Ibrahim and Salih *Human Genome Variation* (2021) 8:11
https://doi.org/10.1038/s41439-021-00141-1

Human Genome Variation

DATA REPORT Open Access

The Y chromosome ancestry marker R1b1b2: a surrogate of the SARS-CoV-2 population affinity

Muntaser Ibrahim[1] and Abdalhameed Salih[1]

Abstract
Individual and population susceptibilities to disease remain a murky area of investigation, clouded by past bias based on ideological differences and wars. The current SARS-CoV-2 pandemic, the largest in living memory, brought this matter to forefront as the disparity in disease burden became apparent. A timeline analysis of the pandemic revealed the presence of country clusters that display a marked preponderance of disease among populations carrying the ancestry marker R1b1b2, notably associated with both infection and mortality. This marker is a relic of past human expansions from western Asia and subsequently Europe and the rest of the world, which may have been accompanied by peculiar biological events rendering these populations vulnerable to SARS-CoV-2.

During epidemics, the human phenotypes of interest are infection, clinical disease, morbidity, and especially death. The current coronavirus disease (SARS-CoV-2) pandemic is associated with this range of phenotypes and the existence of marked human clusters of infection and particularly death. Populations, people within populations, and individuals obviously vary in their propensity to develop clinical symptoms, including death. In fact, epidemics without associated mortality are of little epidemiological value and in many cases may pass unnoticed. Populations are products of genetic history, that is, the history of a group's genes upon interaction with various environments, including viral onslaughts and other environmental and genetic interactions[1]. There is also the history of the population itself, including migration, admixture, culture, and other non-biological determinants.

Genetic backgrounds pertain mainly to ancestry. When seeking markers of ancestry, the choice is usually either mitochondrial DNA or the Y chromosome, as they are both spared from the shuffling impact of recombination. However, given the unique history of males within the context of male-driven migration, they make a better marker of population structure and global human ancestry clusters[2–4].

Populations as evolutionary units were rather cumbersome to define in the post-World War II era, in terms of both health and disease. In many cases, this was aggravated by a lingering tendency to adopt archaic terms for human groupings (Caucasoid, Mongoloid, and Negroid) mostly based on race myths. Entering the era of chronic and noninfectious diseases, also in the aftermath of World War II, heralded a departure from universal paradigms of health into one that accommodates human differences, currently best demonstrated in the arena of pharmacogenomics. Infectious diseases, however, remained aloof to these new approaches, as pathogens are widely believed to affect humans equally, despite the existence of human disease clusters, familial differences, and traits. The current SARS-CoV-2 pandemic raises several questions concerning the possible contribution of human genetic variation to the epidemiology of the disease and the extent to which it could explain the current disparities in disease burden across countries and communities.

Data from the current SARS-CoV-2 pandemic (www.worldmeters.info) for 95 countries as of 29 April, 31 May, 25 June, and 1 July 2020 (Supplementary tables: 1,2,3,4) were ranked by the number of deaths and the fatality rate using data derived from the World Meter Coronavirus

Correspondence: Muntaser Ibrahim (mibrahim@iend.org)
[1]Institute of Endemic Diseases, University of Khartoum, Medical Campus Qasser Street, Khartoum, Sudan

update (www.worldmeters.info). Explaining complex phenomena such as the current pandemic in the biological and environmental dimensions is a cumbersome task and may often require equally complex analytical models. Regression analysis is perhaps the most convenient tool to capture the relative weight of multiple contributors to a complex phenomenon. As the main disease clusters included countries that are mainly in temperate and cold areas, suggesting temperature as an environmental factor potentially contributing to the epidemiology of the

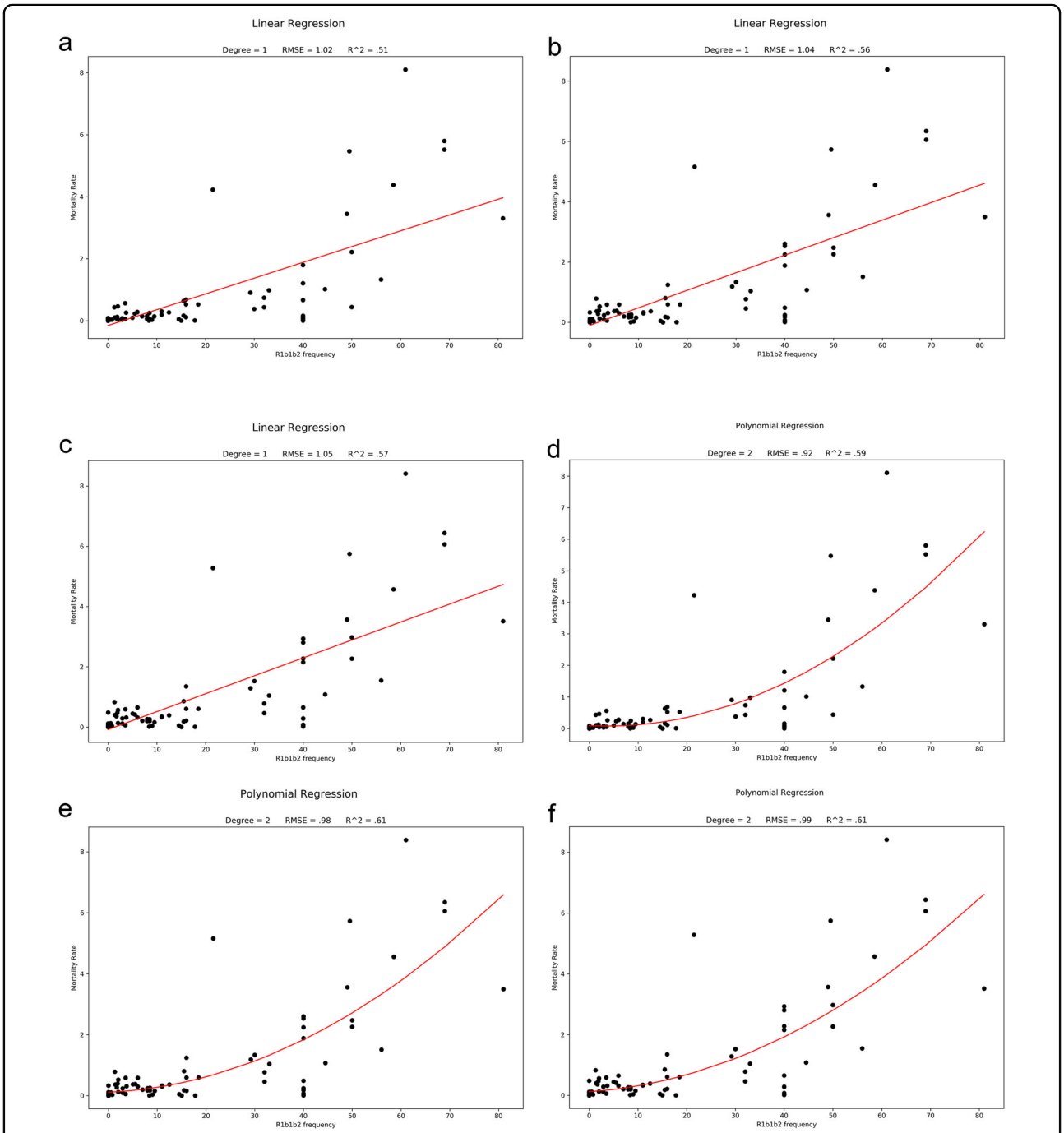

Fig. 1 The correlation between the independent variable R1b1b2 and mortality/cases in the current SARS-CoV-2 pandemic as the dependent variable tested using different regression models on 31 May, 25 June and 1 July. : a, b and c are simple linear regression analyses that preliminarily gave an R^2 = 0.57, p ≤ 0.001 with a cluster at the bottom left of the figure depicting countries where the pandemic has a smaller impact in terms of mortality/cases. Better fitting was achieved with a regression polynomial distribution (d, e and f) (R^2 = 0.61, p ≤ 0.001).

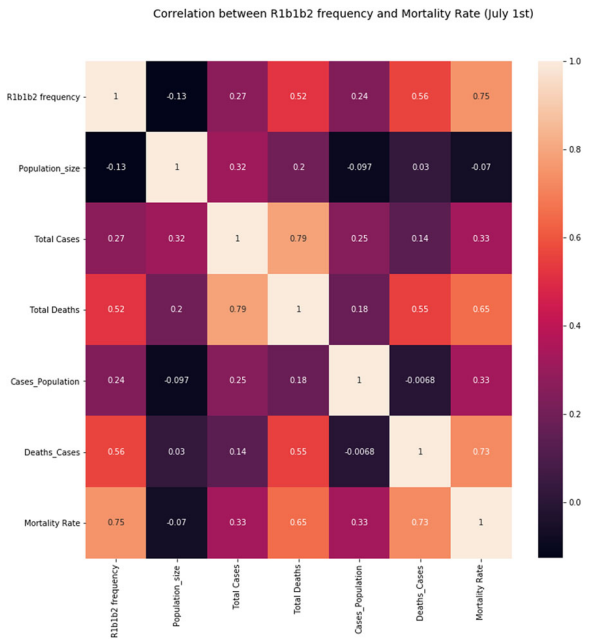

Fig. 2 Correlation Heatmap. Matrix showing correlation between mortality rate, population size, total deaths, total cases, cases to population, deaths to cases, and frequency of R1b1b2 as of July first 2020.

outbreak, multiple regression analysis of temperature readings against disease records by country from *World Bank Data Catalog* was carried out, and results are shown below. Given the current viral diversity and strain variations, neither temperature nor viral diversity seems to be able to adequately explain the current global trends in the pandemic. One of the foremost candidates naturally becomes the genetic background of the populations in the affected countries. To examine the relationship between ancestry and the clinical outcome of SARS-CoV-2, several ancestry markers were investigated, and the most prevalent and informative were included in subsequent analysis of populations affected by the pandemic. Haplotype frequencies were obtained from Y-chromosome haplogroup databases (https://en.wikipedia.org/wiki/Y-DNAhaplogroups_in_populations/ http://thegeneticatlas.com/World_Y-DNA.htm) and the available literature[2,5–8]. For correlation and regression analysis, Python Software in the Statistical Package was used, Python Language Reference, version 3 (available at http://www.python.Org).

Among the 33 countries with the highest number of cases and deaths on 31 April (10,000 cases or more and an average death rate of 6.55% up to a staggering 16%, 23 are countries with populations that have a predominance of the Y chromosome haplotype R1b1b2 (M269)). The haplotype is believed to have originated in western Asia, specifically around the present area of Iran, which was coincidentally one of the primary foci for the first high fatality numbers in the pandemic. In the 33 countries recording between a thousand and 10,000 cases and a mortality index of 3.6, the R1b1b2 haplotype was still up to 40% in 22 countries, which was seen as a result of recent European expansion into the Americas and Northern Africa. Among thirty countries where case numbers were below a thousand and the mortality index was 3.1, most are in Africa, Asia, or South America, with lower frequencies of R1b1b2 that reach 0% in 11 countries. Overall, linear regression based on normal, Poisson, and polynomial distribution models for the correlation of R1b1b2 with both infection and mortality had the following values: 0.68 and 0.47, respectively. The coefficient for death to cases with R1b1b2 was 1.0927, $p \leq 0.001$ when adjusted for country population size. Based on multiple linear regression, the coefficient for death to cases with temperature was −0.008, $p \leq 0.01$ (Supplementary Fig. 1) and with R1b1b2 was 0.0359, with $p \leq 0.001$ when adjusted for country population size (Figs. 1 and 2).

The association with R1b1b2 is obviously noncausal, although recently collected evidence suggests that the Y chromosome influences immune and inflammatory responses in men, translating into genetically programmed susceptibility to diseases with a strong immune component[9]. However, the link in this analysis remains a tag for population history. Interestingly, such a tagging approach was reported for the risk of coronary artery disease by Charchar and his colleagues, who studied 3233 biologically unrelated British men to investigate heart disease risk; an odds ratio of 1.75 (95% CI 1.20–2.54, $p \leq 0.05$) was reported between coronary artery disease and men carrying the haplotype R1b1b2[2]. We adopted a similar tagging haplotype approach earlier during the quest to understand the genetic basis of susceptibility to visceral leishmaniasis (VL) and to reveal hidden structures within seemingly homogenous populations[3]. A stratification of the target population based on Y chromosome haplogroups increased the likelihood of linkage to an impressive LOD score of 5.656, defining a susceptibility locus associated with carriers of haplogroup A1b1b2b formerly A-M13 A3b2; interestingly, this is the most common Y chromosome haplogroup among Nilotics of southern Sudan and other Nilo-Saharan speakers, who happen to be the most vulnerable population to VL in East Africa and worldwide[10].

Population markers are not necessarily a reflection of ethnic identity. For example, African American, Caribbean and South American ethnic populations were found to carry substantial amounts of the R1b1b2 haplotype and its subclades. In some instances, the Y chromosome of Europeans has almost completely replaced the native male contribution to the gene pool[11].

The events took place in time and space for ancestral carriers of R1b1b2 that rendered them vulnerable to SARS-CoV-2 infection and its severe pathology remain to be studied. Possible clues might be collected from the analysis of genomes/exomes of patients expressing the clinical outcome of such susceptibility loci. These loci might prove to be present at similar or higher frequencies in other populations that are phylogenetically divergent yet susceptible but show a high mortality phenotype, such as Indonesia and the Philippines, as outliers with no reported frequencies of R1b1b2. Even there, fatality was significantly lower when the ratio was corrected against the country population size. Similarly, in countries such as Japan and South Korea, mortality was 1.94 and 2.2, respectively, even without correction for the country population size. Interestingly, these ratios remained approximately the same, and the correlation with R1b1b2 even grew stronger over time when the same dataset was analyzed at the end of May, 25th of June, 1st of July and 18th of November (Supplementary Figure 2). Among European countries, a linear positive correlation was found between R1b-S116 allele frequency and basic reproduction numbers. Evidently, disease burden does not vary only between continents, countries and regions in correlation with the average frequency of R1b1b2; even cities may become hot spots due to the unique history of human settlement, as noticed in northern Italy, where the epidemic is more intense than in southern Italy[12].

The claim that the SARS-CoV-2 pandemic has yet to take its toll in Africa is largely speculative if not unfounded. Whether Africa's population demographic pyramid, climate, degree of crowdedness, genetic background, natural or cross immunity and other confounders/contributors might have hindered or slowed the progression of SARS-CoV-2 remains to be studied. Anecdotes and forecasts of a pandemic on the loose with some countries peaking, and others waiting to strike, are embedded in a universal health paradigm that does not cater to human variations in disease susceptibility, a fact modern biomedical sciences is recognizing and embracing progressively and more than ever.

Acknowledgements
We are grateful to Professor Melanie Newport and Dr. Shahd Osman for reading the manuscript and providing insightful remarks.

Conflict of interest
The authors declare that they have no conflict of interest.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41439-021-00141-1.

Received: 17 October 2020 Revised: 21 December 2020 Accepted: 22 December 2020.
Published online: 18 February 2021

References

1. Ali Albsheer, M. M. et al. The Duffy T-33C is an insightful marker of human history and admixture. *Meta Gene* **26**, 100782 (2020).
2. Charchar, F. J. et al. Inheritance of coronary artery disease in men: an analysis of the role of the Y chromosome. *Lancet* **379**, 915–922 (2012).
3. Miller, N. et al. Y chromosome lineages tag village specific genes on chromosomes 1p22 an 6q27 that control leishmaniasis in the Sudan. *PLoS Genet. 1* **3**, e71 (2007).
4. Underhill, P. A. et al. Thephylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann. Hum. Genet* **65**, 43snd–21 (2001).
5. Bekada, A. et al. Introducing the Algerian Mitochondrial DNA and Y-Chromosome Profiles into the North African Landscape". *PLoS ONE* **8**, e56775 (2013).
6. Distribution of European Y-chromosome DNA (Y-DNA) haplogroups by country in percentage. [Online]. Available:https://www.eupedia.com/europe/european_y-dna_haplogroups.shtml. [Accessed: 5-Jul-2020]Eupedia, 2017c.
7. Gaviria, A. et al. Characterization and Haplotype analysis of 11 Y-STR loci in Ecuadorian population. *Forensic Sci. Int.: Genet. Suppl. Ser.* **4**, e310–e311 (2013).
8. Ribeiro, J. et al. Male lineages in Brazilian populations and performance of haplogroup prediction tools. Forensic Science. *Int.: Genet.* **44**, 102163. 10.10 (2019).
9. Maan, A. et al. The Y chromosome: a blueprint for men's health? *Eur. J. HumGenet.* **25**, 1181–1188 (2017).
10. Ibrahim, M. E. The epidemiology of visceral leishmaniasis in east Africa: hints and molecular revelations. In Molecular Epidemiology of Parasitic diseases. Edited by J Alvar and DC Barker. *Trans. R. Soc. Trop. Med Hyg.* **96**, S25 (2002). Suppl 1.
11. Rojas, W. et al. Genetic makeup and structure of Colombian populations by means of uniparental andparental DNA markers. *Am. J. Phys. Anthropol.* **143**, 13–20 (2010).
12. Delanghe, J. R. et al. The potential influence of human Y-chromosome haplogroup on COVID-19 prevalence and mortality. *Ann. Oncol.* **31**(Aug), 1582–1584, https://doi.org/10.1016/j.annonc.2020.08.2096 (2020). Epub ahead of print. PMID: 32835812; PMCID: PMC7442561.