLoS One*, vol. 8, no. 3, p. e59587, Mar. 2013, doi: 10.1371/journal.pone.0059587.

[235] K. Prüfer *et al.*, “The complete genome sequence of a Neanderthal from the Altai Mountains,” *Nature*, vol. 505, no. 7481, pp. 43–49, 2014, doi: 10.1038/nature12886.

[236] C. Lalueza-Fox *et al.*, “Genetic evidence for patrilocal mating behavior among Neandertal groups.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 1, pp. 250–3, Jan. 2011, doi: 10.1073/pnas.1011553108.

[237] Q. Fu *et al.*, “Genome sequence of a 45,000-year-old modern human from western Siberia,” *Nature*, vol. 514, no. 7253, pp. 445–449, Oct. 2014, doi: 10.1038/nature13810.

[238] H. C. Harpending, M. A. Batzer, M. Gurven, L. B. Jorde, A. R. Rogers, and S. T. Sherry, “Genetic traces of ancient demography.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 95, no. 4, pp. 1961–7, Feb. 1998, doi: 10.1073/pnas.95.4.1961.

[239] A. Tenesa *et al.*, “Recent human effective population size estimated from linkage disequilibrium,” *Genome Res.*, vol. 17, no. 4, p. 520, Apr. 2007, doi: 10.1101/GR.6023607.

[240] L. A. Zhivotovsky, N. A. Rosenberg, and M. W. Feldman, “Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers.,” *Am. J. Hum. Genet.*, vol. 72, no. 5, pp. 1171–86, May 2003, doi: 10.1086/375120.

[241] H. Liu, F. Prugnolle, A. Manica, and F. Balloux, “A geographically explicit genetic model of worldwide human-settlement history.,” *Am. J. Hum. Genet.*, vol. 79, no. 2, pp. 230–7, Aug. 2006, doi: 10.1086/505436.

[242] C. D. Huff, J. Xing, A. R. Rogers, D. Witherspoon, and L. B. Jorde, “Mobile elements reveal small population size in the ancient ancestors of Homo sapiens,” doi: 10.1073/pnas.0909000107.

[243] Dadax, “Current World Population.” https://www.worldometers.info/world-population/ (accessed Aug. 18,

2019).

[244] E. Gazave, D. Chang, A. G. Clark, and A. Keinan, “Population growth inflates the per-individual number of deleterious mutations and reduces their mean effect,” *Genetics*, vol. 195, no. 3, pp. 969–978, 2013, doi: 10.1534/genetics.113.153973.

[245] V. M. Cabrera, “Human molecular evolutionary rate, time dependency and transient polymorphism effects viewed through ancient and modern mitochondrial DNA genomes,” *Nature*, vol. 11, no. 5036, 2021, doi: 10.1038/s41598-021-84583-1.

[246] T.-S. C. Jobling Mark, Hollox Edward, Hurles Matthew, Kivisild Toomas, *Human Evolutionary Genetics*, 2nd ed., no. 4. Garland Science, 2014.

[247] H. Harpending and A. Rogers, “Genetic perspectives on human origins and differentiation,” *Annual Review of Genomics and Human Genetics*, vol. 1, no. 2000. Annual Reviews Inc., pp. 361–385, Nov. 28, 2000, doi: 10.1146/annurev.genom.1.1.361.

[248] W. Amos and J. I. Hoffman, “Evidence that two main bottleneck events shaped modern human genetic diversity,” *Proc. R. Soc. B Biol. Sci.*, vol. 277, no. 1678, pp. 131–137, Jan. 2010, doi: 10.1098/rspb.2009.1473.

[249] S. Kohl and C. Hamel, “Clinical utility gene card for: Achromatopsia-update 2013,” *Eur. J. Hum. Genet.*, vol. 21, no. 11, p. 1324, Nov. 2013, doi: 10.1038/ejhg.2013.44.

[250] G. L. N Morton, R Lew, I Hussels, “Pingelap and Mokil Atolls: Historical Genetics,” *Am. J. Hum. Genet.*, vol. 24, no. 3, pp. 304–309, 1972.

[251] I. E. Hussels’ And, N. E. Morton’, I. E. Hussels, and N. E. Morton, “Pingelap and Mokil Atolls: Achromatopsia,” *Am. J. Hum. Genet.*, vol. 24, no. 3, pp. 304–309, 1972.

[252] B. Triggs-Raine *et al.*, “Development of a diagnostic DNA chip to screen for 30 autosomal recessive disorders in the Hutterite population.,” *Mol. Genet. genomic Med.*, vol. 4, no. 3, pp. 312–21, May 2016, doi: 10.1002/mgg3.206.

[253] J. X. Chong, R. Ouwenga, R. L. Anderson, D. J. Waggoner, and C. Ober, “A Population-Based Study of Autosomal-Recessive Disease-Causing Mutations in a Founder Population,” *Am. J. Hum. Genet.*, vol. 91, no. 4, pp. 608–620, Oct. 2012, doi: 10.1016/j.ajhg.2012.08.007.

[254] K. Klinger, G. T. Horn, P. Stanislovitis, R. H. Schwartz, T. M. Fujiwara, and K. Morgan §’ 1i’, “Cystic Fibrosis Mutations in the Hutterite Brethren,” 1990.

[255] A. H. Bittles and M. L. Black, “Consanguineous Marriage and Human Evolution,” *Annu. Rev. Anthropol.*, vol. 39, no. 1, pp. 193–207, Oct. 2010, doi: 10.1146/annurev.anthro.012809.105051.

[256] A. Bener and R. R. Mohammad, “Global distribution of consanguinity and their impact on complex diseases: Genetic disorders from an endogamous population,” *Egypt. J. Med. Hum. Genet.*, vol. 18, no. 4, pp. 315–320, Oct. 2017, doi: 10.1016/J.EJMHG.2017.01.002.

[257] S. Mcbrearty and A. S. Brooks, “The revolution that wasn’t: A new interpretation of the origin of modern human behavior,” *J. Hum. Evol.*, vol. 39, no. 5, pp. 453–563, 2000, doi: 10.1006/jhev.2000.0435.

[258] K. W. Alt *et al.*, “Earliest Evidence for Social Endogamy in the 9,000-Year-Old-Population of Basta, Jordan,” *PLoS One*, vol. 8, no. 6, p. e65649, Jun. 2013, doi: 10.1371/journal.pone.0065649.

[259] H. Hamamy, “Consanguineous marriages - preconception consultation in primary health care settings,” *J. Community Genet.*, vol. 3, no. 3, pp. 185–92, Jul. 2012, doi: 10.1007/s12687-011-0072-y.

[260] A. H. Bittles, “The Role and Significance of Consanguinity as a Demographic Variable,” *Popul. Dev. Rev.*, vol. 20, no. 3, p. 561, Sep. 1994, doi: 10.2307/2137601.

[261] H. Hamamy *et al.*, “Consanguineous marriages, pearls and perils: Geneva International Consanguinity Workshop Report,” *Genet. Med.*, vol. 13, no. 9, pp. 841–847, Sep. 2011, doi: 10.1097/GIM.0b013e318217477f.

[262] A. H. Bittles and M. L. Black, “Evolution in health and medicine Sackler colloquium: Consanguinity, human evolution, and complex diseases.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107 Suppl, no. Suppl 1, pp. 1779–86, Jan. 2010, doi: 10.1073/pnas.0906079106.

[263] G. O. Tadmouri, P. Nair, T. Obeid, M. T. Al Ali, N. Al Khaja, and H. A. Hamamy, “Consanguinity and reproductive health among Arabs.,” *Reprod. Health*, vol. 6, p. 17, Oct. 2009, doi: 10.1186/1742-4755-6-17.

[264] Bittles A H, *Consanguinity in Context*. Cambridge University Press, 2012.

[265] Bittles Alan H, Grant Jonathan C, and Shami Sajjad A, “Consanguinity as a Determinant of Reproductive Behaviour and Mortality in Pakistan,” *Int. J. Epidemiol.*, 1993.

[266] T. S. S. Rao, M. R. Asha, K. Sambamurthy, and K. S. J. Rao, “Consanguinity: Still a challenge.,” *Indian J. Psychiatry*, vol. 51, no. 1, pp. 3–5, Jan. 2009, doi: 10.4103/0019-5545.44897.

[267] A. L. Sandridge, J. Takeddin, E. Al-Kaabi, and Y. Frances, “Consanguinity in Qatar: Knowledge, attitude and practice in a population born between 1946 and 1991,” *J. Biosoc. Sci.*, vol. 42, no. 1, pp. 59–82, Jan. 2010, doi: 10.1017/S002193200999023X.

[268] C. J. Clark, A. Hill, K. Jabbar, and J. G. Silverman, “Violence During Pregnancy in Jordan,” *Violence Against Women*, vol. 15, no. 6, pp. 720–735, Jun. 2009, doi: 10.1177/1077801209332191.

[269] F. F. Fikree, S. N. Jafarey, R. Korejo, A. Afshan, and J. M. Durocher, “Intimate partner violence before and during pregnancy: Experiences of postpartum women in Karachi, Pakistan,” *J. Pak. Med. Assoc.*, vol. 56, no. 6, pp. 253–257, 2006.

[270] F. El-Zanaty and A. Way, “Egypt Demographic and Health Survey 2005,” Cairo, 2006.

[271] J. Hawks, “Neanderthals and Denisovans as biological invaders.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, no. 37, pp. 9761–9763, Sep. 2017, doi: 10.1073/pnas.1713163114.

[272] “Inbreeding depression.” https://evolution.berkeley.edu/evolibrary/article/conservation_03 (accessed Apr. 06, 2021).

[273] Charles Darwin, “The Effects of Cross and Self Fertilisation in the Vegetable Kingdom - Charles Darwin - Google Books,” *D.Appleton and Company*, 1895. https://books.google.ca/books/about/The_Effects_of_Cross_and_Self_Fertilisat.html (accessed Aug. 12, 2019).

[274] L. F. Garrod, A, Oxon M, “The incidence of alkaptonuria: a study in chemical individuality (reproduced from the Lancet, 1902),” *Mol. Med.*, vol. 2, no. 3, pp. 274–282, 1992.

[275] A. M. Erzurumluoglu, H. A. Shihab, S. Rodriguez, T. R. Gaunt, and I. N. M. Day, “Importance of Genetic

Studies in Consanguineous Populations for the Characterization of Novel Human Gene Functions.," *Ann. Hum. Genet.*, vol. 80, no. 3, pp. 187–96, May 2016, doi: 10.1111/ahg.12150.

[276] M. A. Bhinder *et al.*, "Consanguinity: A blessing or menace at population level?," *Ann. Hum. Genet.*, vol. 83, no. 4, pp. 214–219, Jul. 2019, doi: 10.1111/ahg.12308.

[277] F. C. Ceballos, P. K. Joshi, D. W. Clark, M. Ramsay, and J. F. Wilson, "Runs of homozygosity: windows into population history and trait architecture," *Nat. Rev. Genet.*, vol. 19, no. 4, pp. 220–234, Apr. 2018, doi: 10.1038/nrg.2017.109.

[278] W. A. Anwar, M. Khyatti, and K. Hemminki, "Consanguinity and genetic diseases in North Africa and immigrants to Europe," *Eur. J. Public Health*, vol. 24, no. suppl 1, pp. 57–63, Aug. 2014, doi: 10.1093/eurpub/cku104.

[279] J. K. Pritchard, "Are rare variants responsible for susceptibility to complex diseases?," *Am. J. Hum. Genet.*, vol. 69, no. 1, pp. 124–137, 2001, doi: 10.1086/321272.

[280] E. J. Lyons *et al.*, "Homozygosity and risk of childhood death due to invasive bacterial disease," *BMC Med. Genet.*, vol. 10, no. 1, p. 55, Dec. 2009, doi: 10.1186/1471-2350-10-55.

[281] E. J. Lyons, A. J. Frodsham, L. Zhang, A. V. S. Hill, and W. Amos, "Consanguinity and susceptibility to infectious diseases in humans," *Biol. Lett.*, vol. 5, no. 4, pp. 574–576, Aug. 2009, doi: 10.1098/rsbl.2009.0133.

[282] A. H. el-Orfi, M. Singh, and A. S. Giasuddin, "Conjugal leprosy among Libyan patients.," *Dermatology*, vol. 196, no. 2, pp. 271–2, 1998.

[283] I. Rudan *et al.*, "Inbreeding and risk of late onset complex disease," *J. Med. Genet.*, vol. 40, no. 12, pp. 925–932, Dec. 2003, doi: 10.1136/jmg.40.12.925.

[284] M. Fareed and M. Afzal, "Increased cardiovascular risks associated with familial inbreeding: a population-based study of adolescent cohort.," *Ann. Epidemiol.*, vol. 26, no. 4, pp. 283–92, Apr. 2016, doi: 10.1016/j.annepidem.2016.03.001.

[285] J. T. C. Shieh, A. H. Bittles, and L. Hudgins, "Consanguinity and the risk of congenital heart disease," *Am. J. Med. Genet. Part A*, vol. 158A, no. 5, pp. 1236–1241, May 2012, doi: 10.1002/ajmg.a.35272.

[286] P. Christofidou *et al.*, "Runs of Homozygosity: Association with Coronary Artery Disease and Gene Expression in Monocytes and Macrophages," *Am. J. Hum. Genet.*, vol. 97, no. 2, pp. 228–237, Aug. 2015, doi: 10.1016/j.ajhg.2015.06.001.

[287] R. L. McLaughlin *et al.*, "Homozygosity mapping in an Irish ALS case–control cohort describes local demographic phenomena and points towards potential recessive risk loci," *Genomics*, vol. 105, no. 4, pp. 237–241, Apr. 2015, doi: 10.1016/j.ygeno.2015.01.002.

[288] N. E. Morton, "Effect of inbreeding on IQ and mental retardation.," *Proc. Natl. Acad. Sci.*, vol. 75, no. 8, pp. 3906–3908, Aug. 1978, doi: 10.1073/pnas.75.8.3906.

[289] R. Lakhan, R. Bipeta, S. S. R. R. Yerramilli, and V. K. Nahar, "A Family Study of Consanguinity in Children with Intellectual Disabilities in Barwani, India.," *J. Neurosci. Rural Pract.*, vol. 8, no. 4, pp. 551–555, 2017, doi: 10.4103/jnrp.jnrp_104_17.

[290] Y. Wadi Alanazi *et al.*, "Consanguinity between Parents and Risk of Epilepsy among Children in Northern

Saudi Arabia," *Egypt. J. Hosp. Med.*, vol. 70, no. 11, pp. 1925–1928, 2018, doi: 10.12816/0044844.

[291] K. B. Bulayeva *et al.*, "Mapping genes of complex psychiatric diseases in Daghestan genetic isolates.," *Am. J. Med. Genet. B. Neuropsychiatr. Genet.*, vol. 132B, no. 1, pp. 76–84, Jan. 2005, doi: 10.1002/ajmg.b.30073.

[292] W. Fatima *et al.*, "Parental Consanguineous Marriages are Associated with Early Age of Onset of Schizophrenia in a Pakistani Cohort," *Int. J. Hum. Genet.*, vol. 17, no. 3, pp. 135–139, Jul. 2017, doi: 10.1080/09723757.2017.1368221.

[293] M. C. Keller *et al.*, "Runs of homozygosity implicate autozygosity as a schizophrenia risk factor.," *PLoS Genet.*, vol. 8, no. 4, p. e1002656, Apr. 2012, doi: 10.1371/journal.pgen.1002656.

[294] L. A. Farrer, A. Bowirrat, R. P. Friedland, K. Waraska, A. D. Korczyn, and C. T. Baldwin, "Identification of multiple loci for Alzheimer disease in a consanguineous Israeli-Arab community," *Hum. Mol. Genet.*, vol. 12, no. 4, pp. 415–422, Feb. 2003, doi: 10.1093/hmg/ddg037.

[295] B. Ujvari *et al.*, "Genetic diversity, inbreeding and cancer.," *Proceedings. Biol. Sci.*, vol. 285, no. 1875, 2018, doi: 10.1098/rspb.2017.2589.

[296] H. Thomsen *et al.*, "Runs of homozygosity and inbreeding in thyroid cancer," *BMC Cancer*, vol. 16, no. 1, p. 227, Dec. 2016, doi: 10.1186/s12885-016-2264-7.

[297] H. Thomsen *et al.*, "Evidence of Inbreeding in Hodgkin Lymphoma," *PLoS One*, vol. 11, no. 4, p. e0154259, Apr. 2016, doi: 10.1371/journal.pone.0154259.

[298] M. D. Bacolod *et al.*, "The Signatures of Autozygosity among Patients with Colorectal Cancer," *Cancer Res.*, vol. 68, no. 8, pp. 2610–2621, Apr. 2008, doi: 10.1158/0008-5472.CAN-07-5250.

[299] J. L. Simpson, A. O. Martin, S. Elias, G. E. Sarto, and J. K. Dunn, "Cancers of the breast and female genital system: search for recessive genetic factors through analysis of human isolate.," *Am. J. Obstet. Gynecol.*, vol. 141, no. 6, pp. 629–36, Nov. 1981, doi: 10.1016/s0002-9378(15)33302-0.

[300] I. M. Gosadi, E. C. Goyder, and M. D. Teare, "Investigating the Potential Effect of Consanguinity on Type 2 Diabetes Susceptibility in a Saudi Population," *Hum Hered*, vol. 77, pp. 197–206, 2014, doi: 10.1159/000362447.

[301] A. M. Habeb *et al.*, "Permanent neonatal diabetes: different aetiology in Arabs compared to Europeans.," *Arch. Dis. Child.*, vol. 97, no. 8, pp. 721–3, Aug. 2012, doi: 10.1136/archdischild-2012-301744.

[302] I. Rudan *et al.*, "Inbreeding and susceptibility to osteoporosis in Croatian island isolates.," *Coll. Antropol.*, vol. 28, no. 2, pp. 585–601, Dec. 2004.

[303] M. Fareed and M. Afzal, "Evidence of inbreeding depression on height, weight, and body mass index: A population-based child cohort study," *Am. J. Hum. Biol.*, vol. 26, no. 6, pp. 784–795, Nov. 2014, doi: 10.1002/ajhb.22599.

[304] R. McQuillan *et al.*, "Evidence of Inbreeding Depression on Human Height," *PLoS Genet.*, vol. 8, no. 7, p. e1002655, Jul. 2012, doi: 10.1371/journal.pgen.1002655.

[305] Badaruddoza, "Inbreeding effects on metrical phenotypes among North Indian Children.," *Coll. Antropol.*, vol. 28 Suppl 2, pp. 311–9, 2004.

[306] P. K. Joshi *et al.*, "Directional dominance on stature and cognition in diverse human populations.," *Nature*,

vol. 523, no. 7561, pp. 459–462, Jul. 2015, doi: 10.1038/nature14618.

[307] H. Krieger, "Inbreeding effects on metrical traits in Northeastern Brazil," *Am. J. Hum. Genet.*, vol. 21, no. 6, pp. 537–546, 1969.

[308] H. Campbell *et al.*, "Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits," *Hum. Mol. Genet.*, vol. 16, no. 2, pp. 233–241, Jan. 2007, doi: 10.1093/hmg/ddl473.

[309] R. A. Power, C. Nagoshi, J. C. DeFries, R. Plomin, and Wellcome Trust Case Control Consortium 2, "Genome-wide estimates of inbreeding in unrelated individuals and their association with cognitive ability.," *Eur. J. Hum. Genet.*, vol. 22, no. 3, pp. 386–90, Mar. 2014, doi: 10.1038/ejhg.2013.155.

[310] D. P. Howrigan *et al.*, "Genome-wide autozygosity is associated with lower general cognitive ability.," *Mol. Psychiatry*, vol. 21, no. 6, pp. 837–43, Jun. 2016, doi: 10.1038/mp.2015.120.

[311] K. J. H. Verweij *et al.*, "The association of genotype-based inbreeding coefficient with a range of physical and psychological human traits.," *PLoS One*, vol. 9, no. 7, p. e103102, Jul. 2014, doi: 10.1371/journal.pone.0103102.

[312] O. S. Alfi, R. Chang, and S. P. Azen, "Evidence for genetic control of nondisjunction in man," *Am. J. Hum. Genet.*, vol. 32, no. 4, pp. 477–483, 1980.

[313] M. M. Hashishe, "Genealogic Study in Down Syndrome," *Alexandria J. Pediatr.*, vol. 12, no. 1, pp. 199–204, 1998.

[314] A. Ray *et al.*, "Risk of Down syndrome birth: Consanguineous marriage is associated with maternal meiosis-II nondisjunction at younger age and without any detectable recombination error," *Am. J. Med. Genet. Part A*, vol. 176, no. 11, pp. 2342–2349, Nov. 2018, doi: 10.1002/ajmg.a.40511.

[315] A. A. Rezayat *et al.*, "Down syndrome and consanguinity," *J. Res. Med. Sci.*, vol. 18, no. 11, pp. 995–997, 2013.

[316] R. Sayee and I. M. Thomas, "Consanguinity, non-disjunction, parental age and Down's syndrome.," *J. Indian Med. Assoc.*, vol. 96, no. 11, pp. 335–337, 1998.

[317] A. Helgason, S. Palsson, D. F. Guthbjartsson, t. Kristjansson, and K. Stefansson, "An Association Between the Kinship and Fertility of Human Couples," *Science (80-.).*, vol. 319, no. 5864, pp. 813–816, Feb. 2008, doi: 10.1126/science.1150232.

[318] A. H. Bittles, J. C. Grant, S. G. Sullivan, and R. Hussain, "Does inbreeding lead to decreased human fertility?," *Ann. Hum. Biol.*, vol. 29, no. 2, pp. 111–30, Jan. 2002, doi: 10.1080/03014460110075657.

[319] Z. A. Szpiech *et al.*, "Long runs of homozygosity are enriched for deleterious variation," *Am. J. Hum. Genet.*, vol. 93, no. 1, pp. 90–102, Jul. 2013, doi: 10.1016/j.ajhg.2013.05.003.

[320] E. L. Weiser, C. E. Grueber, E. S. Kennedy, and I. G. Jamieson, "Unexpected positive and negative effects of continuing inbreeding in one of the world's most inbred wild animals," *Evolution (N. Y).*, vol. 70, no. 1, pp. 154–166, Jan. 2016, doi: 10.1111/evo.12840.

[321] E. Moreno, J. Pérez-González, J. Carranza, and J. Moya-Laraño, "Better Fitness in Captive Cuvier's Gazelle despite Inbreeding Increase: Evidence of Purging?," *PLoS One*, vol. 10, no. 12, p. e0145111, 2015, doi: 10.1371/journal.pone.0145111.

[322] D. W. Nebert, M. Gálvez-Peralta, Z. Shi, and N. Dragin, "Inbreeding and epigenetics: beneficial as well as deleterious effects.," *Nat. Rev. Genet.*, vol. 11, no. 9, p. 662, Sep. 2010, doi: 10.1038/nrg2664-c2.

[323] M. Thorven *et al.*, "A homozygous nonsense mutation (428G-->A) in the human secretor (FUT2) gene provides resistance to symptomatic norovirus (GGII) infections.," *J. Virol.*, vol. 79, no. 24, pp. 15351–5, Dec. 2005, doi: 10.1128/JVI.79.24.15351-15355.2005.

[324] H. Maghzi, V. Shaygannejad, A. Minagar, A. Hassanzadeh, and A.-H. Maghzi, "Consanguinity and multiple sclerosis susceptibility: A case control study.," *Mult. Scler. Relat. Disord.*, vol. 10, pp. 179–180, Nov. 2016, doi: 10.1016/j.msard.2016.09.013.

[325] T. Cherif W, "Does Consanguinity Protect Against Breast Cancer in Tunisian Population?," *Hered. Genet.*, vol. 03, no. 02, 2014, doi: 10.4172/2161-1041.1000130.

[326] S. Denic, C. Frampton, and M. G. Nicholls, "Risk of cancer in an inbred population," *Cancer Detect. Prev.*, vol. 31, no. 4, pp. 263–269, Jan. 2007, doi: 10.1016/j.cdp.2007.07.006.

[327] A. Bener *et al.*, "Impact of consanguinity on cancer in a highly endogamous population," *Asian Pacific J. Cancer Prev.*, vol. 10, no. 1, pp. 35–40, 2009.

[328] I. A. Hoppenbrouwers *et al.*, "Familial clustering of multiple sclerosis in a Dutch genetic isolate," *Mult. Scler. J.*, vol. 13, no. 1, pp. 17–24, Jan. 2007, doi: 10.1177/1352458506071216.

[329] I. Rudan, "Inbreeding and cancer incidence in human isolates," *Hum. Biol.*, vol. 71, no. 2, pp. 173–187, 1999.

[330] S. A. Shami, R. Qaisar, and A. H. Bittles, "Consanguinity and adult morbidity in Pakistan.," *Lancet (London, England)*, vol. 338, no. 8772, p. 954, Oct. 1991, doi: 10.1016/0140-6736(91)91828-i.

[331] J. E. Powell, P. M. Visscher, and M. E. Goddard, "Reconciling the analysis of IBD and IBS in complex trait studies," *Nat. Rev. Genet.*, vol. 11, no. 11, pp. 800–805, Nov. 2010, doi: 10.1038/nrg2865.

[332] E. A. Thompson, "Identity by descent: Variation in meiosis, across genomes, and in populations," *Genetics*, vol. 194, no. 2. pp. 301–326, 2013, doi: 10.1534/genetics.112.148825.

[333] S. R. Browning and B. L. Browning, "Identity by Descent Between Distant Relatives: Detection and Applications," *Annu. Rev. Genet.*, vol. 46, no. 1, pp. 617–633, Dec. 2012, doi: 10.1146/annurev-genet-110711-155534.

[334] E. L. Stevens, G. Heckenberg, E. D. O. Roberson, J. D. Baugher, T. J. Downey, and J. Pevsner, "Inference of Relationships in Population Data Using Identity-by-Descent and Identity-by-State," *PLoS Genet.*, vol. 7, no. 9, p. e1002287, Sep. 2011, doi: 10.1371/journal.pgen.1002287.

[335] K. Lipscomb Sund and C. W. Rehder, "Detection and Reporting of Homozygosity Associated with Consanguinity in the Clinical Laboratory," *Hum Hered*, vol. 77, pp. 217–224, 2014, doi: 10.1159/000362448.

[336] G. L. Ryland *et al.*, "Loss of heterozygosity: What is it good for?," *BMC Med. Genomics*, vol. 8, no. 1, p. 45, Aug. 2015, doi: 10.1186/s12920-015-0123-z.

[337] T. F. Chaves *et al.*, "Long contiguous stretches of homozygosity detected by chromosomal microarrays (CMA) in patients with neurodevelopmental disorders in the South of Brazil," *BMC Med. Genomics*, vol. 12, no. 1, p. 50, Mar. 2019, doi: 10.1186/s12920-019-0496-5.

[338] T. J. Pemberton, D. Absher, M. W. Feldman, R. M. Myers, N. A. Rosenberg, and J. Z. Li, "Genomic patterns

of homozygosity in worldwide human populations," *Am. J. Hum. Genet.*, vol. 91, no. 2, pp. 275–292, Aug. 2012, doi: 10.1016/j.ajhg.2012.06.014.

[339] J. Gibson, N. E. Morton, and A. Collins, "Extended tracts of homozygosity in outbred human populations," *Hum. Mol. Genet.*, vol. 15, no. 5, pp. 789–795, Mar. 2006, doi: 10.1093/hmg/ddi493.

[340] M. Nothnagel, T. T. Lu, M. Kayser, and M. Krawczak, "Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans," *Hum. Mol. Genet.*, vol. 19, no. 15, pp. 2927–2935, Aug. 2010, doi: 10.1093/hmg/ddq198.

[341] J. Stapley, P. G. D. Feulner, S. E. Johnston, A. W. Santure, and C. M. Smadja, "Variation in recombination frequency and distribution across eukaryotes: patterns and processes," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 372, no. 1736, Dec. 2017, doi: 10.1098/RSTB.2016.0455.

[342] E. L. Goode, "Linkage Disequilibrium," in *Encyclopedia of Cancer*, Springer Berlin Heidelberg, 2011, pp. 2043–2048.

[343] C. Sabatti and N. Risch, "Homozygosity and linkage disequilibrium.," *Genetics*, vol. 160, no. 4, pp. 1707–19, Apr. 2002.

[344] H. Uecker and J. Hermisson, "The Role of Recombination in Evolutionary Rescue.," *Genetics*, vol. 202, no. 2, pp. 721–32, Feb. 2016, doi: 10.1534/genetics.115.180299.

[345] R. C. Elston, "Linkage and association," *Genet. Epidemiol.*, vol. 15, no. 6, pp. 565–576, 1998, doi: 10.1002/(SICI)1098-2272(1998)15:6<565::AID-GEPI2>3.0.CO;2-J.

[346] M. E. Goddard and B. J. Hayes, "Mapping genes for complex traits in domestic animals and their use in breeding programmes," *Nat. Rev. Genet.*, vol. 10, no. 6, pp. 381–391, Jun. 2009, doi: 10.1038/nrg2575.

[347] M. Kirin, R. McQuillan, C. S. Franklin, H. Campbell, P. M. McKeigue, and J. F. Wilson, "Genomic Runs of Homozygosity Record Population History and Consanguinity," *PLoS One*, vol. 5, no. 11, p. e13996, Nov. 2010, doi: 10.1371/journal.pone.0013996.

[348] A. R. Rogers, "How population growth affects linkage disequilibrium.," *Genetics*, vol. 197, no. 4, pp. 1329–41, Aug. 2014, doi: 10.1534/genetics.114.166454.

[349] B. Derrida, S. C. Manrubia-And Damiad, and N. H. Zanette, "On the Genealogy of a Population of Biparental Individuals," *J. theor. Biol*, pp. 303–315, 2000, doi: 10.1006/jtbi.2000.1095.

[350] D. L. T. Rohde, S. Olson, and J. T. Chang, "Modelling the recent common ancestry of all living humans," *Nature*, vol. 431, no. 7008, pp. 562–566, Sep. 2004, doi: 10.1038/nature02842.

[351] L. B. Jorde *et al.*, "The distribution of human genetic diversity: A comparison of mitochondrial, autosomal, and Y-chromosome data," *Am. J. Hum. Genet.*, vol. 66, no. 3, pp. 979–988, 2000, doi: 10.1086/302825.

[352] S. Wright, "Coefficients of inbreeding and relationship," *Am. Nat.*, no. 56, pp. 330–338, 1922.

[353] C. A. B. Smith, "The Detection of Linkage in Human Genetics," *J. R. Stat. Soc. Ser. B*, vol. 15, no. 2, pp. 153–184, 1953, doi: 10.1111/j.2517-6161.1953.tb00133.x.

[354] E. S. Lander and D. Botstein, "Homozygosity mapping: A way to map human recessive traits with the DNA of inbred children," *Science (80-.).*, vol. 236, no. 4808, pp. 1567–1570, 1987, doi: 10.1126/science.2884728.

[355] N. E. MORTON, "Sequential tests for the detection of linkage.," *Am. J. Hum. Genet.*, vol. 7, no. 3, pp. 277–

318, Sep. 1955.

[356] D. R. Nyholt, “All LODs are not created equal.,” *Am. J. Hum. Genet.*, vol. 67, no. 2, pp. 282–8, Aug. 2000, doi: 10.1086/303029.

[357] K. W. Broman and J. L. Weber, “Long homozygous chromosomal segments in reference families from the centre d’Etude du polymorphisme humain.,” *Am. J. Hum. Genet.*, vol. 65, no. 6, pp. 1493–500, Dec. 1999, doi: 10.1086/302661.

[358] S. Wang, C. Haynes, F. Barany, and J. Ott, “Genome-wide autozygosity mapping in human populations,” *Genet. Epidemiol.*, vol. 33, no. 2, pp. 172–180, 2009, doi: 10.1002/gepi.20344.

[359] S. Wang, C. Haynes, F. Barany, and J. Ott, “Genome-wide autozygosity mapping in human populations.,” *Genet. Epidemiol.*, vol. 33, no. 2, pp. 172–80, Feb. 2009, doi: 10.1002/gepi.20344.

[360] A. Blant, M. Kwong, Z. A. Z. A. Szpiech, and T. J. T. J. Pemberton, “Weighted likelihood inference of genomic autozygosity patterns in dense genotype data,” *BMC Genomics*, vol. 18, no. 1, 2017, doi: 10.1186/s12864-017-4312-3.

[361] H. Chen, N. Patterson, and D. Reich, “Population differentiation as a test for selective sweeps,” *Genome Res.*, vol. 20, no. 3, pp. 393–402, Mar. 2010, doi: 10.1101/gr.100545.109.

[362] A. G. Clark, “The Size Distribution of Homozygous Segments in the Human Genome,” *Am. J. Hum. Genet.*, vol. 65, no. 6, pp. 1489–1492, Jul. 2002, doi: 10.1086/302668.

[363] A. Albrechtsen, T. S. Korneliussen, I. Moltke, T. van Overseem Hansen, F. C. Nielsen, and R. Nielsen, “Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium,” *Genet. Epidemiol.*, vol. 33, no. 3, pp. 266–274, 2009, doi: 10.1002/gepi.20378.

[364] A. Auton *et al.*, “Global distribution of genomic diversity underscores rich complex history of continental human populations.,” *Genome Res.*, vol. 19, no. 5, pp. 795–803, May 2009, doi: 10.1101/gr.088898.108.

[365] S. Purcell *et al.*, “PLINK: a tool set for whole-genome association and population-based linkage analyses.,” *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–75, Sep. 2007, doi: 10.1086/519795.

[366] M. C. Keller, P. M. Visscher, and M. E. Goddard, “Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data.,” *Genetics*, vol. 189, no. 1, pp. 237–49, Sep. 2011, doi: 10.1534/genetics.111.130922.

[367] I. M. MacLeod, D. M. Larkin, H. A. Lewin, B. J. Hayes, and M. E. Goddard, “Inferring demography from runs of homozygosity in whole-genome sequence, with correction for sequence errors.,” *Mol. Biol. Evol.*, vol. 30, no. 9, pp. 2209–23, Sep. 2013, doi: 10.1093/molbev/mst125.

[368] 1000 Genomes Project Consortium *et al.*, “A global reference for human genetic variation,” *Nature*, vol. 526, no. 7571, pp. 68–74, Oct. 2015, doi: 10.1038/nature15393.

[369] T. J. Pemberton, C. Wang, J. Z. Li, and N. A. Rosenberg, “Inference of unexpected genetic relatedness among individuals in HapMap Phase III.,” *Am. J. Hum. Genet.*, vol. 87, no. 4, pp. 457–64, Oct. 2010, doi: 10.1016/j.ajhg.2010.08.014.

[370] N. A. Rosenberg, S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard, and M. W. Feldman, “Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure,” *PLoS Genet.*, vol.

1, no. 6, p. e70, 2005, doi: 10.1371/journal.pgen.0010070.

[371] M. Kardos, G. Luikart, and F. W. Allendorf, "Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees.," *Heredity (Edinb).*, vol. 115, no. 1, pp. 63–72, Jul. 2015, doi: 10.1038/hdy.2015.17.

[372] A. Yu *et al.*, "Comparison of human genetic and sequence-based physical maps," *Nature*, vol. 409, no. 6822, pp. 951–953, Feb. 2001, doi: 10.1038/35057185.

[373] the 1000 G. Project *et al.*, "Variation in genome-wide mutation rates within and between human families," *Nat. Genet.*, vol. 43, no. 7, pp. 712–714, Jul. 2011, doi: 10.1038/ng.862.

[374] C. J. Hoggart *et al.*, "Sequence-level population simulations over large genomic regions," *Genetics*, vol. 177, no. 3, pp. 1725–1731, 2007, doi: 10.1534/genetics.106.069088.

[375] T. 1000 G. P. Consortium, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, Oct. 2010, doi: 10.1038/nature09534.

[376] D. P. Howrigan, M. A. Simonson, and M. C. Keller, "Detecting autozygosity through runs of homozygosity: A comparison of three autozygosity detection algorithms," *BMC Genomics*, vol. 12, Sep. 2011, doi: 10.1186/1471-2164-12-460.

[377] V. Narasimhan, P. Danecek, A. Scally, Y. Xue, C. Tyler-Smith, and R. Durbin, "BCFtools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data," *Bioinformatics*, vol. 32, no. 11, pp. 1749–1751, Jun. 2016, doi: 10.1093/bioinformatics/btw044.

[378] R. A. Fisher, *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, 1925.

[379] N. A. Rosenberg and M. G. B. Blum, "Sampling properties of homozygosity-based statistics for linkage disequilibrium," *Math. Biosci.*, vol. 208, no. 1, pp. 33–47, Jul. 2007, doi: 10.1016/j.mbs.2006.07.001.

[380] T. C. Matise *et al.*, "A second-generation combined linkage physical map of the human genome," *Genome Res.*, vol. 17, no. 12, pp. 1783–1786, Dec. 2007, doi: 10.1101/gr.7156307.

[381] C. Fraley, A. E. Raftery, and T. B. Murphy, "mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation," 2012.

[382] R Core Team (2017), "R: A language and environment for statistical computing," *R Foundation for Statistical Computing, Vienna, Austria.* .

[383] J. L. Hintze and R. D. Nelson, "Violin Plots: A Box Plot-Density Trace Synergism," *Am. Stat.*, vol. 52, no. 2, pp. 181–184, May 1998, doi: 10.1080/00031305.1998.10480559.

[384] R. R. Hudson, "Generating samples under a Wright-Fisher neutral model of genetic variation," *Bioinformatics*, vol. 18, no. 2, pp. 337–338, Feb. 2002, doi: 10.1093/bioinformatics/18.2.337.

[385] G. Bhatia, N. Patterson, S. Sankararaman, and A. L. Price, "Estimating and interpreting FST: the impact of rare variants.," *Genome Res.*, vol. 23, no. 9, pp. 1514–21, Sep. 2013, doi: 10.1101/gr.154831.113.

[386] C. Wang *et al.*, "Comparing Spatial Maps of Human Population-Genetic Variation Using Procrustes Analysis," *Stat. Appl. Genet. Mol. Biol.*, vol. 9, no. 1, Jan. 2010, doi: 10.2202/1544-6115.1493.

[387] A. Ferrer-Admetlla, M. Liang, T. Korneliussen, and R. Nielsen, "On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure," *Mol. Biol. Evol.*, vol. 31, no. 5, pp. 1275–1291, May 2014, doi:

10.1093/molbev/msu077.

[388] Z. A. Szpiech and R. D. Hernandez, "selscan: an efficient multithreaded program to perform EHH-based scans for positive selection.," *Mol. Biol. Evol.*, vol. 31, no. 10, pp. 2824–7, Oct. 2014, doi: 10.1093/molbev/msu211.

[389] S. Schiffels and R. Durbin, "Inferring human population size and separation history from multiple genome sequences.," *Nat. Genet.*, vol. 46, no. 8, pp. 919–25, Aug. 2014, doi: 10.1038/ng.3015.

[390] T. J. Pemberton *et al.*, "Using Population Mixtures to Optimize the Utility of Genomic Databases: Linkage Disequilibrium and Association Study Design in India," *Ann. Hum. Genet.*, vol. 72, no. 4, pp. 535–546, Jul. 2008, doi: 10.1111/j.1469-1809.2008.00457.x.

[391] M. Jakobsson *et al.*, "Genotype, haplotype and copy-number variation in worldwide human populations," *Nature*, vol. 451, no. 7181, pp. 998–1003, Feb. 2008, doi: 10.1038/nature06742.

[392] T. J. Pemberton and N. A. Rosenberg, "Population-genetic influences on genomic estimates of the inbreeding coefficient: A global perspective," *Hum. Hered.*, vol. 77, no. 1–4, pp. 37–48, 2014, doi: 10.1159/000362878.

[393] F. Prugnolle, A. Manica, and F. Balloux, "Geography predicts neutral genetic diversity of human populations.," *Curr. Biol.*, vol. 15, no. 5, pp. R159-60, Mar. 2005, doi: 10.1016/j.cub.2005.02.038.

[394] S. Ramachandran, O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, and L. L. Cavalli-Sforza, "Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa," *Proc. Natl. Acad. Sci.*, vol. 102, no. 44, pp. 15942–15947, Nov. 2005, doi: 10.1073/pnas.0507611102.

[395] M. DeGiorgio, M. Jakobsson, and N. A. Rosenberg, "Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa," *Proc. Natl. Acad. Sci.*, vol. 106, no. 38, pp. 16057–16062, Sep. 2009, doi: 10.1073/pnas.0903341106.

[396] T. J. Pemberton, M. DeGiorgio, and N. A. Rosenberg, "Population structure in a comprehensive genomic data set on human microsatellite variation.," *G3 (Bethesda).*, vol. 3, no. 5, pp. 891–907, May 2013, doi: 10.1534/g3.113.005728.

[397] S. Xu *et al.*, "Dissecting linkage disequilibrium in African-American genomes: roles of markers and individuals.," *Mol. Biol. Evol.*, vol. 24, no. 9, pp. 2049–58, Sep. 2007, doi: 10.1093/molbev/msm135.

[398] I. Silva-Zolezzi *et al.*, "Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 21, pp. 8611–6, May 2009, doi: 10.1073/pnas.0903045106.

[399] K. Bryc *et al.*, "Genome-wide patterns of population structure and admixture among Hispanic/Latino populations," *Proc. Natl. Acad. Sci.*, vol. 107, no. Supplement_2, pp. 8954–8961, May 2010, doi: 10.1073/pnas.0914618107.

[400] S. Wang *et al.*, "Geographic patterns of genome admixture in Latin American Mestizos.," *PLoS Genet.*, vol. 4, no. 3, p. e1000037, Mar. 2008, doi: 10.1371/journal.pgen.1000037.

[401] Z. A. Szpiech, A. Blant, and T. J. Pemberton, "GARLIC: Genomic Autozygosity Regions Likelihood-based Inference and Classification," *Bioinformatics*, vol. 33, no. 13, 2017, doi: 10.1093/bioinformatics/btx102.

[402] J. N. Fenner, "Cross-cultural estimation of the human generation interval for use in genetics-based population

divergence studies," *Am. J. Phys. Anthropol.*, vol. 128, no. 2, pp. 415–423, Oct. 2005, doi: 10.1002/ajpa.20188.

[403] J. D. Wall *et al.*, "Estimating genotype error rates from high-coverage next-generation sequence data.," *Genome Res.*, vol. 24, no. 11, pp. 1734–9, Nov. 2014, doi: 10.1101/gr.168393.113.

[404] A. Thomas, M. H. Skolnick, and C. M. Lewis, "Genomic mismatch scanning in pedigrees.," *IMA J. Math. Appl. Med. Biol.*, vol. 11, no. 1, pp. 1–16, 1994.

[405] A. Thomas, N. J. Camp, J. M. Farnham, K. Allen-Brady, and L. A. Cannon-Albright, "Shared genomic segment analysis. Mapping disease predisposition genes in extended pedigrees using SNP genotype assays.," *Ann. Hum. Genet.*, vol. 72, no. Pt 2, pp. 279–87, Mar. 2008, doi: 10.1111/j.1469-1809.2007.00406.x.

[406] A. Kong *et al.*, "A high-resolution recombination map of the human genome," *Nat. Genet.*, vol. 31, no. 3, pp. 241–247, Jul. 2002, doi: 10.1038/ng917.

[407] L. Wu, "Investigation of consanguineous marriages among 30 Chinese ethnic groups," *Hered Dis*, no. 4, pp. 163–166, 1987.

[408] P. S. Gadgil M, Joshi NV, Prasad UVS, Manoharan S, "Peopling of India," in *The Indian human heritage*, A. R. N. Balasubramanian D, Ed. Hyderabad, India: Universities Press, 1998, pp. 100–129.

[409] W. Zhusheng, "Dai | Encyclopedia.com," 1996. https://www.encyclopedia.com/science-and-technology/computers-and-electrical-engineering/computers-and-computing/dai (accessed Aug. 28, 2019).

[410] M. Kirin, R. McQuillan, C. S. Franklin, H. Campbell, P. M. McKeigue, and J. F. Wilson, "Genomic runs of homozygosity record population history and consanguinity.," *PLoS One*, vol. 5, no. 11, p. e13996, Nov. 2010, doi: 10.1371/journal.pone.0013996.

[411] T. M. Karafet *et al.*, "Extensive genome-wide autozygosity in the population isolates of Daghestan," *Eur. J. Hum. Genet.*, vol. 23, no. 10, pp. 1405–1412, Oct. 2015, doi: 10.1038/ejhg.2014.299.

[412] D. F. Conrad *et al.*, "A worldwide survey of haplotype variation and linkage disequilibrium in the human genome," *Nat. Genet.*, vol. 38, no. 11, pp. 1251–1260, Nov. 2006, doi: 10.1038/ng1911.

[413] L. Huang *et al.*, "Haplotype variation and genotype imputation in African populations," *Genet. Epidemiol.*, vol. 35, no. 8, pp. 766–780, Dec. 2011, doi: 10.1002/gepi.20626.

[414] J. Z. Li *et al.*, "Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation," *Science (80-.).*, vol. 319, no. 5866, pp. 1100–1104, Feb. 2008, doi: 10.1126/science.1153717.

[415] S. Gazal, M. Sahbatou, M.-C. Babron, E. Génin, and A.-L. Leutenegger, "High level of inbreeding in final phase of 1000 Genomes Project," *Sci. Rep.*, vol. 5, no. 1, p. 17453, Dec. 2015, doi: 10.1038/srep17453.

[416] Z. Shen *et al.*, "Genetic variation of 17 STR loci in Dai population in mainland China.," *Forensic Sci. Int. Genet.*, vol. 19, pp. 37–38, Nov. 2015, doi: 10.1016/j.fsigen.2015.05.015.

[417] H. Sun *et al.*, "Autosomal STRs Provide Genetic Evidence for the Hypothesis That Tai People Originate from Southern China," *PLoS One*, vol. 8, no. 4, p. e60822, Apr. 2013, doi: 10.1371/journal.pone.0060822.

[418] Y.-C. Li, W. Huang, J.-Y. Tian, X.-Q. Chen, and Q.-P. Kong, "Exploring the maternal history of the Tai people," *J. Hum. Genet.*, vol. 61, no. 8, pp. 721–729, Aug. 2016, doi: 10.1038/jhg.2016.36.

[419] R. Blekhman *et al.*, "Natural Selection on Genes that Underlie Human Disease Susceptibility," *Curr. Biol.*,

vol. 18, no. 12, pp. 883–889, Jun. 2008, doi: 10.1016/j.cub.2008.04.074.

[420] J. S. Berg *et al.*, "An informatics approach to analyzing the incidentalome," *Genet. Med.*, vol. 15, no. 1, pp. 36–44, Jan. 2013, doi: 10.1038/gim.2012.112.

[421] M. Lek *et al.*, "Analysis of protein-coding genetic variation in 60,706 humans," *Nature*, vol. 536, no. 7616, pp. 285–291, Aug. 2016, doi: 10.1038/nature19057.

[422] Y. Itan, B. L. Jones, C. J. Ingram, D. M. Swallow, and M. G. Thomas, "A worldwide correlation of lactase persistence phenotype and genotypes," *BMC Evol. Biol.*, vol. 10, no. 1, p. 36, Feb. 2010, doi: 10.1186/1471-2148-10-36.

[423] P. Gerbault *et al.*, "Evolution of lactase persistence: an example of human niche construction," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 366, no. 1566, pp. 863–877, Mar. 2011, doi: 10.1098/rstb.2010.0268.

[424] G. D. Smith *et al.*, "Lactase persistence-related genetic variant: population substructure and health outcomes," *Eur. J. Hum. Genet.*, vol. 17, no. 3, pp. 357–367, Mar. 2009, doi: 10.1038/ejhg.2008.156.

[425] C. J. E. Ingram, C. A. Mulcare, Y. Itan, M. G. Thomas, and D. M. Swallow, "Lactose digestion and the evolutionary genetics of lactase persistence," *Hum. Genet.*, vol. 124, no. 6, pp. 579–591, Jan. 2009, doi: 10.1007/s00439-008-0593-6.

[426] F. J. Simoons, "Primary adult lactose intolerance and the milking habit: A problem in biologic and cultural interrelations," *Am. J. Dig. Dis.*, vol. 15, no. 8, pp. 695–710, Aug. 1970, doi: 10.1007/BF02235991.

[427] R. D. McCracken, "Lactase Deficiency: An Example of Dietary Evolution," *Current Anthropology*, vol. 12. The University of Chicago PressWenner-Gren Foundation for Anthropological Research, pp. 479–517, doi: 10.2307/2740932.

[428] N. Kretchmer, "Lactose and lactase.," *Sci. Am.*, vol. 227, no. 4, pp. 71–8, Oct. 1972.

[429] Y. G. Wang *et al.*, "Prevalence of primary adult lactose malabsorption in three populations of northern China.," *Hum. Genet.*, vol. 67, no. 1, pp. 103–6, 1984, doi: 10.1007/bf00270566.

[430] I. Gallego Romero *et al.*, "Herders of Indian and European Cattle Share Their Predominant Allele for Lactase Persistence," *Mol. Biol. Evol.*, vol. 29, no. 1, pp. 249–260, Jan. 2012, doi: 10.1093/molbev/msr190.

[431] N. S. Enattah, T. Sahi, E. Savilahti, J. D. Terwilliger, L. Peltonen, and I. Järvelä, "Identification of a variant associated with adult-type hypolactasia," *Nat. Genet.*, vol. 30, no. 2, pp. 233–237, Feb. 2002, doi: 10.1038/ng826.

[432] T. Sahi, "Genetics and epidemiology of adult-type hypolactasia.," *Scand. J. Gastroenterol. Suppl.*, vol. 202, pp. 7–20, 1994.

[433] C. E. Woteki, E. Weser, and E. A. Young, "Lactose malabsorption in Mexican-American children," *Am. J. Clin. Nutr.*, vol. 29, no. 1, pp. 19–24, Jan. 1976, doi: 10.1093/ajcn/29.1.19.

[434] S. A. Tishkoff *et al.*, "Convergent adaptation of human lactase persistence in Africa and Europe," *Nat. Genet.*, vol. 39, no. 1, pp. 31–40, Jan. 2007, doi: 10.1038/ng1946.

[435] G. McVean, "A genealogical interpretation of principal components analysis.," *PLoS Genet.*, vol. 5, no. 10, p. e1000686, Oct. 2009, doi: 10.1371/journal.pgen.1000686.

[436] N. Patterson, A. L. Price, and D. Reich, "Population Structure and Eigenanalysis," *PLoS Genet.*, vol. 2, no.

12, p. e190, 2006, doi: 10.1371/journal.pgen.0020190.

[437] T. J. Pemberton *et al.*, "Impact of restricted marital practices on genetic variation in an endogamous Gujarati group," *Am. J. Phys. Anthropol.*, vol. 149, no. 1, pp. 92–103, Sep. 2012, doi: 10.1002/ajpa.22101.

[438] J. L. Hodges, Jr, and E. L. Lehmann, "The Efficiency of Some Nonparametric Competitors of the t-test," *The Annals of Mathematical Statistics*, vol. 27. Institute of Mathematical Statistics, pp. 324–335, doi: 10.2307/2236996.

[439] G. Chen *et al.*, "Genome-wide association study identifies novel loci association with fasting insulin and insulin resistance in African Americans.," *Hum. Mol. Genet.*, vol. 21, no. 20, pp. 4530–6, Oct. 2012, doi: 10.1093/hmg/dds282.

[440] N. Maeda *et al.*, "Diet-induced insulin resistance in mice lacking adiponectin/ACRP30," *Nat. Med.*, vol. 8, no. 7, pp. 731–737, Jul. 2002, doi: 10.1038/nm724.

[441] J. J. Díez and P. Iglesias, "The role of the novel adipocyte-derived hormone adiponectin in human disease.," *Eur. J. Endocrinol.*, vol. 148, no. 3, pp. 293–300, Mar. 2003.

[442] X. Sim *et al.*, "Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from Southeast Asia.," *PLoS Genet.*, vol. 7, no. 4, p. e1001363, Apr. 2011, doi: 10.1371/journal.pgen.1001363.

[443] E. E. Kenny *et al.*, "A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci.," *PLoS Genet.*, vol. 8, no. 3, p. e1002559, Mar. 2012, doi: 10.1371/journal.pgen.1002559.

[444] S. M. WALKER, C. P. DOWNES, and N. R. LESLIE, "TPIP: a novel phosphoinositide 3-phosphatase," *Biochem. J.*, vol. 360, no. 2, p. 277, Dec. 2001, doi: 10.1042/0264-6021:3600277.

[445] R. J. Clifford *et al.*, "Genetic variations at loci involved in the immune response are risk factors for hepatocellular carcinoma," *Hepatology*, vol. 52, no. 6, pp. 2034–2043, Dec. 2010, doi: 10.1002/hep.23943.

[446] L. Zuo *et al.*, "Sex difference of autosomal alleles in populations of European and African descent.," *Genes Genomics*, vol. 37, no. 12, pp. 1007–1016, Dec. 2015, doi: 10.1007/s13258-015-0332-z.

[447] J. M. VanLiere and N. A. Rosenberg, "Mathematical properties of the measure of linkage disequilibrium," *Theor. Popul. Biol.*, vol. 74, no. 1, pp. 130–137, Aug. 2008, doi: 10.1016/j.tpb.2008.05.006.

[448] T. Fung and K. Keenan, "Confidence Intervals for Population Allele Frequencies: The General Case of Sampling from a Finite Diploid Population of Any Size," *PLoS One*, vol. 9, no. 1, p. e85925, Jan. 2014, doi: 10.1371/journal.pone.0085925.

[449] H. M. Cann, "A Human Genome Diversity Cell Line Panel," *Science (80-.).*, vol. 296, no. 5566, pp. 261b – 262, Apr. 2002, doi: 10.1126/science.296.5566.261b.

[450] L. L. Cavalli-Sforza, "The Human Genome Diversity Project: past, present and future," *Nat. Rev. Genet.*, vol. 6, no. 4, pp. 333–340, Apr. 2005, doi: 10.1038/nrg1596.

[451] T. I. H. 3 Consortium, "Integrating common and rare genetic variation in diverse human populations," *Nature*, vol. 467, no. 7311, pp. 52–58, Sep. 2010, doi: 10.1038/nature09298.

[452] Z. A. Szpiech *et al.*, "Long runs of homozygosity are enriched for deleterious variation.," *Am. J. Hum. Genet.*, vol. 93, no. 1, pp. 90–102, Jul. 2013, doi: 10.1016/j.ajhg.2013.05.003.

[453] D. A. Baerenwald *et al.*, "Multiple functional polymorphisms in the G6PC2 gene contribute to the association

with higher fasting plasma glucose levels," *Diabetologia*, vol. 56, no. 6, pp. 1306–1316, Jun. 2013, doi: 10.1007/s00125-013-2875-3.

[454] S. M. Lieberman *et al.*, "Identification of the β cell antigen targeted by a prevalent population of pathogenic CD8+ T cells in autoimmune diabetes," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 14, pp. 8384–8388, Jul. 2003, doi: 10.1073/pnas.0932778100.

[455] S. S. Strautnieks *et al.*, "A gene encoding a liver-specific ABC transporter is mutated in progressive familial intrahepatic cholestasis," *Nat. Genet.*, vol. 20, no. 3, pp. 233–238, 1998, doi: 10.1038/3034.

[456] D. Weil *et al.*, "Usher syndrome type I G (USH1G) is caused by mutations in the gene encoding SANS, a protein that associates with the USH1C protein, harmonin," *Hum. Mol. Genet.*, vol. 12, no. 5, pp. 463–471, Mar. 2003, doi: 10.1093/hmg/ddg051.

[457] V. Lukacs *et al.*, "Impaired PIEZO1 function in patients with a novel autosomal recessive congenital lymphatic dysplasia," *Nat. Commun.*, vol. 6, no. 1, pp. 1–7, Sep. 2015, doi: 10.1038/ncomms9329.

[458] R. Zarychanski *et al.*, "Mutations in the mechanotransduction protein PIEZO1 are associated with hereditary xerocytosis," *Blood*, vol. 120, no. 9, pp. 1908–1915, Aug. 2012, doi: 10.1182/blood-2012-04-422253.

[459] M. Slatkin, "Gene flow and the geographic structure of natural populations," *Science (80-.).*, vol. 236, no. 4803, pp. 787–792, May 1987, doi: 10.1126/SCIENCE.3576198.

[460] O. R. Jones and J. Wang, "A comparison of four methods for detecting weak genetic structure from marker data," *Ecol. Evol.*, vol. 2, no. 5, p. 1048, May 2012, doi: 10.1002/ECE3.237.

[461] A. Coventry *et al.*, "Deep resequencing reveals excess rare recent variants consistent with explosive population growth," *Nat. Commun.*, vol. 1, no. 8, 2010, doi: 10.1038/ncomms1130.

[462] and M. W. F. Brenna M. Henn, L. L. Cavalli-Sforza, "The great human expansion," *Proc. Natl. Acad. Sci.*, vol. 109, no. 44, pp. 17758–17764, Oct. 2012.

[463] B. J. . Keats and S. L. Sherman, "Population Genetics," *Ref. Modul. Biomed. Sci.*, Jan. 2014, doi: 10.1016/B978-0-12-801238-3.05486-6.

[464] R. McQuillan *et al.*, "Runs of Homozygosity in European Populations," *Am. J. Hum. Genet.*, vol. 83, no. 3, pp. 359–372, Sep. 2008.