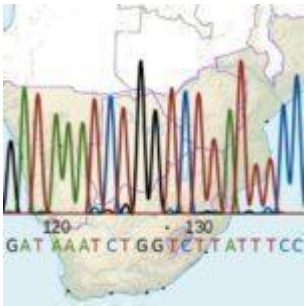
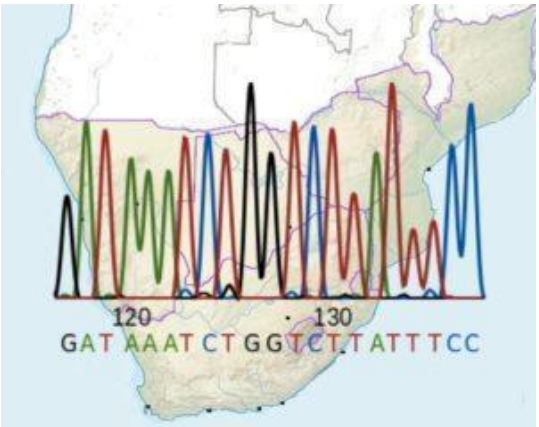
SAHGP Pilot: Southern Africans are more diverse than expected





Southern Africa Map with WGS illustration (adapted from Southern Africa Map B1 (Jon Richfield) and DNA Sequence electrogram (Sjef), Wikimedia Commons)

Genomic medicine is a fast-growing medical discipline that relies on genetic information for clinical care. African populations are the most genetically diverse population and suffer the highest health burden in the world. However, genome studies of disease association rarely include or are conducted on African populations, limiting the practical application of genomic medicine in Africa.

The Southern African Human Genome Programme (SAHGP) is a

multi-disciplinary initiative that aims at determining the genetic diversity of different ethnographic groups in Southern Africa. The SAHGP conducted a pilot study that aimed at studying genetic differences between some of the major ethnolinguistic groups of South Africa. They utilised deep Whole Genome Sequencing (WGS) to provide an unbiased estimate of genetic variation among self-identified Southeastern Bantu# (SEB) and Coloured* (COL) individuals.

Choudhury *et al.* identified greater than 800,000 novel single nucleotide variants (SNV) that were absent from currently available genomic databases. Major SNV density differences observed between SEB and COL were associated with protein-coding regions. Additionally, high genetic differences in genes which include *SEMA4F, EREG, PLN* and *PTF1A*, were observed between self-identified Sotho and Xhosa within SEB. Using the Genecards database. Choudhury *et al.* inferred medical implications of genetic diversity of these genes, which included potential association in the development or pathogenesis of prostate, breast, colorectal and pancreatic cancer as well as pulmonary tuberculosis. It should be noted that these are not true associations, due to the absence of variation of genetic phenotype and data on disease incidence and prevalence. This highlights the need for future studies that can elucidate the role of genetic diversity in the genes above and implication on health.

Finally, genetic exploration of the COL identified better proxies for comparing COL ancestry with Southeast and South Asians. Previous studies utilised Chinese and Gujarati ancestry as proxies for Southeast and South Asian ancestry, respectively. Choudhury *et al.* identified Malay (Southeast Asian) and Bengali (South Asian) ancestry as better proxies for studies that aim to determine the contribution of Southeast and South Asian ancestry to COL genetic diversity.

In summary, the SAHGP pilot study represents the first report on genetic diversity using high coverage WGS. Despite the low

sample size, the pilot project is the first genome-scale study that detected significant genomic differences between two ethnicities in SEB.

#SEB ethnicity included mostly individuals of Sotho or Xhosa heritage.
*Coloured ethnicity is an admixed population of individual with European, Bantu-speaking African, Khoesan, Southeast Asian and South Asian ancestry.

Journal Article: Choudhury *et al.* 2017. Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. Nature communications.

Article by Cheleka AM Mpande