What a Large Scale Genetic Study Reveals About the Afrikaner Population of South Africa and its Diaspora

"Johan Oosthuizen"

BA MA "Cyrus Irani", Xander "Esterházy", Andrei Kharchenko, "Alan Hampton"

August, 2024

Telegram

PNG images of the figures and tables in this study

CONTENTS

DECLARATION OF AUTHORSHIP

This is to certify that the work presented in this research article is my own and that of the co-authors listed on the cover page.

DECLARATION OF INTERESTS

This is to certify that the work presented in this research article is independent and not under the influence of any institution(s) or alternate interest(s). This is also to certify that this research is not being funded by any institution(s), private individual(s) or corporate body(s).

FIGURES AND TABLES

GLOSSARY

cM (centimorgan): https://thednageek.com/fact-check-the-misunderstood-centimorgan/
MRCA (most recent common ancestor)
PCA (principal component analysis): PCA is a technique used to reduce a large set of data into multiple, smaller dimensions known as PCs.
PC(s) (principal component(s))
Y-DNA haplogroup: A Y-DNA haplogroup is a classification in human genetics that groups individuals based on specific mutations in the non-recombining portions of the Y chromosome, which is inherited patrilineally.
Mitochondrial DNA haplogroup: A mitochondrial DNA haplogroup is a classification in human genetics that groups individuals based on specific variations in their mitochondrial DNA, which is inherited matrilineally.
SNP (single nucleotide polymorphism)
RV(s) (risk variant(s)): an allele which is linked to an increased risk or genetic predisposition to a medical condition
EU (European), AFR (African), SEA (Southeast Asian), EA (East Asian), AJ (Ashkenazi Jewish), US (The United States of America), NE (Netherlands), GE (Germany), FR (France)
Yellow highlight in figure 1: Inconclusive due to low sample size.
Red highlight in figure 1: File corruption highly likely, best to be ignored.

ABSTRACT

Background:

Previous research has identified numerous heritable diseases that are relatively rare in other parts of the world to be more prevalent among the Afrikaner population in South Africa and its diaspora. Some studies have posited that this phenomenon may be attributable to the founder effect. Additionally, it has been suggested that Afrikaners possess genetic traits that are commonly found in European populations. However, there is a paucity of studies investigating the specific biomarkers of Afrikaners.

Genealogical analyses have estimated an average inbreeding coefficient akin to that of fourth cousins; however, a recent genetic study suggested that the level of inbreeding among Afrikaners is comparable to that of most contemporary European populations. A singular study examining the mitochondrial DNA and Y-DNA haplogroups of the Afrikaner population in South Africa and its diaspora revealed that 100% of the Y-DNA haplogroups were predominantly found in Europe, while 18% of the mitochondrial DNA haplogroups were most common outside of Europe. This was used to erroneously suggest an average non-European contribution of 9%.

Genealogical studies have estimated the non-European component of the ancestry of the Afrikaner population in South Africa and its diaspora to range from a high of 7.2% to a low of 0.5% (with estimates accounting for unknown ancestry suggesting an average as high as 7.5% and as low as 0.9%). One author proposed that such admixture is ubiquitous, while another study found that this admixture was present in only 75% of the 32 Afrikaner pedigrees examined.

Genetic studies have inferred the non-European component of the ancestry of the Afrikaner population in South Africa and its diaspora to be as high as 6.5% and as low as 0.9%. One relatively representative study found this admixture to be present in approximately 98.7% of the Afrikaners examined, while another source reported its presence in only about 57.1% of the Afrikaners it analysed.

Results:

In the analysis of 72 Afrikaners, a significant number of medical conditions examined revealed an increased risk or genetic predisposition, as well as elevated allele frequencies of pathological or high-risk variants. Conversely, in a minority of instances, a decreased risk or genetic predisposition, along with reduced allele frequencies of pathological or high-risk variants, was noted for the medical conditions assessed. These suggest founder effects. In the majority of cases, average risk or genetic predisposition, as well as average allele frequencies of pathological or high-risk variants, were observed for the medical conditions under investigation.

Regarding genetic traits, it was observed that a slight majority of the genetic traits and phenotypes among the examined Afrikaners exhibited a distribution frequency comparable to that of heterogenous European populations. Nevertheless, in numerous instances, these traits diverged to varying degrees, with some suggesting extra-European genetic input, while others indicated founder effects or some form of sexual or natural selection. In a few cases, the data did not support any of these interpretations or suggested a combination of factors.

With respect to biomarker predispositions, the analysis of the 72 Afrikaners revealed no significant anomalies, with any effects of these predispositions being entirely negligible. However, some evidence of adaptation to the South African environment was observed in a limited number of biomarkers.

An analysis of the shared centimorgans (cM) among six randomly selected genetic Afrikaners and 1,280 other genetic Afrikaners revealed an average total of approximately 48.3 shared cM, with individual averages ranging from approximately 44.3 to 51. The average largest shared segment was found to be approximately 13 cM, with a range of approximately 12.5 to 14 cM. According to the "Shared cM Project 4.0 tool v4," this data suggests a total average relationship most closely resembling that of third cousins once removed, with the typical relationship range falling between third and fourth cousins, and an average most recent common ancestor (MRCA) among less recent but still common ancestors of a fifth or sixth great-grandparent.

Among the 306 Afrikaner individuals with publicly available mitochondrial DNA haplogroups, 36 individuals (approximately 11.8%) reported carrying a haplogroup that is most prevalent outside of Europe. In contrast, of the 197 Afrikaner individuals with publicly available Y-DNA haplogroups, only one individual (approximately 0.5% of the cohort) reported carrying a haplogroup that is most common outside of Europe; however, the paternal progenitor associated with this haplogroup was known to be of European descent.

The inference of admixture proportions for the 1,388 Afrikaner individuals analysed indicated that the majority of their ancestry is European, with an average of approximately 97.7%. Furthermore, non-European admixture was identified in only around three-quarters of the individuals examined. The analysis also revealed that the average European components of Afrikaner individuals without any indications of non-European ancestry were predominantly Dutch, followed by French and German, with a minority of their ancestry derived from other European sources.

The majority of Afrikaners plot within the European cluster on a principal component analysis (PCA) scatter plot.

Conclusion:

After more than three centuries since their ethnogenesis, the Afrikaner population of South Africa and its diaspora exhibits neither stability nor uniformity. They are characterised by a high degree of endogamy; however, they remain heterogeneous, displaying a complex interaction of health and illness, purity and admixture, simultaneously. The sundry origins of their European ancestry, combined with non-European influences in certain individuals, have not effectively mitigated the consequences of their limited founder population and endogamous practices. This endogamy has conferred both advantages and disadvantages concerning heritable diseases, and in some instances, show entirely negligible impact. While the majority of their genetic traits align with those typical of European populations, there are notable divergences that reflect distinct founder effects, and in a minority of cases, evidence of potential extra-European genetic contributions. The Afrikaner population is predominantly of European descent, primarily Dutch, followed by French and German ancestry. Although a significant proportion of Afrikaners studied exhibited minor signals of non-European admixture, such occurrences were not in any manner ubiquitous.

INTRODUCTION

The Afrikaner population of South Africa and its diaspora, less commonly referred to as "The Boere" or "Boerevolk," are descendants of a relatively small group of European colonists, primarily of Dutch, French, and German origin, who settled in the Cape of Good Hope between the mid-17th century and the early 19th century. This population exhibits typically European phenotypical traits, which generally indicate a predominantly European ancestral background. Members of this ethnic group may or may not be able to communicate in Afrikaans. For the purposes of this study, a genetic Afrikaner is defined as an individual with a significant degree of genetic relatedness to other South Africans, as identified in genetic databases such as GEDmatch, and who possesses a publicly accessible genetic profile that predicts a "White passing" phenotype based on admixture analysis. Further details can be found in the "Materials and Methods" section.

The objective of this study is to achieve a comprehensive understanding of the genetic profile of the Afrikaner population. This extensive research examines a wide array of subjects related to the genetic profile of the Afrikaner population in South Africa and its diaspora, including health profiles and heritable disorders prevalent among Afrikaners, genetic traits characteristic of this group, biomarker predispositions, the intensity of endogamy (inbreeding), the geographical distribution of mitochondrial DNA and Y-DNA haplogroups and the frequency of major clades, and the patterns of European and non-European admixture, with a genetic perspective on the types of European admixture present within the population.

Understanding the genetic profile of the Afrikaner population is of paramount importance for several reasons. Firstly, it provides insights into the historical migrations and demographic events that have shaped the current genetic landscape of South Africa. Secondly, this study illuminates the genetic health predispositions specific to the Afrikaner population, which can inform medical research and public health strategies, ultimately leading to enhanced healthcare and targeted interventions for this community. Thirdly, this research contributes to the broader field of population genetics by offering a detailed case study of a unique population. The observed patterns of endogamy and admixture within the Afrikaner population can provide valuable comparisons with other populations, thereby enriching understanding of genetic variance and history. Finally, this research provides a deeper understanding of the distinct and valuable history of the Afrikaner people, fostering a greater appreciation for their heritage and ethnocultural identity.

BACKGROUND

Health profiles, heritable disorders, genetic traits, and biomarkers:

Previous studies by various authors, have identified many heritable diseases, rare elsewhere in the world, to be much more common among the Afrikaner population of South Africa and its diaspora, including but not limited to: familial hypercholesterolaemia, progressive familial heart block, Huntington's chorea, porphyria variegata, Gaucher's disease, cystic fibrosis, familial colonic polyposis, Fanconi's anemia, long QT syndrome, osteogenesis imperfecta, pseudoxanthoma elasticum, and sclerosteosis (Botha & Beighton. 1983), (Rosendorffetal. 1987), (Brink et al. 2005), (Wallace. 2022) (Founders Effect amongst the South African Population. Geni.com. No date). It has been suggested, based on genealogical information, that this is a result of the founder effect (Botha & Beighton. 1983). For example, it was found that a single founding couple was common to 40 out of 48 Afrikaner families with Parkinson's disease, suggesting a strong founder effect (Gelndenhuys et al. 2014). Similarly, the founder Willem Schalk van der Merwe and his wife Elsje Cloete have been credited with introducing at least four diseases to the Afrikaner: Huntington's chorea, pseudoxanthoma elasticum, lipoid proteinosis and schizophrenia (Founders Effect amongst the South African Population. Geni.com. No date). Some of the various genetic traits of the Afrikaner population have also been examined, however this has been limited to lactose persistence and the warrior versus worrier phenotype in MAOA, finding, respectively, an average of 83% (Greef et al. 2020) and a phenotype similar to other European populations (Erasmus et al. 2015). Little research has been done regarding the biomarkers of the Afrikaners, however, one study (Sifunda et al. 2023) did find a mean HBa1c of 5.64% in White South Africans (most of whom are presumably Afrikaners). [1], [2], [3], [4], [5], [6], [7], [8], [9]

Endogamy:

The intensity and presence of endogamy and inbreeding is an interesting aspect of the Afrikaner population in South Africa and its diaspora. Within the Afrikaner/Boer culture, it is seen as a token of respect for juniors to address their seniors as "oom" (uncle) or "tannie" (auntie), regardless of known familial connections. The custom of referring to seniors as "oom" or "tannie" is not only done out of respect but it also reflects the reality of the close familial bond and the small collective genetic pool. Genealogical records suggest that between 1657 and 1807, approximately 1,600 male Afrikaner progenitors, known as "stamvaders," arrived at the Cape of Good Hope (Colenbrander, 1902), making a total of around 3,200 when including females. However, this figure is somewhat misleading, as many of the female progenitors, especially the partners of later male settlers, were born into the proto-Afrikaner community at the Cape, reducing the number of unique progenitors to approximately 2,250 (Colenbrander, 1902). This figure is still deceptive due to the varying degrees of genetic contribution from these progenitors, depending on factors such as the number of offspring and the timing of their arrival. By taking into account the number of children each founding couple had and considering only those born outside the Cape as unique progenitors, while also scaling the intensity of their genetic contribution based on their arrival time, the effective genetic founding population can be estimated at approximately 440 individuals arriving simultaneously, a figure comparable to the estimated 500 founders of the modern Ashkenazi Jewish population. This represents a significant genetic bottleneck, which is consistent with findings from a

2012 study that analysed the pedigrees of 32 influential Afrikaners. Using J. A. Heese's genealogical fantasies and dubitable findings, "Die Herkoms Van Die Afrikaner" (1971), this study calculated an average inbreeding coefficient of 0.005102, indicating a common relationship akin to that of fourth cousins (Philpott, 2012). More recently, a 2020 study by Greef et al. investigated the genetic dimensions of inbreeding within the Afrikaner population, controversially asserting that "Afrikaners today have comparable inbreeding levels to current-day European populations." This claim was made based on calculations of runs of homozygosity for each of the 77 individuals in the study. [9], [10], [11], [12]

Haplogroup distribution and frequency:

Research on the haplogroups of the Afrikaner population in South Africa and its diaspora remains limited. Naive estimates, derived by doubling J.A. Heese's genealogical estimate of the total non-European contribution, suggested that the contribution of haplogroups most common outside of Europe to Afrikaner mitochondrial DNA haplogroups might be approximately 12%-14%, with the contribution to Y-DNA haplogroups estimated at nearly 0%. In a 2013 study, C. Erasmus reported that 18% of the mtDNA haplogroups in a sample of 185 Afrikaners were predominantly found outside Europe, while 0% of the Y-DNA haplogroups were primarily found outside of Europe. These findings were used by the author’s publishers, the GGSA (Genealogical Society of South Africa) to imply an average non-European admixture fraction 1.5 times higher than Heese's earlier estimate, (which was cited as 6%, therfore, 9%). [12], [13]

Patterns of European and non-European admixture:

The patterns of European and non-European admixture, and the types of European admixture in the Afrikaner population of South Africa and its diaspora, or simply put, their ancestry, is the most controversial of all the topics that this large scale genetic study seeks to investigate, and has been debated and investigated for nearly two centuries. It therefore deserves the longest background explanation. It is also the topic with the greatest range of opinions, findings, and the least amount of historical and present consensus, despite the perception of some.

The seventeenth century colonisation of Southern Africa resulted in the settlement of two distinct groups: Europeans and (mostly non-European) slaves. These two groups admixed within their groups, with each other and with local African populations to varied extents resulting in the creation of two distinct populations: Afrikaners and Coloureds. The earliest attempt to estimate the intensity and presence of European and non-European admixture in the Afrikaners was undertaken by G.M. Theal in 1897, based on the genealogical register of C.C. De Villiers (1894). He found that the Afrikaners were approximately two thirds Dutch and one sixth French, with the final sixth being equally divided between (mostly) the Germans and other (European) nationalities. He gave no explanation as to how he calculated these figures, and he only considered the paternal progenitors. One of the other earliest studies regarding this topic was undertaken by the Dutch historian Dr. H.T. Colenbrander in his book "De Afkomst der Boeren" (1902), which studied the period between 1657 and 1807. He based his genealogical findings on various sources available to him at the time, namely C.C De Villiers genealogical register "Geslacht-Register der oude Kaapsche familiën" (1894), the Cape muster-rolls, marriage records, various church records and the other genealogical and historical work of G.M. Theal. In order to calculate the average Afrikaner’s ancestry, he split the period of study into five thirty year blocks, and multiplied the children of each nationality by 16, 8, 4, 2 and 1 for each period. Based on his raw findings, he calculated the average

Afrikaner's ancestry to be approximately 48.6% Dutch, 25.8% German, 17.2% French, 6.1% other European nationalities, 1.8% unknown, and 0.5% non-European. He then adjusted his calculations to account for the unknown portion of their ancestry, assuming 1/4 to be pre-existing cape-born proto-Afrikaner, 1/4 to be slave or non-European, 3/8 to be Dutch and 1/8th to be German, and he then added the ethnically indistinct "other nationalities" such as the German or French Swiss to their respective groups, to get a final average of approximately 50.6% Dutch, 26.5% German, 17.3% French, 4.7% other European nationalities and 0.9% non-European. Following this, there were no significant genealogical studies for nearly seven decades. This silence was shattered by the release of J.A. Heese's book "Die Herkoms van die Afrikaner" (1971) which studied the period between 1657 and 1867. Based on his genealogical fantasies and dubitable findings, he estimated the average Afrikaner's ancestry to be 35.5% Dutch, 34.4% German, 13.9% French, 7.2% non-European, 2.6% British, 2.8% other European and 3.6% unknown. This was scaled to account for unknown ancestry, giving an estimate of: 36.8% Dutch, 35.7% German, 14.4% French, 2.9% other European, 2.7% British, and 7.5% non-European. He calculated these figures in the same way as Dr. H.T. Colenbrander, however, he split the period into six blocks, rather than five. He used the same sources as Dr. H.T. Colenbrander (including his work). Despite using the same sources, the results differed, as Heese came up with different theories and inferences in regards to the origins of various progenitors (mostly those of female sex) compared to Colenbrander, mostly due to his misunderstanding of the implications of the toponym "van die Kaap/van Cabo (de Goede Hoop)", assuming it to always be a reference to slave or non-European ancestry. This unfounded falsehood still persists in many genealogical circles and publications today. His genealogical fantasies and dubitable findings in regards to the origins of many progenitors forms the basis for the vast majority of present day South African genealogical research, and national myth (Krotoa-Eva). 5 years after the release of "Die Herkoms van die Afrikaner", G.F.C De Bruyn reevaluated Heese's estimations using a more robust calculation method in his article "Die samestelling van die Afrikaner Volk" (1976) published in the "Tydskrif vir Geesteswetenskappe". His refined estimate of the average Afrikaner's ancestry was: 34.1% Dutch, 29.2% German, 24.7% French, 5.4% non-European, 3.9% unknown, 2.7% other European, and 0.3% British. Following this, another period of relative silence ensued in this area of study, which was broken in 2007 by J.M. Greef conducting a study on himself. Using the genealogical fantasies and dubitable findings of Heese, and scaling for unknown ancestors or incomplete parts of the pedigree, he found himself to be 37.5% Dutch, 27.4% German, 26.4% French, 0.8% other European, 1.9% British, and 6% non-European, with a very significant portion of his non-European ancestry being recent. Following this, in 2012 a genealogical analysis of the pedigrees of 32 influential Afrikaners was undertaken by D.M. Philpott. Basing his research of each individual's pedigree on the genealogical fantasies and dubitable findings of J.A. Heese, he estimated an average ancestry of 32.6% French, 31.4% Dutch, 23.5% German, 4.8% other European, 2.2% non-European including "van die Kaap/van Cabo (de Goede Hoop)" individuals, 1.6% British and 3.9% unknown/incomplete. Scaled to account for the unknown or incomplete parts of pedigrees: 33.9% French, 32.7% Dutch, 24.5% German, 5% other European, 2.3% non-European including "van die Kaap/van Cabo (de Goede Hoop)" individuals and 1.6% British. He also found that 25% (8/32) of the individuals examined had no non-European ancestors. The median amount of adjusted non-European ancestry was 1.6%. The author repeatedly cited his anti-racist beliefs as motivations for the study. [10], [11], [12], [14], [15], [16], [17], [18], [19]

In 1972, the first genetic study of the Afrikaner population was carried out. Working with a number of blood group gene frequencies, Botha & Pritchard estimated that between 6–7% admixture between Europeans and slaves from Africa and the East, and/or Khoikhoi, would be required to explain the allele frequencies. In 2017, R. Khan of the Gene Expression blog obtained 4 Afrikaner

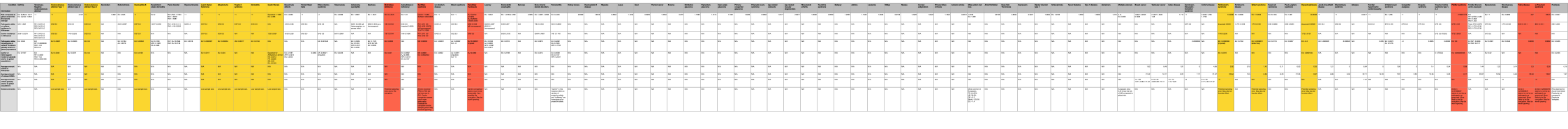
genomes through various donations and crawls through FTDNA. He run unsupervised admixture analysis on them using 5 populations, Gujarati Indians, Dai Chinese, Khoisan South Africans, Esan Nigerians, and an unspecified European population. He found that these 4 individuals had an average ancestry of 94% European, 1% East Asian, 1% Khoisan, 1% non-Khoisan African, and 2% South Asian. It is likely that he made an input error or there was a rounding artefact, as this doesn't add up to 100%. In 2018, he received 8 more genomes, making a total of 12, and repeated his experiment. This time, he found an average of 93.5% European. In the February of 2020, J. M. Greef et al published the first and only somewhat large scale genetic study of the Afrikaner population of South Africa. They found an average of 95.3% European, 1.7% South Asian, 1.3% Khoisan, 1% East Asian and Amerindian and 0.8% West/East African. They also found such non-European admixture to be present in 98.7% of all individuals analysed. 3 months after this, an online data source conducted a small scale genetic study of the Afrikaner population of South Africa and other colonial populations using 23&me results. The sample size was 7, and an average of 2.6% SSA and 0.3% EA & NA was found in the samples investigated. 3 years later in the August of 2023, another online data source uploaded a video about a genetic study conducted on the author and 13 other Afrikaners using MyHeritage v0.95 admixture estimation. An average of 99.1% European was observed, with only 57.1% of Afrikaners examined exhibiting any non-European admixture. Another anonymous online data source who has yet to be identified analysed 22 Afrikaner genomes from a bronze age, neolithic and mesolithic perspective. They found an average ancestral contribution of 47.5% Western Steppe Cattle Herder, 38.5% Anatolian Farmer, 12.9% Western Hunter Gatherer, 0.4% Ancestral South Indian, 0.4% Southeast Asian and 0.3% African. They also found 7/22 (~31.8%) of the samples they analysed to have no non-European admixture. [9], [13], [20], [21], [22], [23], [24]

RESULTS

Health Profiles and Heritable Disorders

The genomic information of 72 Afrikaners was analysed to identify health profiles and heritable diseases prevalent within the Afrikaner population of South Africa and its diaspora. Figure 1 presents information regarding the prevalence of pathogenic risk variants associated with each disease or medical condition analysed. Carrier prevalence is reported for autosomal recessive diseases or in instances where it is necessary to specify heterozygotes. The figure also includes the average number of risk variants (RVs) identified in each file for each disease, the average number of relevant single nucleotide polymorphisms (SNPs) in each file for each disease, and a comparison of incidence rates of the observed pathologies or pathogenic risk variants across different populations. Additional information, notes, or comments pertaining to the results will be provided in the Notes/ Comments section of the table. It is important to note that in some instances, no relevant SNPs are identified for a specific disease in a submitted file; consequently, certain incidence rates will not be out of 72. Common risk variants are not reported. Columns highlighted red ought to be ignored and could be the result of file corruption. Columns highlighted yellow are not conclusive due to low sample size.

Figure 1



Genetic Traits

The genomic information of 72 Afrikaners was analysed to identify genetic traits that are prevalent within the Afrikaner population of South Africa and its diaspora. Certain traits are predominantly polygenic, while others are primarily monogenic. For predominantly polygenic traits (Figure 2a), only the predicted phenotypes for each trait are reported, and comparisons are drawn with other populations, contingent upon the availability of relevant data. In the case of predominantly monogenic traits (Figure 2b), the allele frequencies are reported, and comparisons with other populations are also conducted, provided that relevant data is available.

Figure 2a

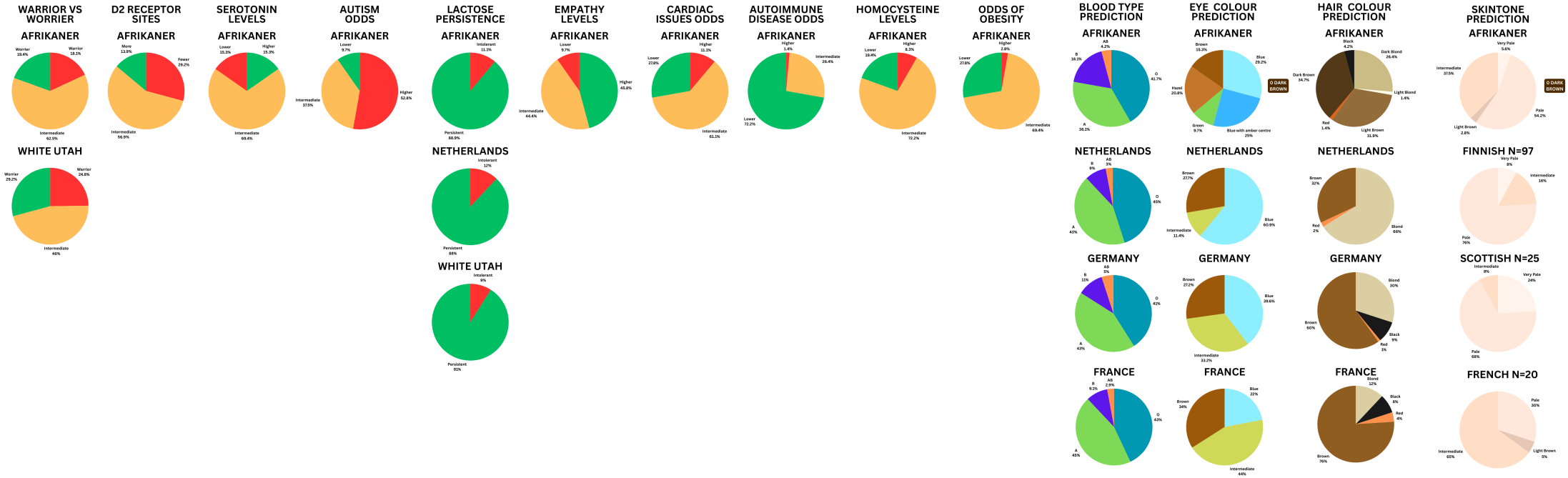
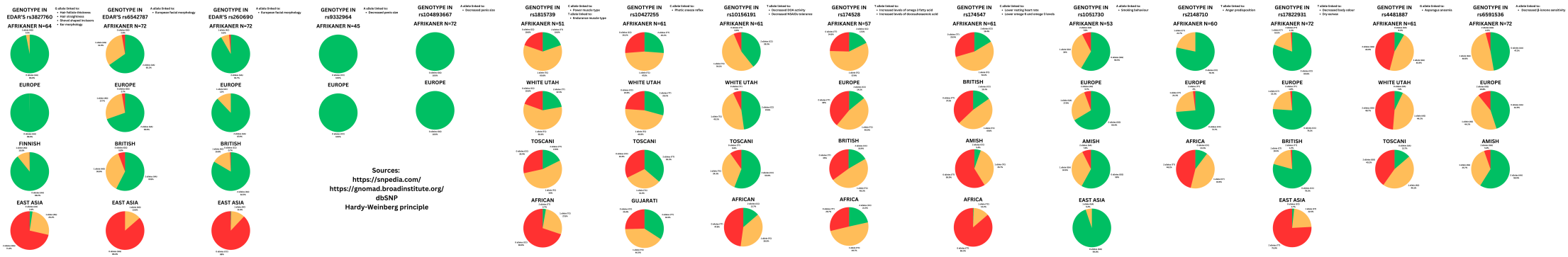

WARRIOR VS WORRIER
AFRIKANER
Worrier 10.4%
Warrior 18.1%
Intermediate 62.5%
WHITE UTAH
Worrier 29.2%
Warrior 24.8%
Intermediate 46%
D2 RECEPTOR SITES
AFRIKANER
More 13.9%
Fewer 29.2%
Intermediate 56.9%
SEROTONIN LEVELS
AFRIKANER
Lower 15.3%
Higher 15.3%
Intermediate 69.4%
AUTISM ODDS
AFRIKANER
Lower 9.7%
Higher 52.8%
Intermediate 37.5%
LACTOSE PERSISTENCE
AFRIKANER
Intolerant 11.1%
Persistent 88.9%
NETHERLANDS
Intolerant 12%
Persistent 88%
WHITE UTAH
Intolerant 9%
Persistent 91%
EMPATHY LEVELS
AFRIKANER
Lower 9.7%
Higher 45.8%
Intermediate 44.4%
CARDIAC ISSUES ODDS
AFRIKANER
Lower 27.8%
Higher 11.1%
Intermediate 61.1%
AUTOIMMUNE DISEASE ODDS
AFRIKANER
Higher 1.4%
Intermediate 26.4%
Lower 72.2%
HOMOCYSTEINE LEVELS
AFRIKANER
Lower 19.4%
Higher 8.3%
Intermediate 72.2%
ODDS OF OBESITY
AFRIKANER
Higher 2.8%
Lower 27.8%
Intermediate 69.4%
BLOOD TYPE PREDICTION
AFRIKANER
AB 4.2%
B 16.3%
O 41.7%
A 36.1%
NETHERLANDS
B 9%
AB 3%
O 45%
A 43%
GERMANY
B 13%
AB 5%
O 42%
A 43%
FRANCE
B 9.2%
AB 2.9%
O 43%
A 45%
EYE COLOUR PREDICTION
AFRIKANER
Brown 15.3%
Blue 29.2%
Hazel 20.8%
Green 9.7%
Blue with amber centre 25%
O DARK BROWN
NETHERLANDS
Brown 27.7%
Blue 60.9%
Intermediate 11.4%
GERMANY
Brown 27.2%
Blue 39.6%
Intermediate 33.2%
FRANCE
Blue 22%
Brown 34%
Intermediate 44%
HAIR COLOUR PREDICTION
AFRIKANER
Black 4.2%
Dark Blond 26.4%
Dark Brown 34.7%
Light Blond 1.4%
Red 1.4%
Light Brown 31.9%
NETHERLANDS
Brown 32%
Blond 66%
Red 2%
GERMANY
Blond 30%
Brown 60%
Black 9%
Red 1%
FRANCE
Blond 12%
Black 8%
Red 4%
Brown 76%
SKINTONE PREDICTION
AFRIKANER
Very Pale 5.6%
Intermediate 37.5%
Pale 54.2%
Light Brown 2.8%
O DARK BROWN
FINNISH N=97
Very Pale 3%
Intermediate 16%
Pale 76%
SCOTTISH N=25
Intermediate 8%
Very Pale 24%
Pale 68%
FRENCH N=20
Pale 30%
Light Brown 5%
Intermediate 65%

Figure 2b


GENOTYPE IN EDAR'S rs3827760
AFRIKANER N=64
EUROPE
FINNISH
EAST ASIA
GENOTYPE IN EDAR'S rs6542787
AFRIKANER N=72
EUROPE
BRITISH
EAST ASIA
GENOTYPE IN EDAR'S rs260690
AFRIKANER N=72
EUROPE
BRITISH
EAST ASIA
GENOTYPE IN rs9332964
AFRIKANER N=45
EUROPE
GENOTYPE IN rs104893667
AFRIKANER N=72
EUROPE
Sources:
https://snpedia.com/
https://gnomad.broadinstitute.org/
dbSNP
Hardy-Weinberg principle
GENOTYPE IN rs1815739
AFRIKANER N=51
WHITE UTAH
TOSCANI
AFRICAN
GENOTYPE IN rs10427255
AFRIKANER N=61
WHITE UTAH
TOSCANI
GUJARATI
GENOTYPE IN rs10156191
AFRIKANER N=61
WHITE UTAH
TOSCANI
AFRICAN
GENOTYPE IN rs174528
AFRIKANER N=57
EUROPE
BRITISH
AFRICA
GENOTYPE IN rs174547
AFRIKANER N=61
BRITISH
AMISH
AFRICA
GENOTYPE IN rs1051730
AFRIKANER N=53
EUROPE
AMISH
EAST ASIA
GENOTYPE IN rs2148710
AFRIKANER N=60
EUROPE
AFRICA
GENOTYPE IN rs17822931
AFRIKANER N=72
EUROPE
BRITISH
EAST ASIA
GENOTYPE IN rs4481887
AFRIKANER N=61
WHITE UTAH
TOSCANI
GENOTYPE IN rs6591536
AFRIKANER N=72
EUROPE
AMISH

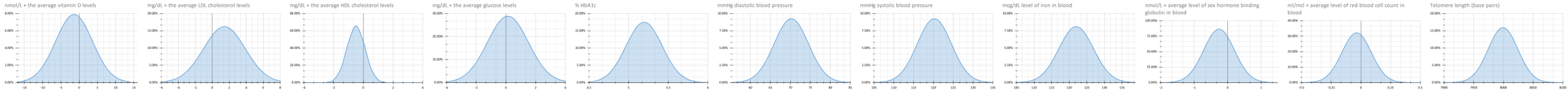
Biomarkers Predispostions

The genomic information of 72 Afrikaners was analysed to identify biomarker predisposition patterns within the Afrikaner population of South Africa and its diaspora. Figure 3a presents information regarding the relationship of each biomarker to an average value (indicated by a +/-) or the quantity of each biomarker, expressed to four decimal places, along with a commentary on its potential effects, if applicable. Figure 3b features a normal distribution graph for each biomarker.

Figure 3a

Biomarker	Vitamin D levels	LDL cholesterol	HDL cholesterol	Glucose levels	% HbA1c	Diastolic blood pressure	Systolic blood pressure	Level of iron in blood	Level of sex hormone binding globulin in blood	Level of red blood cell count in blood	Telomere length (base pairs)
Comparison to average or amount of biomarker	-1.3945 nmol/L	+1.4294 mg/dL	-0.4925 mg/dL	+0.138 mg/dL	GER & AUS: -2.6037 5.1963	70.2349 mmHg	120.2349 mmHg	121.9619 mcg/dL	-0.2584 nmol/L	-0.0359 ml/mcl	7999.243
Effect	Negligible	Minor negative	Negligible	Negligible	Minor positive	Negligible	Negligible	Negligible	Negligible	Negligible	Negligible

Figure 3b


nmol/L + the average vitamin D levels
mg/dL + the average LDL cholesterol levels
mg/dL + the average HDL cholesterol levels
mg/dL + the average glucose levels
% HbA1c
mmHg diastolic blood pressure
mmHg systolic blood pressure
mcg/dL level of iron in blood
nmol/L + average level of sex hormone binding globulin in blood
ml/mcl + average level of red blood cell count in blood
Telomere length (base pairs)

Endogamy

The total amount of centimorgans (cM) and the largest segment of cM shared between 6 random genetic Afrikaners and 1280 other genetic Afrikaners was examined for the purposes of identifying the presence and intensity of endogamy in the Afrikaner population of South Africa and its diaspora. Negligible variance was observed between the average total amount of cM shared and the largest shared segment of cM shared with other genetic Afrikaners, for each of the six source individuals. The average total amount of shared cM was ~48.3 (range of the averages between the six: ~44.3-~51) and the average largest shared segment of cM was ~13 (range ~12.5-~14). The modal total amount of shared cM was between 30-30.9 cM, and the modal largest shared segment was between 10-10.9 cM. The median for both was 43.1 cM and 11.3 cM respectively.

According to "The Shared cM Project 4.0 tool v4", ~48.3 cM is equivalent to an average relationship most similar to that of third cousins once removed, with the typical range falling somewhere between a third or fourth cousin relationship. According to "The Shared cM Project 4.0 tool v4", a largest shared segment of ~13 cM is equivalent to an average MRCA (most recent common ancestor), among less recent but still common ancestors, of a fifth or sixth great grandparent.

The adjusted standard deviation across five of the six source individuals examined was ~28.3 for the total amount of cM they shared with the 1280 other genetic Afrikaners, and ~7.2 for the largest segment of cM they shared with the 1280 other genetic Afrikaners. The 3rd source individual had a skewed SD due to having close family in their matches, so they were not included in the SD calculations, however, they were still included in the average shared cM figures, as these close relationships were not enough to cause any significant effect on the average.

Figure 4a includes a normal distribution graph incorporating an average of each individual’s results for both the total shared cM and the amount of cM shared on the largest shared segment of DNA. Figure 4b includes the same information, but separated for each random source individual’s results. Figure 4c includes a frequency bar graph incorporating each individual’s results for both the total shared cM and the amount of cM shared on the largest shared segment of DNA.

Figure 4a

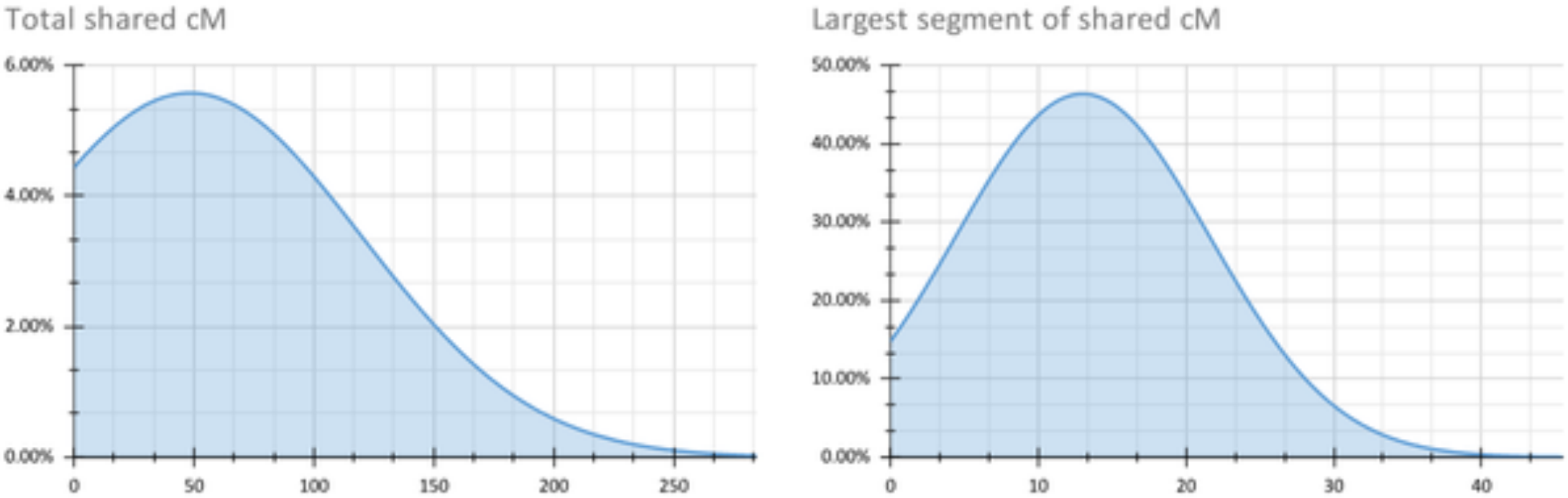
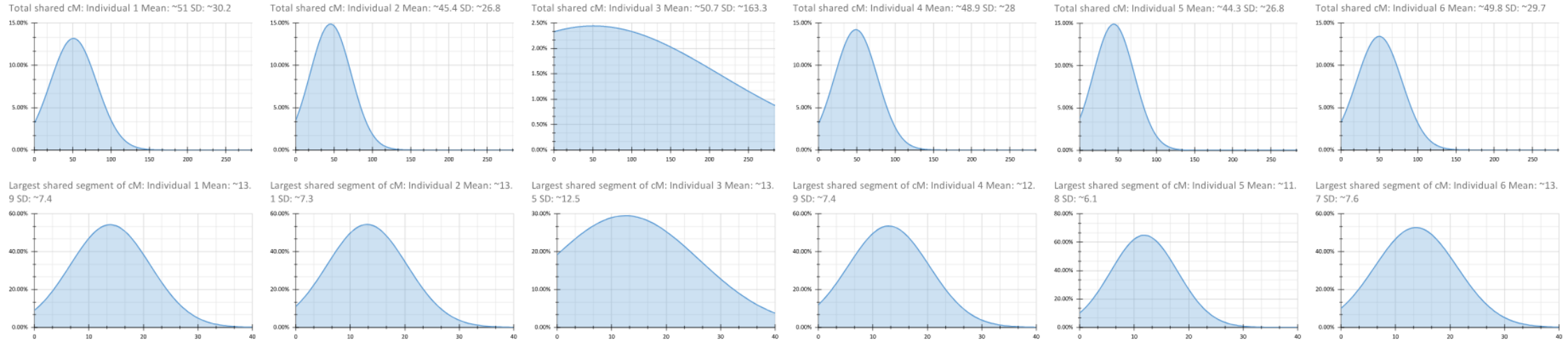

Total shared cM
6.00%
4.00%
2.00%
0.00%
0
50
100
150
200
250
Largest segment of shared cM
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
0
10
20
30
40

Figure 4b


Total shared cM: Individual 1 Mean: ~51 SD: ~30.2
Total shared cM: Individual 2 Mean: ~45.4 SD: ~26.8
Total shared cM: Individual 3 Mean: ~50.7 SD: ~163.3
Total shared cM: Individual 4 Mean: ~48.9 SD: ~28
Total shared cM: Individual 5 Mean: ~44.3 SD: ~26.8
Total shared cM: Individual 6 Mean: ~49.8 SD: ~29.7
Largest shared segment of cM: Individual 1 Mean: ~13.9 SD: ~7.4
Largest shared segment of cM: Individual 2 Mean: ~13.1 SD: ~7.3
Largest shared segment of cM: Individual 3 Mean: ~13.5 SD: ~12.5
Largest shared segment of cM: Individual 4 Mean: ~12.9 SD: ~7.4
Largest shared segment of cM: Individual 5 Mean: ~11.8 SD: ~6.1
Largest shared segment of cM: Individual 6 Mean: ~13.7 SD: ~7.6

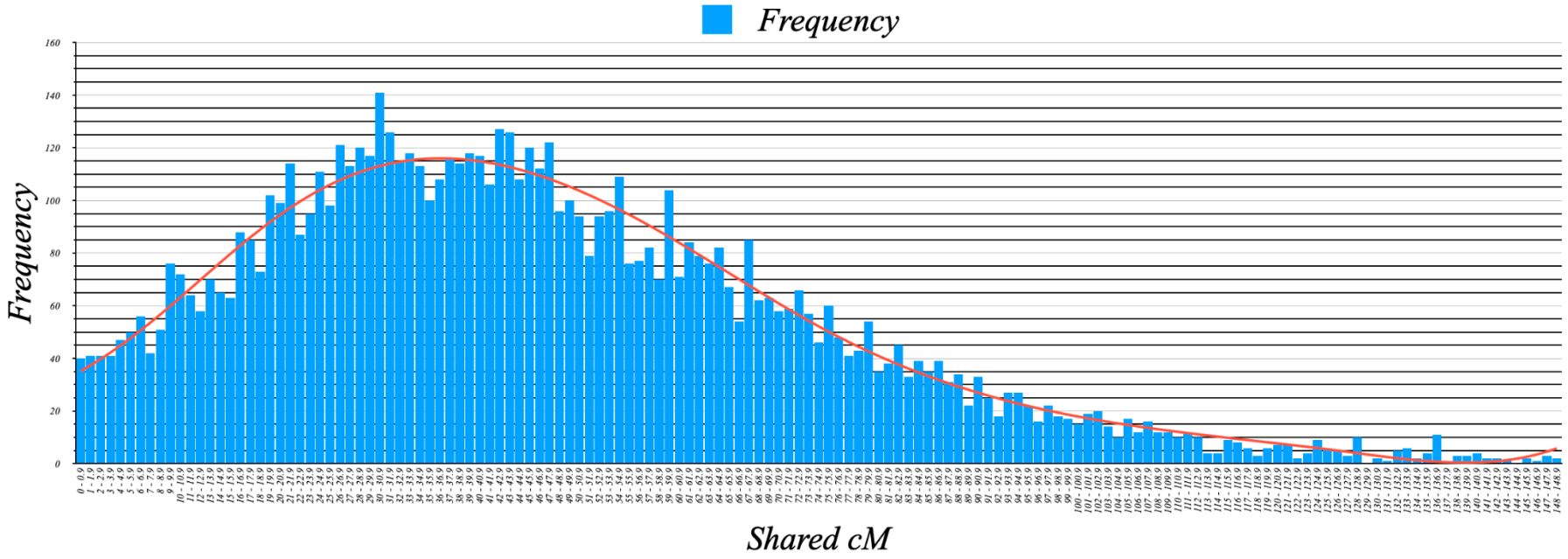
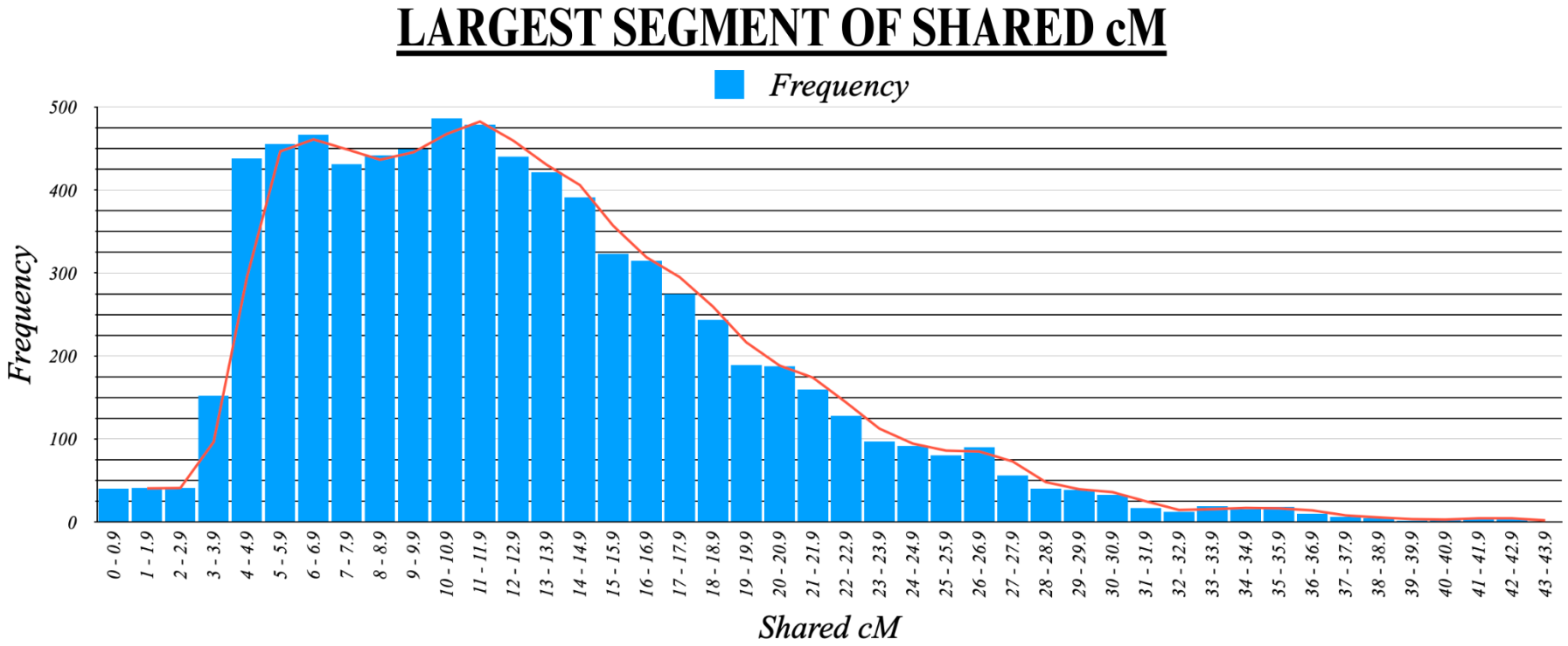

TOTAL SHARED cM
Frequency
Frequency
Shared cM

Figure 4c


LARGEST SEGMENT OF SHARED cM
Frequency
Frequency
0
100
200
300
400
500
0 - 0.9
1 - 1.9
2 - 2.9
3 - 3.9
4 - 4.9
5 - 5.9
6 - 6.9
7 - 7.9
8 - 8.9
9 - 9.9
10 - 10.9
11 - 11.9
12 - 12.9
13 - 13.9
14 - 14.9
15 - 15.9
16 - 16.9
17 - 17.9
18 - 18.9
19 - 19.9
20 - 20.9
21 - 21.9
22 - 22.9
23 - 23.9
24 - 24.9
25 - 25.9
26 - 26.9
27 - 27.9
28 - 28.9
29 - 29.9
30 - 30.9
31 - 31.9
32 - 32.9
33 - 33.9
34 - 34.9
35 - 35.9
36 - 36.9
37 - 37.9
38 - 38.9
39 - 39.9
40 - 40.9
41 - 41.9
42 - 42.9
43 - 43.9
Shared cM

Haplogroups

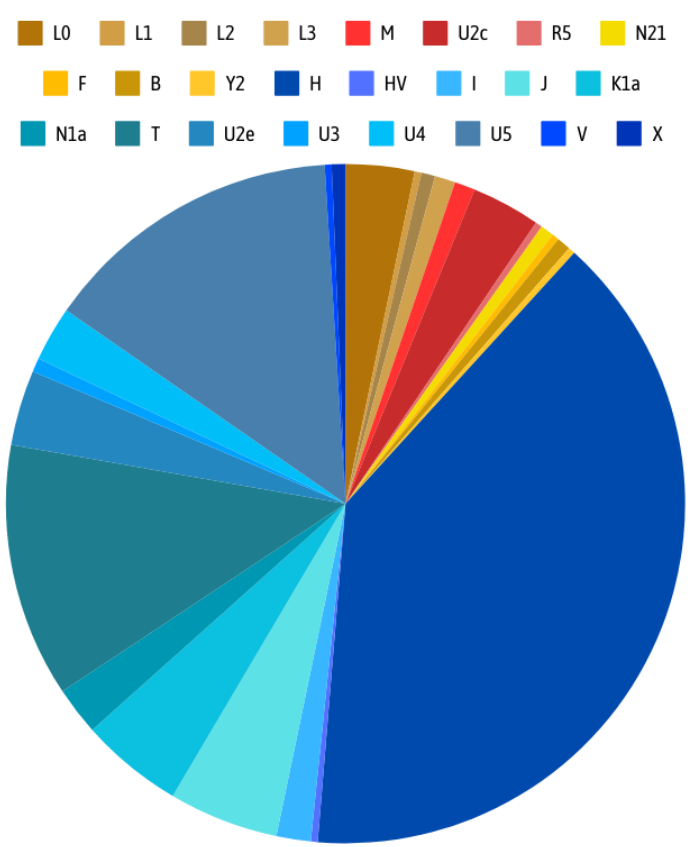
The Y-DNA and mitochondrial DNA haplogroup information were obtained from genetically confirmed Afrikaner individuals who had publicly shared their haplogroup information on GEDmatch. This data collection was undertaken to estimate the prevalence and geographical distribution of mitochondrial DNA and Y-DNA haplogroups within the Afrikaner population of South Africa and its diaspora. The data scraping process resulted in 197 unique Afrikaner samples with publicly available Y-DNA haplogroups and 306 unique samples with publicly available mitochondrial DNA haplogroups.

The greatest diversity was observed in the mitochondrial DNA haplogroups of the genetic Afrikaners analysed, with approximately ~11.8% (36) of the haplogroups observed being most common outside of Europe (Africa: ~5.2%, South Asia: ~4.6%, Southeast Asia & Austronesia: 2%), with a great variety of haplogroups most common inside Europe, which made up the remaining ~88.2% (270) of haplogroups observed. The Y-DNA haplogroups observed were much less diverse, with only ~0.5% (1) of the haplogroups observed being most common outside of Europe (Southeast Asia & Austronesia: ~0.5%), with small variance observed in the remaining ~99.5% (196) haplogroups, which were most common in Europe.

Figure 5a includes a pie chart and table including the distribution of and areas of highest frequency for the Mitochondrial DNA haplogroups in the representative sample cohort of genetic Afrikaners who had made their haplogroup public. Figure 5b includes a pie chart and table including the distribution of and areas of highest frequency for the Y-DNA haplogroups in the representative cohort of genetic Afrikaners who had made their haplogroup public.

Figure 5a

- L0: ~3.3%
- L1: ~0.3%
- L2: ~0.7%
- L3: ~1%
- M: ~1%
- U2c: ~3.3%
- R5: ~0.3%
- N21: ~0.7%
- F: ~0.3%
- B: ~0.7%
- Y2: ~0.3%
- H: ~39.5%
- HV: ~0.3%
- I: ~1.6%
- J: ~5.2%
- K1a: ~4.9%
- N1a: ~2.3%
- T: 12.1%
- U2e: ~3.6%
- U3: ~0.7%
- U4: ~2.6%
- U5: ~14.4%
- V: 0.3%
- X: 0.7%

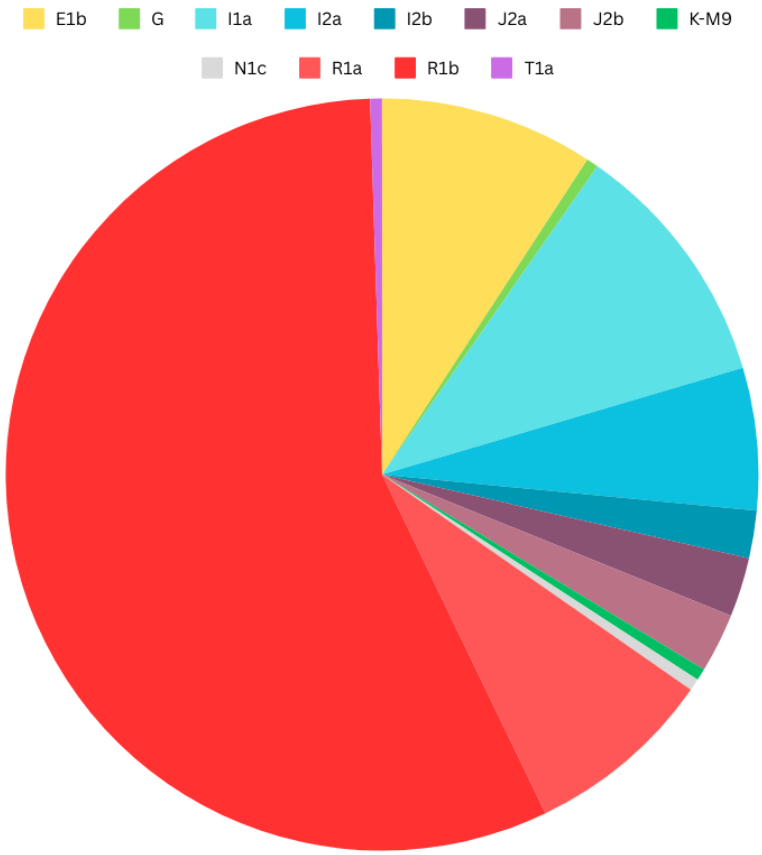


Haplogroups most common in:
Africa: ~5.2%
South Asia: ~4.6%
Southeast Asia & Austronesia: ~2%
Europe: ~88.2%

N=306

Figure 5b

- E1b: ~9.1%
- G: ~0.5%
- I1a: ~10.7%
- I2a: ~6.1%
- I2b: ~2%
- J2a: ~2.5%
- K-M9: ~0.5%
- N1c: ~0.5%
- R1a: ~8.1%
- R1b: ~56.3%
- T1a: ~0.5%



Haplogroups most common in:
Southeast Asia & Austronesia: ~0.5%
Europe: ~99.5%

N=197

Admixture

The individual admixture proportions of 1,388 genetically identified Afrikaners were inferred in order to estimate the patterns of European and non-European admixture. The average admixture proportions were determined as follows (with values rounded to three decimal places in parentheses): 97.7% (97.75%) from European populations, 0.7% (0.68%) from South Asian populations, 0.6% (0.62%) from West African populations, 0.5% (0.45%) from Southern African populations, 0.3% (0.29%) from East and Southeast Asian populations, 0.2% (0.18%) from East African populations, and 0% (0.03%) from Oceanian populations. The standard deviation was approximately 3.5%. The median level of European admixture was 98.9%.

Regarding the distribution of European ancestry, approximately 25.8% (358 individuals) exhibited no detectable non-European admixture. About 23.1% (321 individuals) had a European admixture fraction between 99.9% and 99%, 16.9% (235 individuals) between 98.9% and 98%, 11% (152 individuals) between 97.9% and 97%, 6.2% (86 individuals) between 96.9% and 96%, 4.6% (64 individuals) between 95.9% and 95%, and 12.4% (172 individuals) had 94.9% or less European admixture. The individual with the lowest European admixture had an inferred European admixture fraction of 69.2%.

The most commonly observed non-European admixture component was West African, with 668 individuals showing no evidence of West African admixture. This was followed by Southern African (741 individuals), South Asian (919 individuals), East or Southeast Asian (1,061 individuals), East African (1,111 individuals), and Oceanian (1,305 individuals) admixture.

For the 358 genetically identified Afrikaners who exhibited no non-European admixture, the average proportions of their European ancestry were genetically inferred in order to estimate the patterns of their European admixture. The adjusted average proportions of their European components were as follows: ~51% Dutch, ~24.8% French (from various regions), ~19.3% German (from various regions), and ~4.9% from other European populations. When integrating these findings into the overall average, the detailed admixture proportions are as follows: ~49.8% Dutch, ~24.2% French, ~18.9% German, ~4.8% from other European populations, ~0.7% South Asian, ~0.6% West African, ~0.5% Southern African, ~0.3% East and Southeast Asian, and ~0.2% East African.

In a two-dimensional principal component analysis (PCA) chart plotting the first and second principal components (PC1 and PC2), the majority of Afrikaners cluster within the European cluster.

Figure 6a presents a stacked bar graph depicting the total European admixture fraction, starting at 65%, for each of the 1,388 genetically identified Afrikaners. Figure 6a.i sorts these individuals alphabetically by name, while Figure 6a.ii sorts them by total European admixture. Figure 6b includes a pie chart representing the imputed average admixture proportions. Figure 6c presents two PCA graphs plotting PC1 and PC2 for the 1,388 Afrikaners alongside individuals from various global population groups. Figure 6c.i focuses on European individuals, the 1,388 Afrikaners, 79 Coloured individuals, and averages from relevant non-European populations. Figure 6c.ii extends this analysis by including irrelevant non-European populations to expatiate the PCA plot.

Figure 6a.i

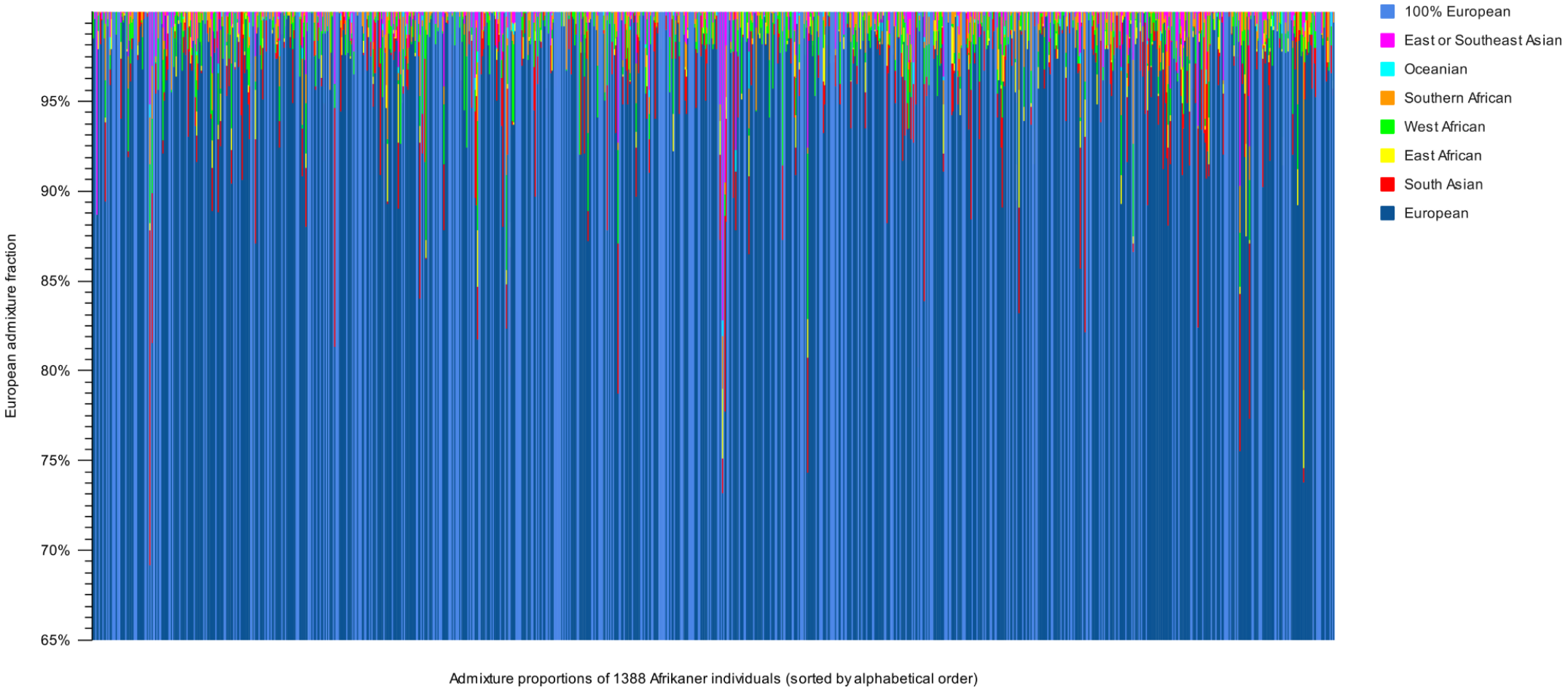

100% European
East or Southeast Asian
Oceanian
Southern African
West African
East African
South Asian
European
95%
90%
85%
80%
75%
70%
65%
European admixture fraction
Admixture proportions of 1388 Afrikaner individuals (sorted by alphabetical order)

Figure 6a.ii

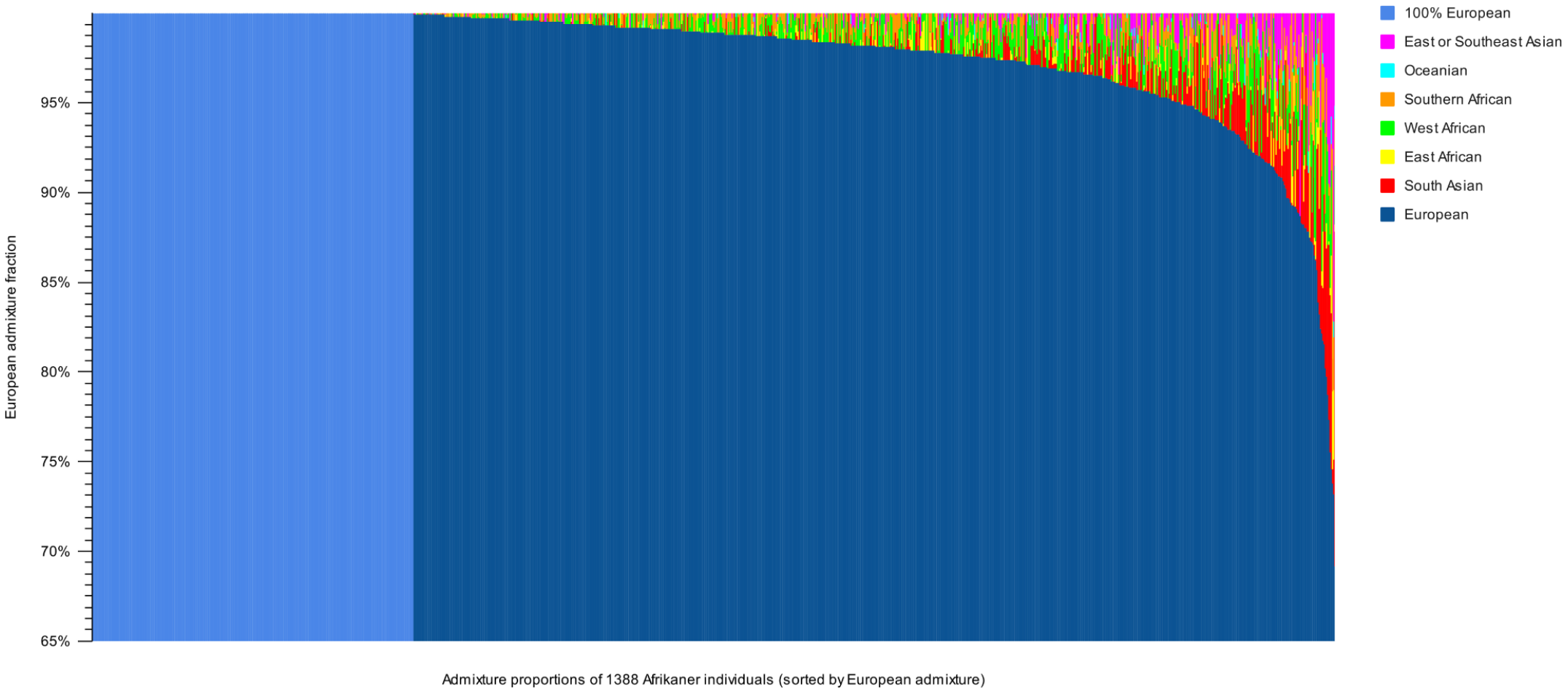

100% European
East or Southeast Asian
Oceanian
Southern African
West African
East African
South Asian
European
95%
90%
85%
80%
75%
70%
65%
European admixture fraction
Admixture proportions of 1388 Afrikaner individuals (sorted by European admixture)

Figure 6b

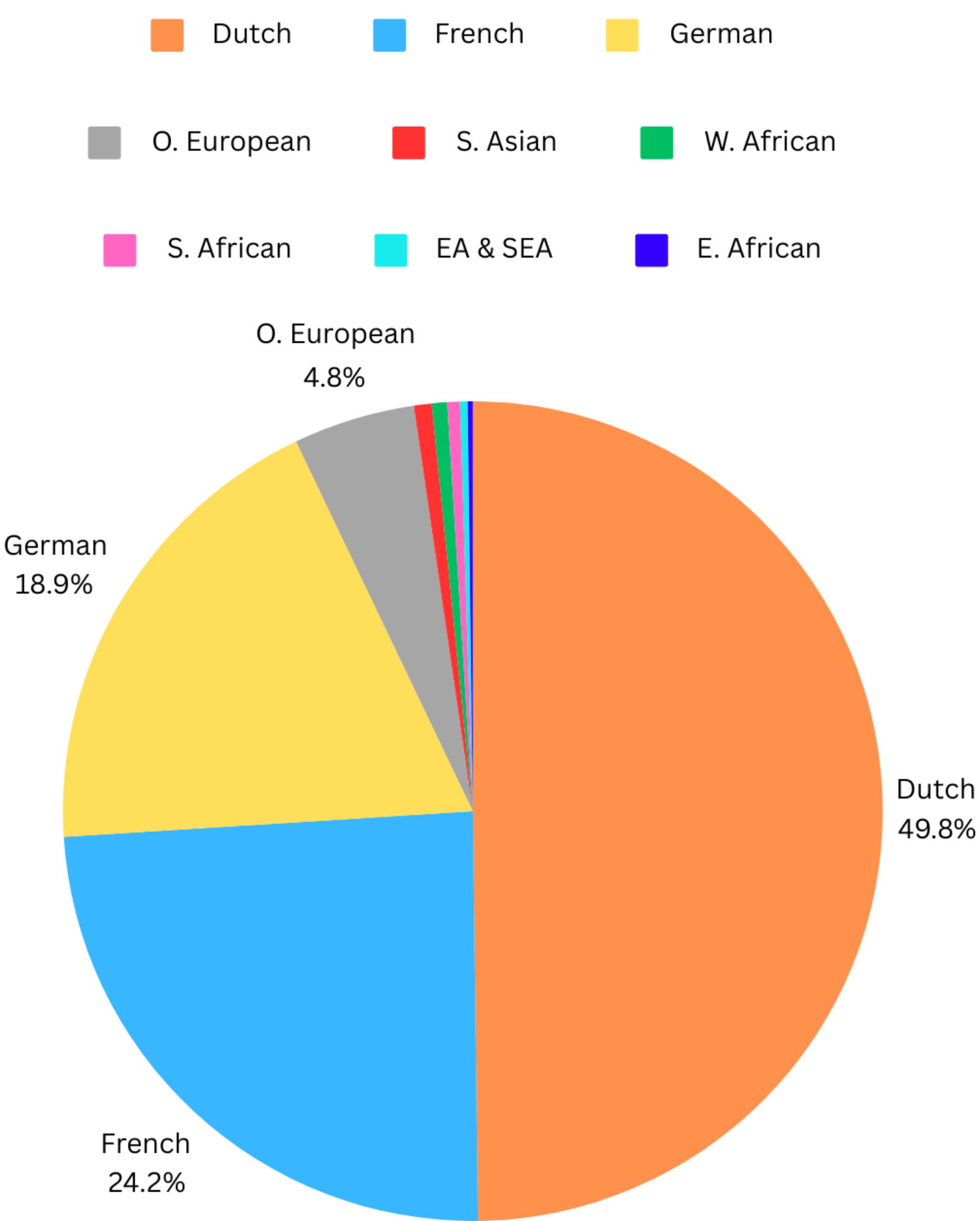
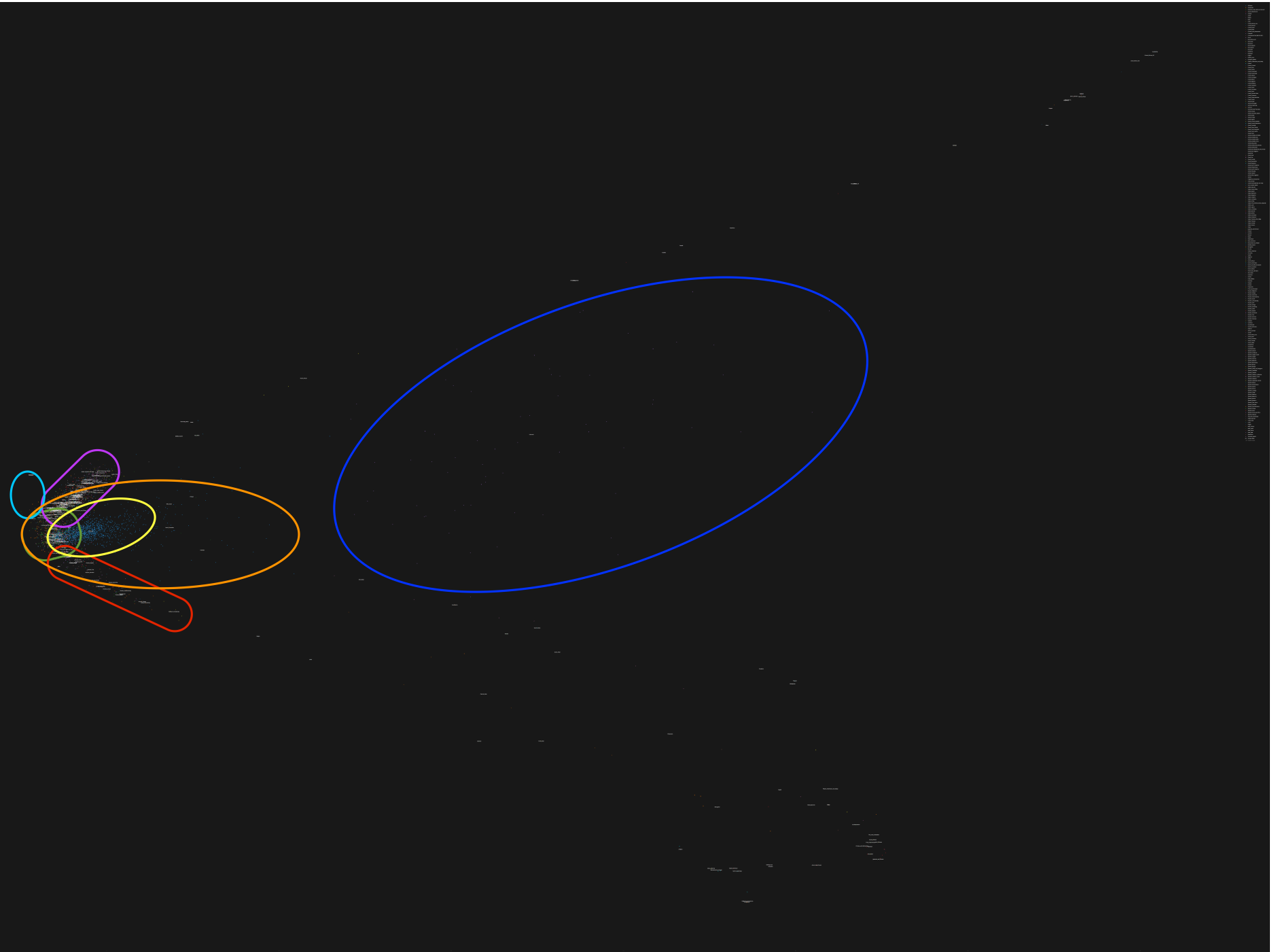

Dutch
French
German
O. European
S. Asian
W. African
S. African
EA & SEA
E. African
O. European
4.8%
German
18.9%
Dutch
49.8%
French
24.2%

Figure 6c.i

Green cluster is South Eastern Europe. Yellow cluster is Southern Europe. Teal cluster is North and Western Europe. Red cluster is Eastern Europe. Blue cluster includes typical Afrikaners. Magenta cluster includes almost all 1388 Afrikaner samples. Orange cluster includes almost all South African Coloured samples. Clusters Heuristically assigned and may be inaccurate.

Figure 6c.ii



Teal cluster is West Mediterranean. Green cluster is North and Western Europe. Magenta cluster is Southern Europe. Red cluster is Eastern Europe. Yellow cluster includes most typical Afrikaners. Orange cluster includes almost all Afrikaners. Blue Cluster includes almost all South African Coloureds. Clusters heuristically assigned and may be inaccurate.

DISCUSSION

Previous studies have observed many founder effects in regards to heritable medical conditions in the Afrikaner population of South Africa and its diaspora, which have implied a high level of inbreeding. [1], [2], [3], [4], [5], [6], [7], [8], [9] , [10], [11], [12] This genetic study has confirmed the presence of these observed founder effects when heritable medical conditions were studied in parallel with previous studies (porphyria, cystic fibrosis), in one case observed no founder effect where previous studies had observed a founder effect (schizophrenia), and also identified both benevolent (below average risk) and malevolent (above average risk) novel founder effects that had not yet been observed elsewhere. The implication of the high frequency of founder effects in the Afrikaner population is that the population is inbred. This has been found to be true. The intensity of endogamy or inbreeding, determined by the amount of shared centimorgans, in the Afrikaner population of South Africa and its diaspora, cannot be compared to any present day European population, with their average of ~48.3 cM exceeding even that of Ashkenazi Jews (23 cM) by more than 200%. Compared to homogenous (mostly regional) European populations: Tuscans (9.3 cM), Finns (10 cM), Sardinians (12.4 cM) and Flemings (10.3 cM), the Afrikaner average of ~48.3 cM is about 4-5 times greater. Compared to heterogenous European populations: Italians (0.1 cM) and European Canadians (0.6 cM), Afrikaners have a considerably higher level of inbreeding. [25], [26] Considering the previously identified effective founding population equivalent to around 440 individuals arriving simultaneously, this intense endogamy and the founder effects in regards to heritable medical conditions associated with it are not surprising. The methods used by J. M. Greef et al. to calculate the intensity of endogamy/inbreeding in the Afrikaners they investigated, runs of homozygosity, is not as effective as investigation of shared centimorgans and is not as conclusive as suggested. This may explain the difference in findings. High rates of skin cancer have been observed in the Afrikaner population of South Africa. It has been suggested that this is a result of their fair skin and higher levels of UV light exposure relative to other European populations. [27], [28], [29] Interestingly, this study found that the 72 Afrikaners it examined for heritable medical conditions had an ~8.1 times higher risk variant allele frequency (or, risk) for basal cell carcinoma (a type of skin cancer) compared to other European populations. This suggests that the elevated rate of skin cancer observed in the Afrikaner population is more-so a result of genetic predispositions than environmental factors such as UV exposure.

Interesting discoveries were also made as a result of the investigation into the genetic traits of the Afrikaners. An intermediate phenotype prediction in regards to the warrior versus worrier phenotype was more common in the Afrikaner population relative to other European populations (Whites from Utah). A higher frequency of a lactose persistent phenotype in the 72 Afrikaners analysed (88.9%), concordant to that of the Dutch (88%), was observed relative to a previous study of 77 Afrikaners by Greef et al. which found a lactose persistent phenotype in 83% of the individuals it examined. [9] The difference is likely a result of sampling error. Such a high level of lactose tolerance in the population is expected due to their North Western European heritage. Contrary to the findings of Botha & Pritchard (1972), there does not appear to be any sign of non-European admixture based on the distribution of blood type genetic predispositions for the Afrikaners analysed. Although the prevalence of a predisposition for blood type B is slightly higher than the prevalence in the major founding populations (Dutch, German, French) of the Afrikaners, the distribution still falls within typical ranges for Europeans, and is most likely entirely explained by an over estimation of the predisposition for the B blood type. [49] This study also found 54.2% of the Afrikaners examined to have some form of blue eyes. This is among one of the highest rates of blue eyes in the world, dethroning Poland from its 9th position globally for blue eye prevalence. [44]

Based on the fractions of their specific (Dutch, German, French, and other) European admixture, one would expect about 40-45% of the eye colours in the population to be blue. The higher than expected frequency is likely the result of a sampling founder effect due to the small effective founding population, and sexual selection for blue eyes. Conversely, such an effect was not observed for hair colour, with the observed frequency of some form of blond hair (27.4%) being lower than the expected 44-35% based on the fractions of their specific European admixture. The reasons for this are probably the same as the reasons for the the increased observed frequency of blue eyes, however, it may also be the result of individuals with dyed blond hair being counted as blond in the comparison data. [45], [46], [47], [48]

A greater alternative allele frequency was observed in EDAR's rs3827760 relative to other European populations, which can suggest Asian genetic input. Although signs of admixture appear to be present based upon this SNP, in every other EDAR SNP examined (rs6542787, rs260690), the 72 Afrikaners had a distribution frequency within the typical margin of error for the alleles associated with European traits, suggesting that non-European input is not to blame for the increased heterozygous allele frequency in rs3827760. There was one Afrikaner who had no European alleles in any of these 3 EDAR SNPs. Comparing to the Finnish, who have approximately 3.2% Asian/Siberian-like admixture (Materials and methods) and a purely non-Asian (homozygous A) allele frequency of 89.2% in rs382776, it can be calculated that an average Southeast or East Asian input of around 0.9% is required to explain the allele frequency of this SNP alone. Although this is higher than the 0.3% inferred in this study, accounting for the lower level of East Asian alleles observed in the other SNPs compared to other similar European populations (British) by measuring the relative distance (1-AFR allele freq. / 1 - BRI allele freq.) and multiplying by the predicted Southeast or East Asian input due to the allele frequencies in rs382776, a Southeast or East Asian input of ~0.5% is required to explain the allele frequencies observed. Although this is negligibly higher than the inferred average Southeast or East Asian admixture fraction of ~0.3%, this may be explained by the overrepresentation of recently mixed individuals in the 72 Afrikaners who donated their undisclosed genetic data to the study, who, as a result of this, had a higher than average inferred Southeast or East Asian contribution of 0.5%.

Similarly to the warrior versus worrier phenotype, when it comes to muscle type in rs1815739, Afrikaners are also more balanced than other comparable European population, and also have a higher than typical heterozygous allele frequency compared to those observed in comparable European populations. This is likely a result of a sampling related founder effect, rather than inbreeding related founder effect. Similar patterns were observed for: photic sneeze reflex in rs10427255, DOA activity and NSAID tolerance in rs10156191, omega 3 unctuous acid and docosahexaenoic acid in rs174528 and resting heart rate and omega 6 & 3 acid levels in rs174547.

In regards to genetic predisposition to smoking behaviour in rs1051730, a small minority of the Afrikaners examined were homozygous for the alternative allele linked to smoking behaviour, which is typical of European populations.

In regards to one's genetic predisposition to anger based on the T allele in rs2148710, the 72 Afrikaners, overall, were less predisposed to anger than is typical for other European populations. This is surprising, for one would expect them to be more predisposed to anger, due to the minor amount of African ancestry in some Afrikaners, as the allele associated with anger predisposition is most common in Africans. [30], [31], [32], [33]

Nothing truly out of the ordinary was observed for the biomarker predispositions of the Afrikaner population. An average genetic predisposition to lower (5.1963%) HBa1c levels relative to other European population was observed, which is in accord with other studies, which found an average of 5.64% in White South Africans. [6] This may be a result of early natural selection (the conditions on the ships arriving at the early Cape necessitated a decreased requirement for sustenance due to the long journeys and rationed food supplies), but it may also be a result of later natural selection (the conditions of the great trek inland also necessitated the ability to go for long periods of time without nourishment). An average genetic predisposition to higher (+1.4294 mg/dL) LDL cholesterol levels was observed, which is in accord with previous studies which have found a higher frequency of medical conditions such as hypercholesterolaemia. [7]

The observed mitochondrial DNA haplogroups most common outside of Europe were lower (11.8%) than previous studies have estimated (12%-14%) or found (18%). [12], [13] The discrepancy between the results of this study and previous estimations are likely a result of naive estimation on their part, and the discrepancy between the observed haplogroup distribution frequencies is likely a result of sampling error on their part (185 versus 306). It should be noted that, just because a maternal progenitor's haplogroup is most common outside of Europe, does not mean they were necessarily non-European, so a higher than expected amount is not indicative of much. For example, Maria Bollart (Bellardy), who was an English 1680s settler born in the Netherlands, had the mitochondrial DNA haplogroup L4b2b, most common in Mali, Regina Christina de Joncker van Ahrensdorp (von Ahrendtsdorf, a village in Germany) had the mitochondrial DNA haplogroup L0a1b2a, most common in Southern Africa, Catharina Vlasvart (Flachsbart) from the "Bengel" municipality in the Rhineland, Germany, had the mitochondrial DNA haplogroup M33, most common in South Asia, and Lijsbeth Sanders, born in Germany in 1655, had the haplogroup L3b, most common around East, West, and South Africa. There are various other examples, but these serve their purpose well enough. The observed Y-DNA haplogroups most common outside of Europe was slightly higher (0.5%) than other studies had estimated or observed (0%). [12], [13] This can likely be entirely explained by sampling error. It should be noted that the one haplogroup (K-M9) observed that was most common outside of Europe, was brought to South Africa by a German progenitor (Adolph Jonker), and although its highest frequency is outside of Europe (Southeast Asia and Austronesia), it has been observed at relatively low frequencies throughout Europe (0-2.2%). [34]

Previous genetic studies and interpretations of genealogical records have inferred or estimated the non-European contribution to the Afrikaners to be as low as 0.5% and as high as 7.5% (predicted to be 9% by the GGSA based on misinterpretation of unrepresentative mitochondrial DNA findings), with its purported/estimated prevalence ranging from 57.1%-100%. [9], [10, [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24] This study including 1388 genetically identified Afrikaners inferred an average non-European admixture fraction of 2.3%, and a preponderancy of 74.2%, falling on the lower and middle sections of this spectrum respectively. It also inferred most of the non-European admixture components (1.8%) to originate from outside Southern Africa, which is not surprising considering there was only one recorded union between a Khoisan woman, Eva/Krotoa van Goringhaicona, and a European man, Pieter van Meerhoff, during the early years at the Cape, however her last confirmed living progeny (Johannes and Christian Zaaiman) died before they could reproduce as a result of the 1710s smallpox epidemic. It also observed a lower relative Asian input, with previous studies studies inferring its contribution to the non-European input to be ~55% [9], whereas this study inferred it to be ~43.5% (1/2.3). Previous genetic studies have inferred a West and East African admixture contribution to the average Afrikaner of 0.8%. [9] This study genetically

inferred a concordant figure of 0.6% West African and 0.2% East African. Based on genetic analysis of the 358 100% European individuals, it was inferred that nearly half of the contribution to the average Afrikaner's admixture proportions is genetically Dutch, with just under a quarter being genetically French, with about a fifth being genetically German and one twentieth being from other European populations. This differs from older genealogical estimates due to the lower German and higher French contribution, and it differs from newer genealogical estimates due to the higher Dutch admixture fraction. [10], [11], [12], [14], [15], [16], [17], [18], [19] The main reasons why this study's genetic inferences differ to those of previous genetic studies are sampling errors and the fact different methods of inferring admixture fractions have been used. The 2020 study including 77 Afrikaners by J.M. Greef et al. clearly incorporated a non-representative proportion of more recently admixed individuals. For instance, individuals with an inferred non-European admixture fraction greater than 8% make up over 15.5% (12/77) of the sample cohort in J. M Greef et al.'s consummately unrepresentative study, whereas in this thoroughly representative study, they make up 6.3% (88/1338), which is less than half the amount observed due to sampling error in J.M. Greef et al.'s sample cohort. [9] The same is true for both of R. Khan's small scale investigations, where such individuals made up 25% of the sample cohort (3/12). [21], [22] Although there are evidently recently admixed individuals included in this study, due to its excessively populous sample cohort, their genetic contribution is not exaggerated, and they do not over inflate the average, as they have in every previous study. If this study were limited to only the individuals who had donated their undisclosed genetic data, the amount of recently mixed individuals would have also been overexaggerated, as they made up around 10% of the 72 individuals, which is a testament to the effect sampling error can have in sample sizes like these. Furthermore, over 70% (54/77) of the individuals included in the 2020 Study J.M. Greef et al. were recently related, sharing the same surname (as they were used in parallel with a study on non-paternity rates). [9] The statistical skew as a result of the considerable number of family groupings in J.M. Greef et al.'s study is more than enough to invalidate its findings, and to explain the differences of its results with those of this study. Although there are some some family groupings included in this study due to the nature of the DNA uploading patterns of users of genealogical and genetic databases such as GEDmatch, they are so few in number that they do not skew the results. To infer the admixture proportions of the 77 Afrikaners J.M. Greef et al. investigated, Admixture and Admixtools were used, whereas this study used PCA, R (to run the nMonte script), and Euclidean Optimisation (more information in "Materials and Methods") to infer the admixture proportions of each of the 1388 Afrikaners investigated. This may also explain the divergence in findings regarding admixture. The main reason as to why the results of this study in regards to the Dutch versus German admixture proportions differ from genealogical estimates (Dutch: 31.4%-50.6% German: 24.5%-35.7%) is due to the limits of genetic inferences, and the rather arbitrary distinction between Dutch and German. Most of the Germans employed by the VOC came from the North Western regions of Germany, such as German Friesland or Saxony, or were Dutchmen living in Germany (e.g. Abraham de Hartogh). These individuals were preferred due to their mutually intelligible dialects, regional proximity to the Netherlands and cultural compatibility. [35], [36] These Germans, though geographically distinct, were genetically more similar (or in some cases, identical) to their Dutch employers and neighbours, than to their German compatriots. This may explain the variation between this study's genetic inferences and previous studies' genealogical estimates. The primary reason for the discrepancies between the genetic inferences of this study concerning the average non-European admixture proportion of the Afrikaner and certain genealogical estimates lies in the fact that much of prevailing early South African genealogy is largely based on misguided conjecture, biased speculation, and spurious theoretical assumptions. These elements are often erroneously presented as indisputable facts by dishonest individuals and genealogists lacking

integrity. Subsequently, this reduces the reliability of these genealogical estimates, making it no surprise that the genetic inferences of this study would differ from previous genealogical estimates, especially the genealogical fantasies and dubitable findings of J.A. Heese. Another reason may be that the methods, formulas, and calculations used to estimate the admixture proportions based on these genealogical registers may be flawed, as suggested by C.F.C De Bruyn, whose estimate, using the exact same genealogical register, differed greatly from the fantastical estimates of J.A. Heese. On a 2 dimensional PCA chart plotting PC1 and PC2, there is an uncanny drift observed in a large proportion of samples in the Afrikaner cluster that is similar to that of the Coloured cluster, suggesting such admixture. Regardless of this drift, the distribution of many (but not all) of the Afrikaner samples along PC1 and PC2 do suggest minor (and in very rare cases, major) African and Asian admixture. Despite this, a large proportion of the Afrikaner samples still plot inside the (North-Western) European cluster, which is not surprising considering their high average (97.7%) and median (98.9%) European ancestry. The genealogical study which is most concordant with this genetic study (this study's results in brackets) is D. M. Philpott's pedigree analysis of 32 influential Afrikaner individuals (2012), which found (even when using the genealogical fantasies and dubitable findings of J. A. Heese) an average non-European contribution of 2.3% (2.3%), with a median of 1.6% (1.1%) and a ubiquity/pervasiveness of 75% (74.2%). [12]

MATERIALS AND METHODS

Sample Collection:

The samples used in this study were scraped from the GEDmatch database. The following was done to extract potential (low confidence) Afrikaners from the GEDmatch database:

1. Obtain the top 3000 closest matches of 70 known Afrikaners and store in a database.
2. Remove individuals without one of 2250 Afrikaner surnames (obtained from the "Cape Dutch / Kaaps-Hollands" project on Family Tree DNA).
3. Remove duplicates.

These three processes reduced the amount of potential (low confidence) Afrikaners to 3434 individuals or "kits". In order to positively identify who was an Afrikaner, and who just happened to have a surname common in the Afrikaner population of South Africa and its diaspora, the following was done:

1. Run each individual kit on the one-to-many tool on GEDmatch.
2. For each individual's 3000 closest matches, count the number of matches with a ".za" email and store in a database.
3. Remove individuals below an arbitrarily set threshold (18) of .za emails in top 3000 closest matches.

These reduced the amount of potential (high confidence) Afrikaners to 1527 individuals or "kits". A threshold of 18 was chosen, because that was the lowest amount of .za emails observed in an individual who, when searched for and identified on Facebook, was from South Africa and appeared to be an Afrikaner. The average amount of .za emails in those below the threshold (id est, non-Afrikaners) was 3.1. In those above the threshold (id est, South Africans (specifically, Afrikaners and a small minority of Coloured individuals)) the average amount of .za emails was 238.2. This also effectively distinguished between English White South Africans (most of whose matches would be continental Europeans or Anglosphere individuals) and Afrikaner White South Africans (most of whose matches would be other Afrikaners and continental Europeans). In order to differentiate between "Coloured" individuals (who on rare occasions appear in the matches of Afrikaners and in almost all cases pass the .za threshold) and "White" individuals in the remaining kits, a phenotype prediction was undertaken for each individual using the admixture analysis method described later in this section of the study to calculate the impact of each human phenotype. The phenotype morphs which were weighted and averaged based on an individual's admixture proportions to predict their phenotype were obtained from humanphenotypes.net or the now deleted physical anthropology wiki. Each phenotype prediction was algorithmically compared with OpenCV and dlib to the phenotype morphs from humanphenotypes.net, classified as either "European" or "Non-European". If an individual's phenotype prediction was classified as "mixed" or "non-European" by the program, they were flagged for manual review, which in some cases included a background check. If the individual phenotype prediction was indeed not "White passing", according to the assumed standards of the untrained eye and crowd wisdom from anonymous Afrikaner individuals, they were excluded from the dataset. If, after manual review, it was determined that the program had fallaciously classified an individual's predicted phenotype as mixed or non-European (which was extremely rare), they remained in the dataset. This reduced the amount of high confidence genetically identified Afrikaners to 1448 individuals or "kits". Below are the average phenotype predictions for the 1388 Afrikaners and 79 Coloured individuals (Figure 7a) and a few examples of predicted phenotypes, mostly those which had been flagged, some of which were excluded from the dataset, some of which were remained in the dataset (Figure 7b).

Figure 7a

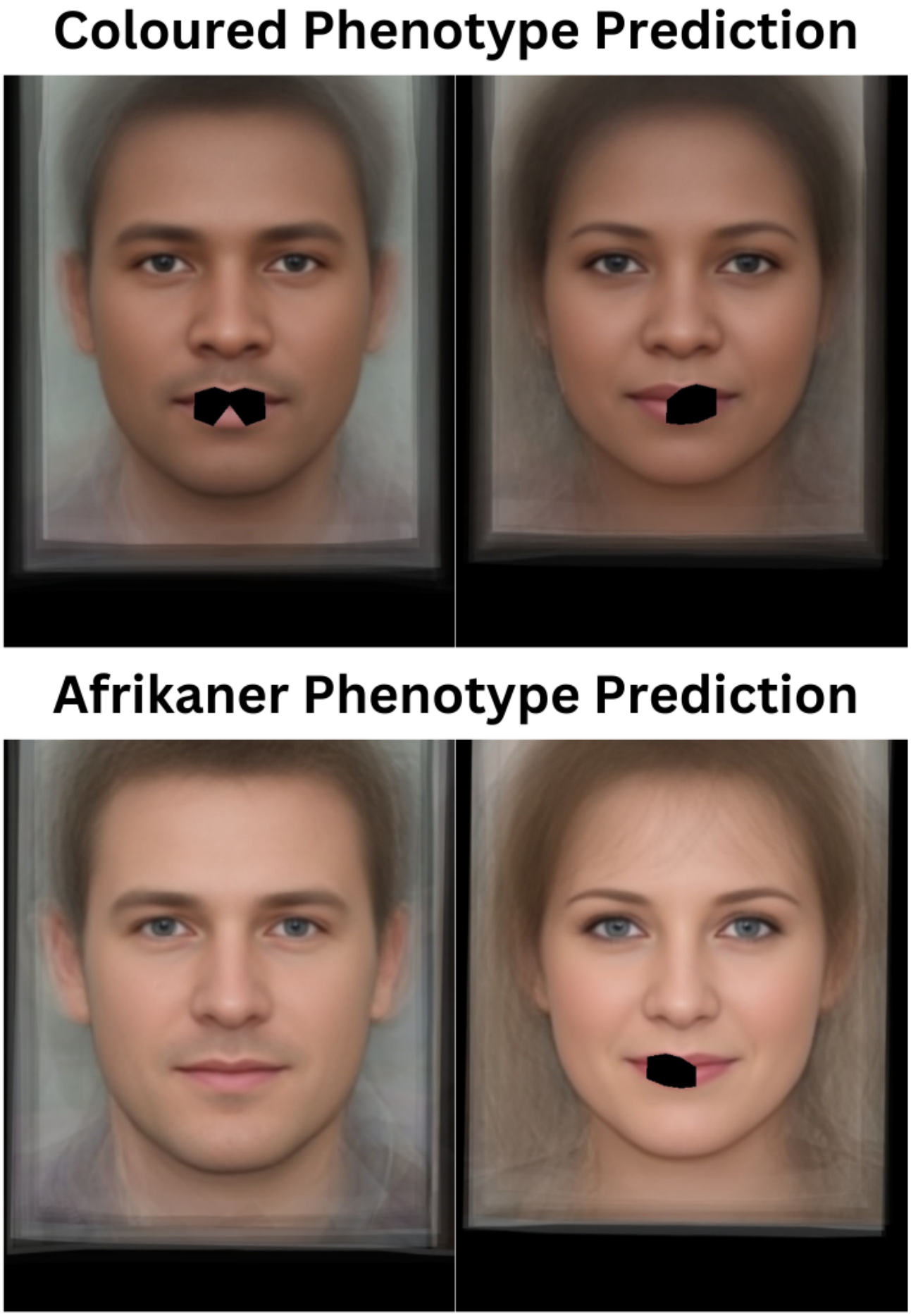
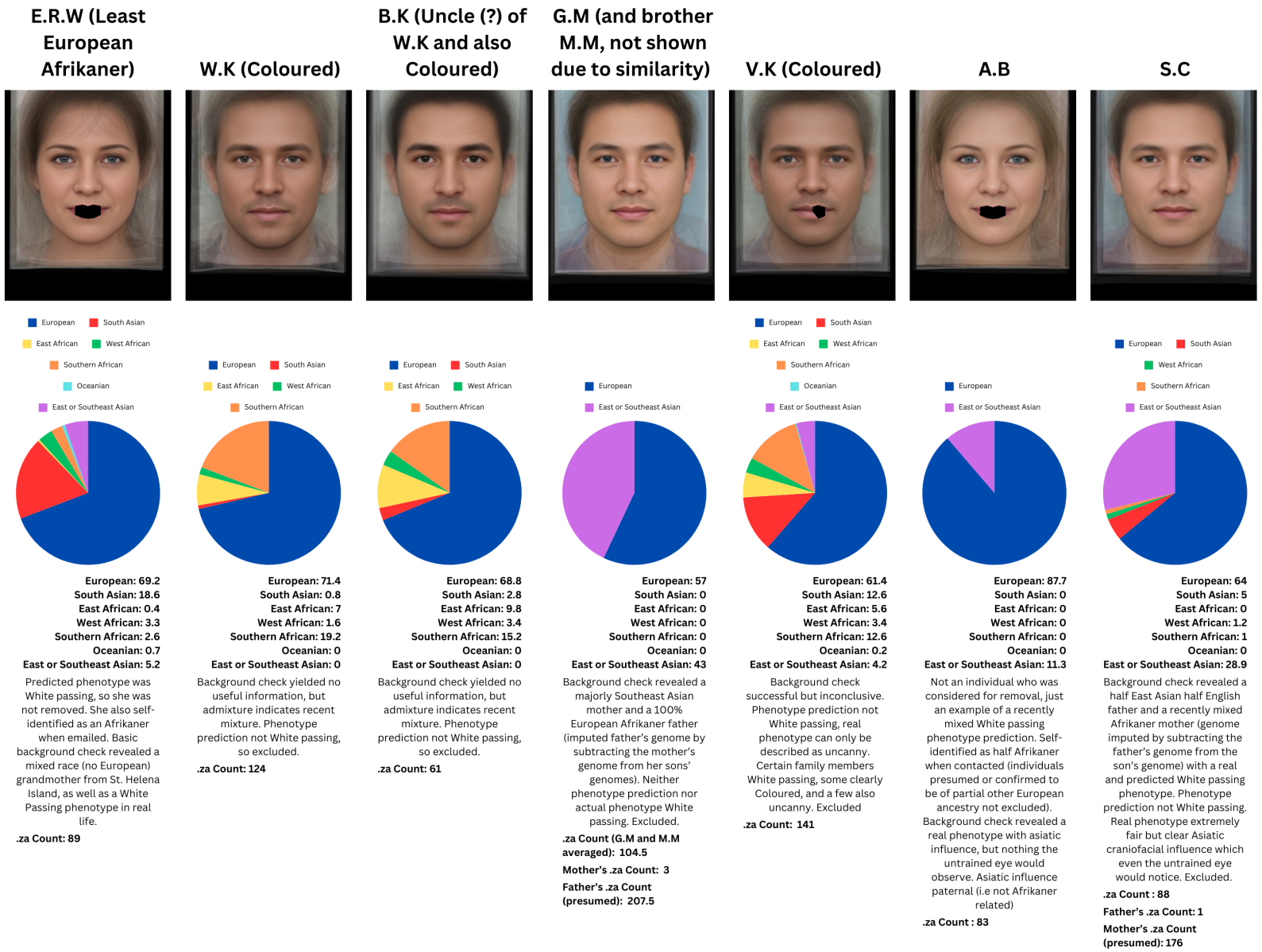

Coloured Phenotype Prediction
Afrikaner Phenotype Prediction

Figure 7b


E.R.W (Least European Afrikaner)
W.K (Coloured)
B.K (Uncle (?) of W.K and also Coloured)
G.M (and brother M.M, not shown due to similarity)
V.K (Coloured)
A.B
S.C
European
South Asian
East African
West African
Southern African
Oceanian
East or Southeast Asian
European: 69.2
South Asian: 18.6
East African: 0.4
West African: 3.3
Southern African: 2.6
Oceanian: 0.7
East or Southeast Asian: 5.2
Predicted phenotype was White passing, so she was not removed. She also self-identified as an Afrikaner when emailed. Basic background check revealed a mixed race (no European) grandmother from St. Helena Island, as well as a White Passing phenotype in real life.
.za Count: 89
European: 71.4
South Asian: 0.8
East African: 7
West African: 1.6
Southern African: 19.2
Oceanian: 0
East or Southeast Asian: 0
Background check yielded no useful information, but admixture indicates recent mixture. Phenotype prediction not White passing, so excluded.
.za Count: 124
European: 68.8
South Asian: 2.8
East African: 9.8
West African: 3.4
Southern African: 15.2
Oceanian: 0
East or Southeast Asian: 0
Background check yielded no useful information, but admixture indicates recent mixture. Phenotype prediction not White passing, so excluded.
.za Count: 61
European: 57
South Asian: 0
East African: 0
West African: 0
Southern African: 0
Oceanian: 0
East or Southeast Asian: 43
Background check revealed a majorly Southeast Asian mother and a 100% European Afrikaner father (imputed father's genome by subtracting the mother's genome from her sons' genomes). Neither phenotype prediction nor actual phenotype White passing. Excluded.
.za Count (G.M and M.M averaged): 104.5
Mother's .za Count: 3
Father's .za Count (presumed): 207.5
European: 61.4
South Asian: 12.6
East African: 5.6
West African: 3.4
Southern African: 12.6
Oceanian: 0.2
East or Southeast Asian: 4.2
Background check successful but inconclusive. Phenotype prediction not White passing, real phenotype can only be described as uncanny. Certain family members White passing, some clearly Coloured, and a few also uncanny. Excluded
.za Count: 141
European: 87.7
South Asian: 0
East African: 0
West African: 0
Southern African: 0
Oceanian: 0
East or Southeast Asian: 11.3
Not an individual who was considered for removal, just an example of a recently mixed White passing phenotype prediction. Self-identified as half Afrikaner when contacted (individuals presumed or confirmed to be of partial other European ancestry not excluded). Background check revealed a real phenotype with asiatic influence, but nothing the untrained eye would observe. Asiatic influence paternal (i.e not Afrikaner related)
.za Count : 83
European: 64
South Asian: 5
East African: 0
West African: 1.2
Southern African: 1
Oceanian: 0
East or Southeast Asian: 28.9
Background check revealed a half East Asian half English father and a recently mixed Afrikaner mother (genome imputed by subtracting the father's genome from the son's genome) with a real and predicted White passing phenotype. Phenotype prediction not White passing. Real phenotype extremely fair but clear Asiatic craniofacial influence which even the untrained eye would notice. Excluded.
.za Count : 88
Father's .za Count: 1
Mother's .za Count (presumed): 176

Finally, a second manual review of the remaining kits in the dataset was undertaken to remove any duplicates of the same individual, with different company files. The matches of the original 70 source Afrikaners were also manually reviewed in order to identify any Afrikaner individuals using aliases or some other name not recognised by the sorting program. These additional individuals were also assessed using the previous .za count and phenotype prediction method discussed. This reduced the amount of individuals or "kits" to the final figure of 1388, who are all with almost 100% certainty of at least partial Afrikaner heritage, and White passing.

In order to obtain the samples that had their heritable medical conditions, genetic traits, and biomarker predispositions analysed, the managers of the 1388 Afrikaners GEDmatch kits were emailed. The email, which was updated and improved at various stages, informed them about the identification of the kit as an Afrikaner, the research use of the publicly available genetic data of the kit in question, and the purposes and aims of the study. The email also included a request for the undisclosed raw DNA file of the kit, information about the analyses that would be conducted if the genomic information was provided, and a statement promising to share any findings or results related to the individual associated with the kit if the raw DNA file was provided. None of the individuals in the database who were not mistakenly emailed denied having at least partial Afrikaner heritage. This yielded the raw DNA files of (at the time of writing) 76 individuals, however, only 72 were included in this study, as 4 of the individuals had sent their files after the analyses had been undertaken.

Analysis of heritable medical conditions, genetic traits, and biomarker predispositions:

To analyse the above, the raw DNA files obtained from email requests were run through Andrei Kharchenko's "AndreiDNA Trait Predictor". [43] The HTML reports generated were parsed and the information extracted was stored in a spreadsheet. The information in the spreadsheet was then interpreted and reported on in various modes as seen in Figures 1 to 3b.

Analysis of the presence and intensity of endogamy/inbreeding:

In order to analyse the above, 6 genetically identified Afrikaner individuals were chosen at random to act as source individuals. These 6 source individuals had the total amount and largest segment of centimorgans they shared with 1280 genetically identified Afrikaners obtained through comparison using the GEDmatch "One-to-one Autosomal DNA Comparison" tool. Default settings were used for: SNP window size threshold, Size (in SNPs) of Mismatch-Bunching limit and GAP to break stop a segment. A minimum cM segment size of 3 cM was chosen over the default of 7cM, as a 3cM segment would represent, under normal conditions, the contribution of a 13th generation ancestor. Most of the shared Ancestry within Afrikaners is found in the early cape colonial period, approximately 9-13 generations ago. A minimum segment size of 3 cM would account for this distant ancestry. The amount of individuals the 6 source individuals were compared to is lower than the total amount of Afrikaner kits identified, because some had been made private or deleted when this analysis was undertaken.

Analysis of the frequency of the areas of highest distribution and the frequency of major clades for Y-DNA and mitochondrial DNA haplogroups:

In order to analyse the above, each of the 1388 genetically identified Afrikaners were once again examined using GEDmatch's "One-to-Many - Limited Version" tool in order to identify those who

had made their haplogroup(s) publicly available. The tool in question was used because it gives access to the haplogroup information of kits (if any is available), while others do not. This yielded a total of 312 individuals with publicly available mitochondrial DNA haplogroups and 206 individuals with publicly available Y-DNA haplogroups. To prevent statistical skew from the substantial number of familial groups within the dataset of the same maternal progeny (in the case of mitochondrial haplogroups) or paternal progeny (in the case of Y-DNA haplogroups), such pairs were reduced to one individual at random if neither was more useful than the other (eg. two brothers or two sisters), or to the most useful individual (e.g. a brother and a sister share the same mitochondrial haplogroup information, but the brother also contains Y-DNA haplogroup information, which his sister would not due to her sex, so the brother would be kept as he would be more useful). In a case where two male individuals share one relevant piece of haplogroup information, but not the other (e.g. father and son: same Y-DNA but different Mitochondrial DNA), either of the individuals would simply have their Y-DNA haplogroup information removed. The individuals with Y-DNA haplogroups were further filtered by removing all those without an Afrikaner surname (one individual). This reduced the total mitochondrial DNA haplogroup samples to the 306 in the study, and the total Y-DNA haplogroup samples to the 197 in the study. The haplogroup each individual provided was also summarised into its major clade (eg. R1b1a2a1a2c1f summarised as R1b) for simplicity.

Inference of admixture proportions:

For each of the 1388 Afrikaner kits, their publicly available genomic information relevant to the prediction of their principal components was summarised into 36 distinct numerical values using various tools on GEDmatch. The amount of SNPs summarised varied depending on the source of the kit. The average amount of SNPs summarised as well as the frequency for each source/company are below:

BH2BU: 52,000 - 55,000
23&me: 52,000 - 52,500
AncestryDNA: ~155,000
MH/FTDNA: ~72,400
Ambiguous/Other: N/A (weighted average of BH2BU, 23&me, MH/FTDNA & AncestryDNA: 93677 SNPs summarised)

BH2BU: 70/1291
23&me: 121/1291
AncestryDNA: 282/1291
MH/FTDNA: 445/1291
Ambiguous/Other: 373/1291

These 36 numerical values were used to predict a 25 dimensional set of principal component values using a 25 wide 36 deep linear regression matrix for each of the 1388 Afrikaner kits. The aforementioned 25 wide 36 deep linear regression matrix was trained on over 2650 individuals from a wide range of diverse geographical backgrounds, 70 of which were of Afrikaner ancestry (the raw DNA files received had real principal component analysis undertaken by BA MA "Cyrus Irani"), and over 350 population averages from a similarly wide variety of populations. [42] The scaled euclidean distance between the predicted average of 72 Afrikaners, and the real average of those same 72 Afrikaners, was 0.00352965, which is the lowest euclidean distance of any predicted

population average to a real population average (that had a comparison made). Such a low distance suggests (to a very strong extent) an accurate prediction for each individual in the study.

To ensure the calculator used to infer the patterns of European and non-European admixture in this study was accurate, the source populations were selected in consideration to previous studies and the history of South Africa. The calculator created included 382 source individuals from relevant populations across the world, and 54 source population averages. The broad population clusters included in the calculator were: European, South Asian, East or Southeast Asian, Oceanian, West African, Southern African and East African. 211 of these individuals were Europeans from various geographic clusters (Germanic, Insular Celtic, Continental Celtic, Slavic, Baltic, Volgan, Iberian, Sardinian, Italic, Balkan and European Anatolian). 93 of the individuals were East or Southeast Asian from various geographic clusters (Mongolic, Hezhen, Tibetan, Malaysian, Laotian, Taiwanese, Thai, SEA Chinese, SEA Indian, Philippine, Sinitic, Japanese and Hmongic–Mienic). 35 of these individuals were South Asian from various geographic clusters (Indian and Pakistani). 8 Individuals were Oceanian from various geographic clusters (Australian and Papuan). 35 individuals and 11 population averages were East African from various geographic clusters ("Cushitic", Nubian, Nilo-Saharan, Hadzabe and Sandawe). 36 populations averages were West African from various geographic clusters (West Africa, Bantu). 7 population averages were Southern African from various geographic clusters (Pygmy, Khoisan).
To ensure the calculator used to infer the type of European admixture in Afrikaners without any inferred non-European admixture in this study was accurate, the source populations were selected in consideration to previous studies and the history of South Africa. The calculator created included 24 population averages from relevant European populations. The broad population clusters were: Dutch, German, French, English, Danish and Belgian. 1 population average was Dutch. 14 population averages were French, from the following regions: weighted France average, Alsace, Auvergne, Bearn, Bigorre, Brittany, Chalosse, Nord, Occitanie, Paris, Pas de Calais, Provence, Seine Maritime and the South. 4 population averages were German, from the following cities or regions: Hamburg, Berlin, Erlangen and East Germany. 1 population average was English. 1 population average was Danish. 3 population averages were Belgian, of varied genetic profiles.

The 25 predicted principal components of each of the 1388 Afrikaners were scaled by multiplying each principal component by the square root of its corresponding eigenvalue (which represents the total variance contained in each principal component) which are as follows:

129.557, 103.13, 14.222, 10.433, 9.471, 7.778, 5.523, 5.325, 4.183, 3.321, 2.637, 2.246, 2.21, 1.894, 1.842, 1.758, 1.7, 1.605, 1.58, 1.564, 1.557, 1.529, 1.519, 1.452, 1.434

Square rooted to 4 d.p: 11.3823, 10.1553, 3.7712, 3.23, 3.07748, 2.789, 2.35017, 2.3076, 2.04522, 1.82236, 1.62385, 1.498675, 1.4866, 1.3762, 0.13572, 1.32586, 1.30384, 1.2669, 1.257, 1.2506, 1.2478, 1.23654, 1.2324, 1.20498, 1.1974

The same was done for the 25 principal components of the individuals and population averages in the calculators. Scaling ensures each principal component receives an accurate weight according to the amount of variance it contains, whereas not scaling the principal components would lead to an unbalanced weight for each principal component. For example if left unscaled, the first global dimension (PC1) distinguishing Africa and Eurasia would receive a weight of 1, the second global dimension (PC2) distinguishing West and East Eurasia would also receive a weight of 1 and so on. This leads to less realistic euclidean distances and models. As a first case study, take 3 extreme

poles of world variation: Sardinians, Han Chinese, and Yorubans. Their euclidean distances from one another when unscaled are as follows:

Distance to: Yoruba
0.10343781 Han
0.10482554 Sardinian

Distance to: Han
0.09992067 Sardinian
0.10343781 Yoruba

Distance to: Sardinian
0.09992067 Han
0.10482554 Yoruba

As can be seen, the distance between the two Eurasians is almost the same as both of their distances from a West African population. This is not what one would expect from the known phylogeny of humans, id est, pure Eurasians forming a clade and thus being closer to one another than they are to Africans. Using scaled sets of principal components produces much more realistic results:

Distance to: Yoruba
0.77636788 Sardinian
0.84430462 Han

Distance to: Han
0.64350770 Sardinian
0.84430462 Yoruba

Distance to: Sardinian
0.64350770 Han
0.77636788 Yoruba

The realistic results in question being that Eurasians are closer to one another than they are to Africans. When inferring admixture using euclidean optimisation of the principal components, this also means that populations which one would expect to be distant from another, are not considered as distant as they truly are, which can lead to some very bizarre results. For an example, two models of the Finnish using bronze age, mesolithic, and neolithic ancestral Eurasian components (PIE, ANF, WHG, EHG, Krasnoyarsk) and a West African source, one unscaled, one scaled:

Unscaled:
Target: Finnish_Southwest
Distance: 3.2433% / 0.03243311
53.0 PIE
23.6 ANF
10.2 WHG
7.2 Krasnoyarsk_BA
3.6 Yoruba
2.4 EHG

Scaled:
Target: Finnish_Southwest

Distance: 5.7361% / 0.05736089
43.4 PIE
26.4 ANF
18.8 WHG
8.2 EHG
3.2 Krasnoyarsk_BA
0.0 Yoruba

Now that the background information has been provided, an explanation of the methods utilised for analysing admixture can proceed. 1744 cycles of euclidean optimisation using the nMonte 3 script on R of the 1388 sets of scaled 25 dimensional predicted principal components for each sample's publicly available genomic information were undertaken for the purpose of inferring admixture proportions. nMonte is an R-program that takes as input a calculator file and a target file. The program then calculates mixtures of all the populations in the calculator file; it searches for the mixture which has the shortest Euclidean distance to the sample(s) in the target file (euclidean optimisation). The average euclidean distance between the inferred admixture proportions and the individuals being analysed was ~0.002, with a maximum of ~0.012 and a minimum of ~0.00002, which suggests a relevant and thorough selection of source populations for the calculator, and an accurate inference of admixture proportions for each individual analysed.

PCA plot:

The PCA scatter graphs in Figure 6.c.i and Figure 6.c.ii plotting PC1 and PC2 were generated with plotly. The other populations plotted were chosen based on the capricious basis that they were tier relevant, or would improve the quality of the PCA plot by acting as roots. Due to the predicted nature of the Afrikaner PC coordinates, variation and drift to the left side of the European plot could not be entirely captured, leading to the impression of a greater non-European drift than is actually present. Fortunately, this did not affect the admixture inferences.

VALIDITY CHECKS

Mitochondrial haplogroups:

In order to confirm whether or not mitochondrial haplogroup results in regards to the proportion of haplogroups most common outside of Europe differed from previous studies due to sampling error on their part or the part of this study, a formula to predict the mitochondrial DNA haplogroups of a (colonial) population with a sex and time bias in admixture was necessary. previous studies have predicted the mitochondrial haplogroups of the Afrikaners by multiplying their genealogically estimated non-European ancestral contribution by 2. This is naive as it does not take into account the sex bias of the non-European contribution and the complex relationship between admixture and mitochondrial haplogrouping in sex and time biased colonial populations. In order to create a formula to predict the proportion of haplogroups most common outside of the area of origin of the dominant ancestral fraction of a colonial population with a sex and time bias based on their admixture fraction, parallel populations needed to be identified. The parallel populations chosen were White and African Americans, due to their similar sex and time bias (White Americans: mixture with African and Native American females during early settlement, African Americans: mixture with European males during early slavery) and colonial status.

A genetic study investigating the admixture proportions based on the autosomal DNA of various American ethnic groups inferred the contribution from Amerindian and African populations to the average of 148,789 White American genomes to be approximately 0.4% split equally between the two groups. [37] A similarly large scale study regarding the mitochondrial haplogroups of various American ethnic groups including (8,537) White Americans found approximately 1% of their mitochondrial DNA haplogroups to be most common in Africa, and 1.1% to be most common in Native Americans. [38] Another genetic study investigated the mitochondrial haplogroups, Y-DNA haplogroups and admixture proportions of Americans of various ethnic groups, 246 of which were African Americans. It found around 5.1% of African American mitochondrial haplogorups to be most common in Europe, 1.8% to be most common in Native Americans or Asians, and the remainder in Africans. It also found approximately 29.7% of Y-DNA haplogroups to be most common in Europe, 0.1% to be most common in Native Americans, and 69.2% to be most common in Africa. It also inferred approximately 7.8% of the autosomal admixture of the African Americans it investigated to be European (about 6.6% paternal related), 5.5% to be Asian or Native American, and the remainder African. [39] Another study including 599 African Americans found 33.6% of the Y-DNA haplogroups in the population to be most common in Europe. [40]

Using these figures, a formula can be created to predict the proportion of haplogroups most common outside of the area of origin of the dominant ancestral fraction of a colonial population with a sex bias based on their admixture fraction by dividing the proportion of areas of origin for mitochondrial DNA haplogroups (in the case of White American) or the proportion of areas of origin for Y-DNA haplogroups (in the case of African Americans) by the total inferred admixture from those groups. Doing this gives the following equations:

White Americans:
EA/NA: 1.1/0.2 = 5.5
African: 1/0.2 = 5
EA/NA & African: 2.1/0.2 = 5.25

African Americans:
Y-DNA 1: 29.7/6.6 = 4.5
Y-DNA 2: 33.6/6.6 = 5.1
Afrikaner:
11.8/2.3 = 5.13
Average (not including Afrikaner):
25.35/5 = 5.07

So the formula for predict the proportion of haplogroups most common outside of the area of origin of the dominant ancestral fraction of a colonial population with a sex bias is as follows: admixture from outside dominant ancestral component * 5.07 = approximate proportion of Y-DNA/ mitochondrial DNA (depending on sex bias) haplogroups most common in the area(s) of origin of non-dominant component.

Inserting the total non-European admixture fraction inferred in the study into the equation:

2.3 * 5.07 = 11.661% of mitochondrial DNA haplogroups most common outside of Europe.

The difference between this estimated figure and the figure observed (11.8%) is entirely negligible (<0.15%). As well as suggesting the mitochondrial haplogroup proportions observed were accurate and not subject to sampling error, this also suggests the non-European admixture proportions inferred were accurate.

To conclude, the previous study of 185 Afrikaner individuals' mitochondrial haplogroups which found 18% to be most common outside of Europe had a sampling error or bias. This study which found 11.8% of the mitochondrial haplogroups of 306 Afrikaners to be most common outside of Europe did not suffer from such sampling error or bias.

Ubiquity of non-European admixture:

In order to confirm the findings of this study in regards to the ubiquity of non-European admixture in the Afrikaner population of South Africa and its diaspora (74.2%), an equation to calculate the growth in the proportion of mixed (to any degree) individuals every generation was created.

$$proportion\ of\ P\,(n+1) = \frac{\left(P\,(n) - \left(M\,(n) \cdot \left(1 - \frac{M(n)}{1}\right)\right)\right)}{M\,(n) + \left(P\,(n) - \left(M\,(n) \cdot \left(1 - \frac{M(n)}{1}\right)\right)\right)}$$

$$proportion\ of\ M\,(n+1) = \frac{M\,(n)}{M\,(n) + \left(P\,(n) - \left(M\,(n) \cdot \left(1 - \frac{M(n)}{1}\right)\right)\right)}$$

In Python format:

```
def simulate_ancestry(N_M, N_P, num_generations):
    # Initial proportions
```

```
    P_M = N_M / (N_M + N_P)
    P_P = N_P / (N_M + N_P)
    proportions_M = [P_M]
    proportions_P = [P_P]

    for generation in range(1, num_generations + 1):
        # Calculate growth in the proportion of the mixed population

        P_P_new = (P_P - (P_M * (1 - (P_M / 1))))

        # Normalise the proportions
        total_population_new = P_M + P_P_new
        P_P = P_P_new / total_population_new
        P_M = P_M / total_population_new

        # Store proportions
        proportions_M.append(P_M)
        proportions_P.append(P_P)

    return proportions_M, proportions_P

# Example usage
N_M = XX
N_P = YY
num_generations = ZZ

proportions_M, proportions_P = simulate_ancestry(N_M, N_P, num_generations)

# Print results
print(f"Generation\tProportion of M\tProportion of P")
for gen in range(num_generations + 1):
    print(f"{gen}\t\t{proportions_M[gen]:.4f}\t\t{proportions_P[gen]:.4f}")
```

Due to the sex and time bias of the non-European admixture in the Afrikaner, the observed frequency of mitochondrial haplogroups most common outside of Europe could be reasonably used as a figure to represent the amount of mixed couples in a hypothetical generation 0 (although, as discussed previously in the study, this is not entirely accurate, as not all mitochondrial haplogroups observed which are most common outside of Europe, came from non-European maternal progenitors).

Although the longest path of the average Afrikaner can stretch back as far as 13 or even 14 generations before coalescing into a founding couple, this is not representative of the average path length before reaching a founding couple. In order to calculate the average path length before converging into a founding couple, multiple calculations had to be done. First, the relative contribution of founders from four 30 year (one 19 year) generational blocks was calculated by scaling the total amount of children born in each generational block and then dividing the scaled total for each generation by the total scaled amount of children across all generations. Secondly, the average year at which a path ended was calculated by multiplying the relative contribution of each

generation by their upper and lower bound of the years of their generational block. This gave an average year at which a path ended of 1711 (1696-1726). [10] For the final calculation, the age of the average Afrikaner was required. This figure was calculated to be approximately 40 (for White South Africans, likely representative of Afrikaners) based on the statistics provided in a 2017 mid year population estimate of South Africa. [41] Subtracting 40 from 2017 gives a date of birth of 1977 for the average Afrikaner. Using these two figures, 1711 (1696-1726) and 1977, the average path can be calculated to be between 281 and 251 years long, or 266 years long. Assuming an average generation time of 27.5 years (25 for women, 30 for men), the average path length in generations is between ~10.2 and ~9.1 generations long (avg. 9.67). Substituting these values into the equation from earlier:

Generation	Proportion of M	Proportion of P
0	0.1180	0.8820
1	0.1317	0.8683
2	0.1487	0.8513
3	0.1703	0.8297
4	0.1983	0.8017
5	0.2358	0.7642
6	0.2876	0.7124
7	0.3617	0.6383
8	0.4702	0.5298
9	0.6262	0.3738
10	0.8176	0.1824

Using linear interpolation, the proportions of M and P for generation 9.67 are as follows:
M: 0.7544
P: 0.2456

The difference in the proportion of P in this simulation (24.56%) versus the proportion of P observed in reality (25.8%) is minimal, and is likely the result of a slightly inaccurate figure when it comes to path length (for generations after 1807 were not considered in the equation), or assumptions made (e.g. every couple has the the same amount of children and each population breeds equally) which may not have been true in reality.

The concordance of these two figures suggests:

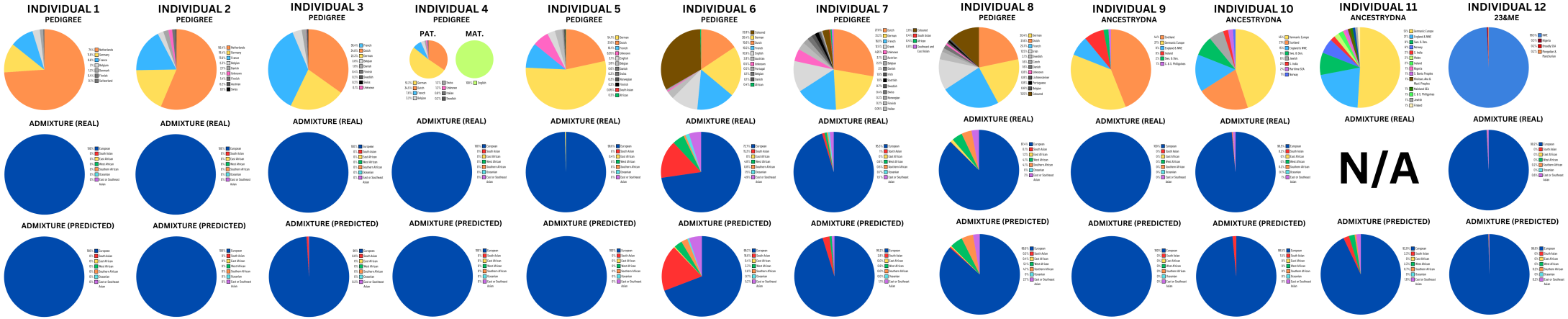
1. The findings of this study in regards the frequency of the distribution of the areas of highest frequency for the mitochondrial DNA haplogroups of the Afrikaner population of South Africa and its diaspora are correct.
2. The findings of this study in regards to the ubiquity of non-European admixture in the Afrikaner population of South Africa and its diaspora are correct.
3. The findings of this study in regards to the patterns of non-European admixture in the Afrikaner population of South Africa and its diaspora are correct.
4. The findings of other studies in regards to these issues are incorrect.

Admixture inference:

In order to confirm the validity of the admixture inferences in this study, Afrikaners who provided DNA test results or (at least almost entirely complete) pedigrees had a their admixture inference

based on their predicted and real principal components (the latter, if applicable) compared with their admixture estimate based on their pedigree. European population admixture inferences not included. Various genealogical registers and theories were used to interpret the information in the pedigrees provided. Coloured refers to any non-European ancestral contribution which is ambiguous in origin, or (in rare cases) to an ancestor(s) classified as Coloured. The comparisons (Figure 8) are below:

Figure 8


INDIVIDUAL 1
PEDIGREE
INDIVIDUAL 2
PEDIGREE
INDIVIDUAL 3
PEDIGREE
INDIVIDUAL 4
PEDIGREE
PAT.
MAT.
INDIVIDUAL 5
PEDIGREE
INDIVIDUAL 6
PEDIGREE
INDIVIDUAL 7
PEDIGREE
INDIVIDUAL 8
PEDIGREE
INDIVIDUAL 9
ANCESTRYDNA
INDIVIDUAL 10
ANCESTRYDNA
INDIVIDUAL 11
ANCESTRYDNA
INDIVIDUAL 12
23&ME
ADMIXTURE (REAL)
ADMIXTURE (REAL)
ADMIXTURE (REAL)
ADMIXTURE (REAL)
ADMIXTURE (REAL)
ADMIXTURE (REAL)
ADMIXTURE (REAL)
ADMIXTURE (REAL)
ADMIXTURE (REAL)
ADMIXTURE (REAL)
N/A
ADMIXTURE (REAL)
ADMIXTURE (PREDICTED)
ADMIXTURE (PREDICTED)
ADMIXTURE (PREDICTED)
ADMIXTURE (PREDICTED)
ADMIXTURE (PREDICTED)
ADMIXTURE (PREDICTED)
ADMIXTURE (PREDICTED)
ADMIXTURE (PREDICTED)
ADMIXTURE (PREDICTED)
ADMIXTURE (PREDICTED)
ADMIXTURE (PREDICTED)
ADMIXTURE (PREDICTED)

Based on these comparisons, it can be observed that, in most cases, the admixture proportions inferred based on euclidean optimisation of the predicted principal components of each individual's genome were concordant with those estimated based on pedigree analysis and admixture estimates from consumer DNA testing companies. Although a negligible amount of false positives (Individual 3) and false negatives (Individual 5) were observed on an individual level for the admixture proportions inferred by euclidean optimisation of the predicted principal components of the individuals in question's genomes, it is unlikely that these affected the results as a whole, as they more than likely nullified each other.

The information above suggests that:

1. The overall findings of this study in regards to the patterns of European and non-European Admixture in the Afrikaner population of South Africa and its diaspora are correct.
2. The findings of this study in regards to the ubiquity of non-European admixture in the Afrikaner population of South Africa and its diaspora are correct.
3. Genealogical registers which are accurate ought to reflect the findings of this study in regards to the patterns of European and non-European admixture. Inaccurate genealogical registers ought not to reflect these findings.

CONCLUSION

After more than three centuries since their ethnogenesis, the Afrikaner population of South Africa and its diaspora exhibits neither stability nor uniformity. They are characterised by a high degree of endogamy; however, they remain heterogeneous, displaying a complex interaction of health and illness, purity and admixture, simultaneously. The sundry origins of their European ancestry, combined with non-European influences in certain individuals, have not effectively mitigated the consequences of their limited founder population and endogamous practices. This endogamy has conferred both advantages and disadvantages concerning heritable diseases, and in some instances, show entirely negligible impact. While the majority of their genetic traits align with those typical of European populations, there are notable divergences that reflect distinct founder effects, and in a minority of cases, evidence of potential extra-European genetic contributions. The Afrikaner population is predominantly of European descent, primarily Dutch, followed by French and German ancestry. Although a significant proportion of Afrikaners studied exhibited minor signals of non-European admixture, such occurrences were not in any manner ubiquitous.

ACKNOWLEGEMENTS

I thank "Alan Hampton" and Xander "Esterházy" for their ethical and academic oversight. I thank Andrei Kharchenko for assisting with the interpretation of the results regarding health profiles and heritable disorders prevalent among the Afrikaners, genetic traits prevalent among the Afrikaners and the biomarker predispositions of the Afrikaners. I thank BA MA "Cyrus Irani" for assisting with the conduction of PCA on the various raw DNA files generously donated to the study by volunteers, without this, the PC predictor would have been much less accurate and robust, putting a large part of the study into jeopardy. I thank all the former for editing and reviewing the study and its findings. I thank every individual included in this study who made their genetic data public on GEDmatch. Without your privacy choice, this study would have been impossible. Finally, I thank the generous volunteers who donated their genomic information not publicly disclosed on GEDmatch to the study. Without your donation, the scope and accuracy of this study would have been severely limited.

REFERENCES

1. https://www.geni.com/projects/South-Africa-Founders-Effect/18512
2. Botha MC, Beighton P. Inherited disorders in the Afrikaner population of southern Africa. Part I. Historical and demographic background, cardiovascular, neurological, metabolic and intestinal conditions. S Afr Med J. 1983 Oct 8;64(16):609-12. PMID: 6226121.
3. Botha MC, Beighton P. Inherited disorders in the Afrikaner population of southern Africa. Part II. Skeletal, dermal and haematological conditions; the Afrikaners of Gamkaskloof; demographic considerations. S Afr Med J. 1983 Oct 15;64(17):664-7. PMID: 6414096.
4. Wallace, Stephanie E, Resources for Genetics Professionals — Genetic Disorders Associated with Founder Variants Common in the Afrikaner Population, 2022, https://www.ncbi.nlm.nih.gov/books/NBK583036/
5. Geldenhuys G, Glanzmann B, Lombard D, Boolay S, Carr J, Bardien S. Identification of a common founder couple for 40 South African Afrikaner families with Parkinson's disease. S Afr Med J. 2014 May 12;104(6):413-9. doi: 10.7196/samj.7747. PMID: 25214250.
6. Sifunda S, Mbewu AD, Mabaso M, Manyaapelo T, Sewpaul R, Morgan JW, Harriman NW, Williams DR, Reddy SP. Prevalence and Psychosocial Correlates of Diabetes Mellitus in South Africa: Results from the South African National Health and Nutrition Examination Survey (SANHANES-1). Int J Environ Res Public Health. 2023 May 12;20(10):5798. doi: 10.3390/ijerph20105798. PMID: 37239526; PMCID: PMC10218408.
7. Delport R. Familial hypercholesterolaemia in South Africans: tracking findings and developments over time - with reference to : prevalence of hypercholesterolaemia in young Afrikaners with myocardial infarction. Ischaemic heart disease risk factors. Cardiovasc J Afr. 2009 Jan-Feb;20(1):18-22. PMID: 19287810; PMCID: PMC4202176.
8. Erasmus, J. Christoff, Klingenberg, Anton, & Greeff, Jaco M.. (2015). Allele frequencies of AVPR1A and MAOA in the Afrikaner population. South African Journal of Science, 111(7-8), 1-6. https://dx.doi.org/10.17159/SAJS.2015/20150074
9. Hollfelder, N., Erasmus, J.C., Hammaren, R. et al. Patterns of African and Asian admixture in the Afrikaner population of South Africa. BMC Biol 18, 16 (2020). https://doi.org/10.1186/s12915-020-0746-1
10. Colenbrander, HT, De afkomst der Boeren. The Hague: M. Nijhoff, 1902. https://archive.org/details/deafkomstderboer00cole
11. Philpott DM. Analysis of the heritage and genetic diversity of influential Afrikaners. Pretoria: University of Pretoria; 2012.
12. Greeff JM. Deconstructing Jaco: genetic heritage of an Afrikaner. Ann Hum Genet. 2007 Sep;71(Pt 5):674-88. doi: 10.1111/j.1469-1809.2007.00363.x. Epub 2007 May 22. PMID: 17521310.
13. Erasmus, JC, 2013. https://web.archive.org/web/20200311154208/https://genza.org.za/index.php/en/?option=com_content&view=article&id=360%3Agenetic-heritage-mtdan-and-ychromosome&catid=129%3Antvl-branch-activities
14. De Villiers, CC, & Theal, GM, 1893, Geslacht-register der oude Kaapsche familien (Vol. 1: A-J). Kaapstad: Van de Sandt de Villiers & Co. Retrieved from University of Pretoria, Library Services.
15. De Villiers, CC, & Theal, GM, 1893, Geslacht-register der oude Kaapsche familien (Vol. 2: A-O). Kaapstad: Van de Sandt de Villiers & Co. Retrieved from University of Pretoria, Library Services.

16. De Villiers, CC, & Theal, GM, 1893, Geslacht-register der oude Kaapsche familien (Vol. 3: P-Z). Kaapstad: Van de Sandt de Villiers & Co. Retrieved from University of Pretoria, Library Services.
17. Van Aswegen, HJ, 1990. History of South Africa to 1854. Academica.
18. Heese, JA, (1972). Die herkoms van die Afrikaner, 1657-1867. South Africa: A. A. Balkema.
19. De Bruyn, CFC, 1976, Die samestelling van die Afrikaner Volk, Tydskrif vir Geesteswetenskappe.
20. Botha MC. Blood group gene frequencies. An indication of the genetic constitution of population samples in Cape Town. The American journal of roentgenology, radium therapy, and nuclear medicine. 1972 May;115(1):1-27.
21. Khan, R, 2017, https://www.gnxp.com/WordPress/2017/09/22/the-non-european-ancestry-of-afrikaners/
22. Khan, R, 2018, https://www.gnxp.com/WordPress/2018/09/12/the-genetics-of-afrikaners-again/
23. https://www.theapricity.com/forum/showthread.php?323633-23andme-prediction-ancestry-tool
24. https://youtu.be/E36s18RkWPg
25. Henn BM, Hon L, Macpherson JM, Eriksson N, Saxonov S, Pe'er I, Mountain JL. Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. PLoS One. 2012;7(4):e34267. doi: 10.1371/journal.pone.0034267. Epub 2012 Apr 3. PMID: 22509285; PMCID: PMC3317976.
26. Carmi S, Hui KY, Kochav E, Liu X, Xue J, Grady F, Guha S, Upadhyay K, Ben-Avraham D, Mukherjee S, Bowen BM, Thomas T, Vijai J, Cruts M, Froyen G, Lambrechts D, Plaisance S, Van Broeckhoven C, Van Damme P, Van Marck H, Barzilai N, Darvasi A, Offit K, Bressman S, Ozelius LJ, Peter I, Cho JH, Ostrer H, Atzmon G, Clark LN, Lencz T, Pe'er I. Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. Nat Commun. 2014 Sep 9;5:4835. doi: 10.1038/ncomms5835. PMID: 25203624; PMCID: PMC4164776.
27. Norval M, Wright CY. The Epidemiology of Cutaneous Melanoma in the White and Black African Population Groups in South Africa. In: Ward WH, Farma JM, editors. Cutaneous Melanoma: Etiology and Therapy [Internet]. Brisbane (AU): Codon Publications; 2017 Dec 21. Chapter 2. Available from: https://www.ncbi.nlm.nih.gov/books/NBK481848/ doi: 10.15586/codon.cutaneousmelanoma.2017.ch2
28. Wright, C. Y., Kapwata, T., & Albers, P. N. (2023). Melanoma crude incidence rates among white South Africans climbing as in other countries: An urgent call for targeted skin cancer prevention and awareness campaigns. South African Medical Journal, 113(3). https://hdl.handle.net/10520/ejc-m_samj_v113_n3_a4
29. Dlova, N.C., Gathers, R., Tsoka-Gwegweni, J., & Hift, R.J. (2018). Skin cancer awareness and sunscreen use among outpatients of a South African hospital: need for vigorous public education. South African Family Practice, 60(4), 132–136. https://doi.org/10.1080/20786190.2018.1426900.
30. https://selfdecode.com/app/snp/rs…/
31. https://gnomad.broadinstitute.org/
32. https://www.snpedia.com/
33. https://www.ncbi.nlm.nih.gov/snp/rs…/
34. Francalacci, Paolo & Sanna, Daria. (2008). History and geography of human Y-chromosome in Europe: A SNP perspective. Journal of anthropological sciences = Rivista di antropologia : JASS / Istituto italiano di antropologia. 86. 59-89.
35. https://www.chavonnesbattery.co.za/blog/post/people-at-the-chavonnes-battery/
36. https://ghil.hypotheses.org/23

37. Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. Am J Hum Genet. 2015 Jan 8;96(1):37-53. doi: 10.1016/j.ajhg.2014.11.010. Epub 2014 Dec 18. PMID: 25529636; PMCID: PMC4289685.
38. Mitchell SL, Goodloe R, Brown-Gentry K, Pendergrass SA, Murdock DG, Crawford DC. Characterization of mitochondrial haplogroups in a large population-based sample from the United States. Hum Genet. 2014 Jul;133(7):861-8. doi: 10.1007/s00439-014-1421-9. Epub 2014 Feb 1. PMID: 24488180; PMCID: PMC4113317.
39. Lao O, Vallone PM, Coble MD, Diegoli TM, van Oven M, van der Gaag KJ, Pijpe J, de Knijff P, Kayser M. Evaluating self-declared ancestry of U.S. Americans with autosomal, Y-chromosomal and mitochondrial DNA. Hum Mutat. 2010 Dec;31(12):E1875-93. doi: 10.1002/humu.21366. PMID: 20886636; PMCID: PMC3051415.
40. Kayser M, Brauer S, Schädlich H, Prinz M, Batzer MA, Zimmerman PA, Boatin BA, Stoneking M. Y chromosome STR haplotypes and the genetic structure of U.S. populations of African, European, and Hispanic ancestry. Genome Res. 2003 Apr;13(4):624-34. doi: 10.1101/gr.463003. PMID: 12671003; PMCID: PMC430174.
41. https://southafrica-info.com/infographics/infographic-people-south-africa-age-sex-race/
42. Taboga, Marco (2022). "SimpleR: multiple linear regression calculator", StatLect. https://www.statlect.com/fundamentals-of-statistics/SimpleR.
43. Kharchenko, Andrei, (2023) "Trait Predictor", AndreiDNA, https://andreikharchenko1.github.io/Trait-Predictor/
44. https://www.worldatlas.com/articles/countries-with-the-most-blue-eyed-people.html
45. Katsara, Alexandra & Nothnagel, Michael. (2019). True colors: A literature review on the spatial distribution of eye and hair pigmentation. Forensic Science International: Genetics. 39. 10.1016/j.fsigen.2019.01.001.
46. https://www.colorwithleo.com/what-hair-color-is-most-common-in-germany/
47. https://worldpopulationreview.com/country-rankings/blonde-hair-percentage-by-country
48. https://worldpopulationreview.com/country-rankings/percentage-of-redheads-by-country
49. https://en.wikipedia.org/wiki/Blood_type_distribution_by_country