medRxiv preprint doi: https://doi.org/10.1101/2020.04.05.20054627; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Investigatingthe likely association between genetic ancestry and COVID-19 manifestation

Ranajit Das* and Sudeep D. Ghate

Yenepoya Research Centre, Yenepoya (Deemed to be University), Mangalore, Karnataka, India.

*All correspondence to

Dr. Ranajit Das

Yenepoya Research Centre

Yenepoya (Deemed to be University)

University Road, Deralakatte

Mangalore 575018

Karnataka, India

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.05.20054627; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Abstract

The novel coronavirus 2019-nCoV/SARS-CoV-2 infection has shown discernible variability across the globe. While in some countries people are recovering relatively quicker, in others, recovery times have been comparatively longer and numbers of those succumbing to it high. In this study, we aimed to evaluate the likely association between an individual's ancestry and the extent of COVID-19 manifestation employing Europeans as the case study. We employed 10,215 ancient and modern genomes across the globe assessing 597,573 single nucleotide polymorphisms (SNPs). Pearson's correlation coefficient (r) between various ancestry proportions of European genomes and COVID-19 death/recovery ratio was calculated and its significance was statistically evaluated. We found significant positive correlation (p=0.03) between European Mesolithic hunter gatherers (WHG) ancestral fractions and COVID-19 death/recovery ratio and a marginally significant negative correlation (p=0.06) between Neolithic Iranian ancestry fractions and COVID-19 death/recovery ratio. We further identified 404 immune response related single nucleotide polymorphisms (SNPs) by comparing publicly available 753 genomes from various European countries against 838 genomes from various Eastern Asian countries in a genome wide association study (GWAS). Prominently, we identified that SNPs associated with Interferon stimulated antiviral response, Interferon-stimulated gene 15 mediated antiviral mechanism and 2′-5′ oligoadenylate synthase mediated antiviral response show large differences in allele frequencies between Europeans and East Asians. Overall, to the best of our knowledge, this is the first study evaluating the likely association between genetic ancestry and COVID-19 manifestation. While our current findings improve our overall understanding of the COVID-19, we note that the development of effective therapeutics will benefit immensely from more detailed analyses of individual genomic sequence data from COVID-19 patients of varied ancestries.

Key words: Genomic ancestry, COVID-19 manifestation, novel coronavirus, Genome Wide Association study (GWAS)

Introduction

The novelcoronavirus 2019-nCoV/SARS-CoV-2, was officially identified on 7th January 2020 in Wuhan, China [1]. Since then, the virus has spread rapidly throughout the world

medRxiv preprint doi: https://doi.org/10.1101/2020.04.05.20054627; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

leading to catastrophic consequences. Notably, COVID-19, caused by 2019-nCoV has shown significant worldwide variability, especially in terms of the death/recovery ratio that pertains to a ratio of the number of deaths caused by the novel coronavirus to numbers of infected individuals recovered within the same time interval. While infected individuals from some countries have recovered relatively quickly with lower morbidities, those from some other parts of the world appear to remain affected for longer with slower recovery times and demonstrate relatively higher death/recovery ratios [2].This is suggestive of a likely population level genetic variation in terms of susceptibility to the coronavirus and COVID-19 manifestation.

A case in point in this context appears to be *ACE2* that encodes for angiotensin-converting enzyme-2 and has been speculated to be the host receptor for the novel corona virus [3,4]. Recently, Cao et al. analysed 1700 variants in *ACE2* from China MAP and 1000 Genomes Project databases [5].While they reported the absence of natural resistant mutations for coronavirus S-protein binding across their study populations, their study revealed significant variation in allele frequencies (AF) among the populations assessed, including between people of East Asian and European ancestries. They speculated that the differences in AFs of *ACE2* coding variants, likely associated with its elevated expression may influence *ACE2* function and putatively impact SARS infectivity among the evaluated populations. Additionally, a majority of the 15 eQTL variants they identified had discernibly higher AFs among East Asian populations compared to Europeans, which may suggest differential susceptibility towards coronavirus in these two populations under similar conditions.

In addition to *ACE2*, COVID-19 manifestation may also be modulated by several immune response modulating genes. Li et al. (2020) mentioned that coronavirus infection causes secretion of large quantities of chemokines and cytokines (such as IL□1, IL□6, IL□8, IL□21, TNF□β and MCP□1) in the infected cells, additionally, the host's major antiviral machinery comprising of Interferons (IFNs) may also the limit of coronavirus [6]. Overall it can be speculated that variations in host genes encoding for interleukins and interferons may be responsible for differential manifestation of COVID-19.

Further recent studies have shown that variants associated with immune responsiveness display significant population-specific signals of natural selection [7]. In particular, the authors showed that admixture with Neanderthals may have differentially shaped the immune systems in European populations, introducing novel regulatory variants, which may affect preferential responses to viral infections. European genomes were further

medRxiv preprint doi: https://doi.org/10.1101/2020.04.05.20054627; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

modulated and distinguished owing to multiple waves of migration throughout their ancient past. We note here that while Neolithic European populations predominantly descended from Anatolian migrants, European ancestry shows a distinct West–East cline in indigenous European Mesolithic hunter gatherers (WHGs) [8]. At the same time Eastern Europe was in the rudimentary stages of the formation of Bronze Age steppe ancestry, which later spread into Central and Northern Europe through East-West expansion [9]. Further, high genetic substructure between North-Western and South-Eastern hunter-gatherers has also been documented recently [10].

Ancestry specific variation in immune responses and its potential genetic and evolutionary determinants are poorly understood especially in the context of the novel coronavirus infection and COVID-19 manifestation.While COVID-19 has affected huge numbers of individuals worldwide, this pandemic has cause catastrophic damage in Europe. Here we aimed to unravel the potential association between genetic ancestry and COVID-19 employing Europeans as the case study. In addition, using a genome wide association study (GWAS) approach we sought to discern the putative SNP markers and genes that may be likely associated with differential novel coronavirus infection by comparative analyses of the European and East Asian genomes.To this end we employed 10,215 ancient and modern genomes, obtained from the databank of Dr. David Reich, Harvard Medical School, USA (https://reich.hms.harvard.edu/datasets) to evaluate the likely correlation between European ancestry and COVID-19. Our findings revealed significant positive correlation between WHG-related ancestry and COVID-19 death/recovery ratio and marginally significant negative correlation between Neolithic Iranian ancestry and COVID-19. Finally, we identified several SNPs that may influence immune responsiveness and displaying significant differential AF between Europeans and East Asians. Our current results shine light on ancestry driven genetic factors underlying variabilities in COVID-19 infection and manifestation; it strongly advocates concerted sequencing/genotyping of COVID-19 affected patients worldwide not only to expand and substantiate our current knowledge but to develop focused population specific therapeutic approaches to mitigate this worldwide challenge.

Method

Dataset

medRxiv preprint doi: https://doi.org/10.1101/2020.04.05.20054627; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Genome data for the current study was obtained from the personal database of Dr. David Reich's Lab Harvard Medical School, USA (https://reich.hms.harvard.edu/datasets). The final dataset comprised of 10,215 ancient and modern genomes across the globe assessing 597,573 single nucleotide polymorphisms (SNPs). File conversions and manipulations were performed using EIG v7.2 [11] and PLINK v1.9 [12].

COVID-19 data was obtained from 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository by Johns Hopkins CSSE [2]. The data available in this web portal as of Sunday, 5th April 2020, 6 PM was used in the analyses. Indian data was obtained from Government of India coronavirus portal (https://www.mygov.in/corona-data/covid19-statewise-status).

Determination of ancestry proportions in European genomes in a global context

We used *qpAdm* [13] implemented in AdmixTools v5.1 [14,15] to estimate ancestry proportions in the European genomes originating from a mixture of 'reference' populations by utilizing shared genetic drift with a set of 'outgroup' populations. 14 ancient genomes namely *Luxembourg_Loschbour.DG*, *Luxembourg_Loschbour_published.DG*, *Luxembourg_Loschbour*, *Iberia_HG* (N=5), *Iberia_HG_lc*, *Iberia_HG_published*, *LaBrana1_published.SG*, *Hungary_EN_HG_Koros*(N=2), *Hungary_EN_HG_Koros_published. SG* were grouped together as European Mesolithic hunter gatherers (WHGs); three ancient genomes namely *Russia_EHG*, *Russia_HG_Samara* and *Russia_HG_Karelia* (N=2) were grouped together as Mesolithic hunter-gatherers of Eastern Europe (EHGs); two ancient genomes: *KK1.SG*, *SATP.SG* were grouped as Caucasus hunter-gatherers (CHGs) and *Sweden_HG_Motala* (N=8) were renamed as Scandinavian Hunter-Gatherers (SHGs) in *qpAdm* analysis. After repeating the analyses with several population combinations, we inferred that Europeans could be best modelled as a combination of three source populations namely EHGs, WHGs and Neolithic Iranians (*Iran_GanjDareh_N*) as *Left* (EHG, WHG,Iran_GanjDareh_N). We used a mixture of eight ancient and modern-day populations comprising of *Ust Ishim*, *MA1*, *Kostenki14*, *Han*, *Papuan*, *Chukchi*, *Karitiana*, *Mbuti*asour '*Right*' outgroup populations (O8).

Pearson's correlation coefficient (r) between various ancestry proportions of European genomes and COVID-19 death/recovery ratio was calculated and its significance was statistically evaluated using GraphPad Prism v8.4.0, GraphPad Software, San Diego, California USA, www.graphpad.com. Additionally, we developed several linear regression

medRxiv preprint doi: https://doi.org/10.1101/2020.04.05.20054627; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

models with different combinations of WHG ancestry fraction (x_1) alongside climatic factors such as 15 days' mean temperatures (x_2) and 15 days' mean humidity (x_3) of the ten European countries employed in this study to statistically evaluate their impact on COVID-19 death/recovery ratio (y) using *glm* function implemented in R v3.6.3 (Table 1).One-tailed tests were performed considering the alternative hypothesis of positive association of x_1, x_2 and x_3 with y ($H_A: \beta > 0$). The best model was determined by likelihood ratio test (LRT), employing *lrtest* function, implemented in R package '*epicalc*'. Temperature and humidity data were obtained from Time and Date.com (https://www.timeanddate.com/), which garners meteorological data from leading meteorological institutes such as The World Meteorological Organization (https://public.wmo.int/en) and MADIS (https://madis.ncep.noaa.gov/) .

Notably, the death to recovery ratio is used here as the test statistic reflecting the extent to which individuals are succumbing versus recovering following the COVID-19 infection. We surmise this ratio to be indicative of the host immune responsiveness against viral assault. Notably since the death to recovery ratio does not take into account the number of active cases, statistical bias arising due to under-reporting/under-testing of COVID-19 cases is avoided.

Genome-wide association analyses (GWAS)

GWAS was performed to investigate SNPs with significant AF variations among European and East Asian genomes and likely correlated with differential COVID-19 infectivity. To this end 753 European genomes were compared to 838 Eastern Asian genomes. Standard case (higher death/recovery ratio: Europeans) – control (lowerdeath/recovery ratio: East Asians) based association analyses was performed in PLINK v1.9 using --assoc command. A Manhattan plot was generated in Haploview [16] by plotting Chisquare values of all assessed SNPs to identify the SNPs that are likely associated with COVID-19 manifestation. Highly significant SNPs were annotated using SNPnexus web-based server [17].

Results

Ancestry proportions in European genomes and COVID-19 manifestation

medRxiv preprint doi: https://doi.org/10.1101/2020.04.05.20054627; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

We modelled all Europeans as a combination of three source populations namely EHGs, WHGs and Neolithic Iranians (see Methods) in *qpAdm* analysis. Among 10 European populations employed in this study, GBR (British in England and Scotland) genomes were found to have the highest WHG ancestry proportions (26.1%) (Table 1) and the lowest Neolithic Iranian ancestry fractions (51.5%) (Table 2). Notably, among the European populations evaluated, COVID-19 death to recovery ratio till date, is the highest among the British people: death/recovery ratio=31.94. In contrast, Russian and Finnish populations which could not be modelled with WHG ancestry so far appear to have less detrimental COVID-19 manifestation with 0.13 and 0.08 death to recovery ratio respectively. Consistent with this, we found significant positive correlation (Pearson's Correlation; r=0.58; p=0.03) between WHG ancestry fraction and COVID-19 death/recovery ratio and marginally significant negative correlation (Pearson's Correlation; r= -0.52; p=0.06) between Neolithic Iranian ancestry fractions and COVID-19 death/recovery ratio. We did not find any correlation between EHG ancestry fraction and COVID-19 manifestation.

Congruent with correlation results, we found significant linear positive association between WHG ancestry fraction and COVID-19 death/recovery ratio (Model 0) (R^2=0.34, p=0.039). Further, we found significant association between the combination of WHG ancestry and 15 days' mean temperature and COVID-19 death/recovery ratio (Model 4) (R^2=0.39, p=0.042) (Supplementary Table 1). LRT determined Model 0 to be the best model, indicating COVID-19 manifestation can be best explained by variability in WHG ancestry fractions among European populations. Model 0 was found to be highly significantly better than Model 1 (Temperature alone) and Model 2 (Humidity alone) in determining the variability of COVID-19 death/recovery ratio among the study populations, suggesting that climatic factors such as colder temperature and variation in humidityalone cannot explain the variability of COVID-19 manifestation among European countries. However, in combination with peoples' ancestral make-up, temperature disparities can explain the variation in COVID-19 death/recovery ratio.

To evaluate the robustness of our results, we repeated all our analyses with the data available in Johns Hopkins CSSE[2] web portal as of Wednesday, 8th April 2020. Notably, we did not find any discernible difference between the results obtained employing 5th April data and those obtained employing 8th April data.

Genome-wide association analyses

medRxiv preprint doi: https://doi.org/10.1101/2020.04.05.20054627; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

For GWAS, 753 genomes from across Europe with high COVID-19 death/recovery ratio (case) were compared to 838 Eastern Asian genomes with relatively lower COVID-19 death/recovery ratio (control). Out of 597,573 SNPs employed in this study, 385,450 (64.5%) markers revealed highly significant variation (Chi square≥ 6.63; P<0.01) between Europeans and East Asians, indicating the discernible differences in the genetic tapestry of the two populations (Fig. 1). For stringency, we annotated only the top 10,000 ranked SNPs (N=10,013, Chi square≥ 739; $P \leq 9.69 \times 10^{-163}$) using SNP nexus web-based server. Among top 10,000 ranked SNPs, 404 were associated with host immune responsiveness including Interferon (IFN) stimulated antiviral response, Interferon-stimulated gene 15 (*ISG15*) mediated antiviral mechanism and 2′-5′ oligoadenylate synthase (OAS) mediated antiviral response. These SNPs were associated with 161 immune system related genes including those related to innate immunity (eg.*IFI27, IFIH1* and *IFNG*), interleukins (eg.*IL16, IL17RB, IL17RD, IL1RAP, IL4R* and *IL6*), antiviral response (eg.*OAS3, SEC13, MX1, NUP54, TRIM25, NUP88, HERC5, FLNB, SEH1L, EIF4E3* and *USP18*) and receptors (eg.*EDAR*). Furthermore, six SNPs (rs2243250, rs1800872, rs1800896, rs1544410, rs1800629 and rs1805015) that display large AF variability between Europeans and East Asians are likely associated with the development of immune responses in the firstyear of life contingent upon the gene-environment interactions (https://www.snpedia.com). Finally, the SNPs with the highest Chi square values (≥1500; P≈0) were found to be associated with immune system related pathways such as Senescence-Associated Secretory Phenotype (SASP), Interleukin-4 and Interleukin-13 signaling, TNFR2 non-canonical NF-kB pathway andClass I MHC mediated antigen processing & presentation.

Discussion

The novel coronavirus 2019-nCoV/SARS-CoV-2 infection has shown discernible variability across the globe. While in some countries people are recovering relatively quicker, in others, recovery times have been comparatively longer, and numbers of those succumbing are higher.Here we sought to investigate the likely association between genetic ancestry determinants and the extent of COVID-19 manifestation employing Europeans as a case study.

The European story

medRxiv preprint doi: https://doi.org/10.1101/2020.04.05.20054627; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Till date, Europe has been most severely impacted owing to the COVID-19 infection. Our findings revealed asignificant positive correlation between WHG-related ancestry and COVID-19 death/recovery ratio (p=0.03) and marginally significant negative correlation between Neolithic Iranian ancestry and COVID-19 (p=0.06). Further, we found significant linear positive association between WHG ancestry fraction and COVID-19 death/recovery ratio (R^2=0.34, p=0.039). Further, we found significant association between the combination of WHG ancestry and 15 days' mean temperature and COVID-19 death/recovery ratio (R^2=0.39, p=0.042). Previously it has been shown that European genomes evolved uniquely with regards to immune responsiveness, particularly pertaining to responses against viral infections [8]. Admixture with Neanderthals is believed to have introduced unique regulatory variants into European genomes, which likely modulated immune responses in European populations [7].We surmise that the European genome architecture has been extensively modulated by the complex origin and migration history of modern-day Europeans during Paleolithic, Mesolithic and Neolithic periods that in turn contributed via introduction of novel variants in European genomes.

Modern-day European genomic diversity has been thought to be shaped by variable proportions of local hunter-gatherer ancestry (WHGs) [18]. While Neolithic, Iberian genomes (modern day Spanish and Portuguese populations) revealed widespread evidence of WHG admixture (10-27%), Neolithic German populations (Linearb and keramik– LBK culture) revealed ~4–5% of the same [18].Notably the variation in WHG ancestry fractions among German and Spanish people correlated with the observed variabilities in COVID-19 death/recovery ratio in these two countries; German populations with discernably lower WHG fractions have a current death/recovery ratio of 0.05, however, in Spain the ratio is 0.33, which corresponded to high WHG fraction among Spanish populations. Interestingly, the other Iberian country, Portugal, despite reporting fewer numbers of COVID-19 cases compared to neighboring Spain, has a significantly high death to recovery ratio (3.93) underscoring the likely association between high WHG ancestry proportions in Iberians and the severity of COVID-19 infection. The WHG ancestry proportion is discernibly high among the Basque people in Spain (33.9%) and congruent with our hypothesis, the death percentage in Basque country is significantly higher compared to other Spanish provinces (https://www.statista.com/statistics/1102882/cases-of-coronavirus-confirmed-in-spain-in-2020-by-region/). However, owing to unavailability of data we were unable to calculate the death to recovery ratio in Basque people. Nevertheless, taking our findings into consideration, since the WHG fraction was found to be the highest among the Basque people

medRxiv preprint doi: https://doi.org/10.1101/2020.04.05.20054627; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

among all populations assessed in this study, we surmise the death to recovery ratio in Basque country will be higher than Spain's countrywide mean. A similar correlation between WHG ancestry and the extent of COVID-19 manifestation was observed for the UK, where GBR genomes depicted high WHG ancestry fractions (>25%) and was correlated with high COVID-19 death/recovery ratio in the country (31.94).

Italy, one of the worst hit countries with the COVID-19 pandemic, showed significant nationwide variation in the severity of disease manifestation (death/recovery ratio=0.73). While COVID-19 related deaths in Northern Italy has surpassed 12,000, less than 1,000 deaths have been so far reported in Southern Italy. Interestingly, these numbers are consistent with WHG fractions among Italian genomes, thus while WHG ancestry fractions were found to be ~23.1% and 9.2% among Northern and Southern Italian genomes respectively.

In contrast, only a handful of COVID-19 cases have been so far reported in Russia, the largest country in the world, and we assessed whether this is owing to their genomic make-up. All modern-day Russians originated from two groups of East Slavic tribes: Northern and Southern and have 0% WHG ancestry fraction. They are genetically similar to modern-day other Slavic populations such as Belarusians. We note that COVID-19 death/recovery ratios are significantly low in both Russia (0.13) and Belarus (0.09). Furthermore, COVID-19 manifestation was found to be far less detrimental in the Central Asian countries that were once part of former Soviet Union (USSR) such as Kazakhstan (death to recovery ratio=0.16), Kyrgyzstan (0.11), Tajikistan (0), Turkmenistan (0) and Uzbekistan (0.08). Finally, Finnish genomes that were shaped by migrations from Siberia ~3500 years ago [19] with no WHG ancestral proportions, have a relatively smaller COVID-19 death to recovery ratio (0.08).

Overall, our results have revealed a clear association between WHG and Neolithic Iranian ancestry fractions and the acuteness of COVID-19 manifestation suggesting the presence of unique underlying genetic variants in European genomes that maybe correlated with their variable susceptibility and immune responsiveness.

The curious case of Central and South Americans

The COVID-19 infection has so far shown an interesting pattern in Central and South America. While in countries like Ecuador and Brazil the death to recovery ratio is significantly higher (Ecuador: 1.72 and Brazil: 3.5), it is intermediate in Columbia (0.38), and

medRxiv preprint doi: https://doi.org/10.1101/2020.04.05.20054627; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

discernibly low in countries such as Chile (0.05), Uruguay (0.05), Peru (0.07), Mexico (0.12) and Argentina (0.15). We hypothesized that this variation in COVID-19 severity may be attributed to the genetic ancestry of Central and South American populations. It has been reported that all present-day Native Americans have descended from at least four distinct ancient migration waves from Asia [20]. Two of them being ancient migrations (~15,000 years ago) from south-central Siberia (Mal'ta Upper Palaeolithic site: MA1 related) and the more recent (~5000 years ago) when Palaeo-Eskimos spread throughout the American Arctic [21]. MA1, is genetically proximal to all modern-day Native Americans (14-38% Native American ancestry) and has been found to be basal to modern-day West Eurasians, without any close affinity to East Asians [21]. MA-1 mitochondrial genome was determined to be associated with haplogroup U, which is found at high frequency among Mesolithic European hunter-gatherers (WHGs) [21], depicting a connection between the WHG and modern-day South American genomes. In our analyses Americans could be best modelled as a combination of two source populations namely MA1 and Eskimos. We used a mixture of seven ancient and modern-day populations comprising of CHG, WHG, Sweden_HG_Motala, Papuan, Chukchi, Karitiana and Mbuti as our 'Right' outgroup populations (O7). Interestingly, our results indicated that countries such as Mexico and Peru with lower death to recovery ratio have higher Eskimo ancestry fraction (13.2% and 20.3%) and lower MA1 fraction (86.8% and 79.7%) compared to our other study populations. Further, we found ~zero Eskimo ancestry fraction among Columbians and Purto Ricans where death to recovery ratio is relatively high.The high death to recovery ratio in Brazil (3.5) is reminiscent of similarly high ratios in Portugal (3.9) and may be attributed to the predominant Portuguese ancestry in the former.

Intrinsic 'protection' for East Asians?

As noted above, significant variation in AFs of *ACE2* gene, the presumed receptor of novel coronavirus has been reported among people from East Asian and European ancestry [5]. The authors of this study speculated that the differences in AFs of *ACE2* coding variants are associated with variable expression of ACE2 in tissues. Additionally, they reported that most eQTL variants identified by them had discernibly higher AFs among East Asian populations compared to Europeans, which may result in differential susceptibility towards the novel coronavirus in these two populations under similar conditions.Consistent with these findings we identified 404 SNPs that are associated with host immune response such as Interferon

medRxiv preprint doi: https://doi.org/10.1101/2020.04.05.20054627; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

(IFN) stimulated antiviral response, Interferon-stimulated gene 15 (ISG15) mediated antiviral mechanism and 2′-5′ oligoadenylate synthase (OAS) mediated antiviral response and show large differences in AFs between Europeans and East Asians. The genetic differences between East and Southeast Asians with Europeans can be largely attributed to their distinctive ancestral origins. Most East Asians derive their ancestry from Mongolian hunter-gatherers who dispersed over Northeast Asia 6000-8000 years ago [22]. Similarly, Southeast Asians exhibit a mixture of East Asian ancestry (Southern Chinese agriculturalist) and a diverged form of Eastern Eurasian hunter-gatherer ancestry (EHGs) [23]. People with similar ancestry can be found as far south as Indonesia [23]. As noted above EHG ancestry fraction does not appear to be significantly associated with COVID-19 manifestation. In congruence with their unique ancestral make-up, all East and Southeast Asian countries exhibit low COVID-19 death/recovery ratio (Japan=0.15, South Korea=0.03, Hong Kong=0.02, Taiwan=0.1, Thailand=0.03, Malaysia=0.06 and Singapore=0.02). Notably, in mainland China, where the novel corona virus originated and the catastrophe first initiated, COVID-19 death to recovery ratio has remained discernibly low (0.04). Overall, our results indicate that people of East and Southeast Asian ancestry appear to be intrinsically protected against the most debilitating effects of the novel coronavirus infection.

South Asians: a variegated canvas for COVID-19

South Asians have a long and complex history of admixture between immigrant gene-pools originating primarily in West Eurasia, Southeast Asia and the South Asian hunter-gatherer lineage with close genetic proximity to the present-day Andamanese people (Ancient Ancestral South Indians: AASI) who likely arrived in India through the “southern exit” wave out of Africa [24]. People of AASI ancestry admixed with an undivided ancient Iranian lineage that subsequently split and lead to the formation of early Iranian farmers, herders, and hunter gatherers approximately in the 3rd millennium BCE. This gene pool is referred to as the ‘Indus Periphery’ gene pool [24,25], which is thought to be the major source of subsequent peopling of South Asia. Modern-day South Asian genome is composed of largely four ancestral components: Ancestral North Indian (ANI), Ancestral South Indian (ASI), Ancestral Tibeto-Burman (ATB) and Ancestral Austro-Asiatic (AAA) [26]. ANI and ASI gene pools likely arose around 2nd millennium BCE during the decline of Indus Valley Civilization (IVC), which prompted multiple waves of migrations across the South Asia. The southward migration of Middle to Late Bronze Age people from Steppe (Steppe MLBA) into

medRxiv preprint doi: https://doi.org/10.1101/2020.04.05.20054627; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

South Asia is thought to have coincided with the decline of IVC. It is speculated that the people of Indus-Periphery-related ancestry, while migrating northward, admixed with Steppe MLBA immigrants to form the ANI, while the others, who migrated southward and eastward admixed with AASI and formed the ASI [25]. Austroasiatic speakers originated in Southeast Asia and subsequently migrated to South Asia during the Neolithic period. The admixture between local South Asians and incoming Southeast Asians took place ~2000-3800 years ago giving rise to the AAA ancestry. Notably, the South Asian population(s) with whom the incoming Southeast Asians mingled were AASI related with little to no West Eurasian ancestry fraction [27]. Finally, it has been showed that the Tibeto-Burmans (ATBs) derived their ancestry through admixture with low-altitude East Asians who migrated from China and likely across Northern India or Myanmar [28] leading to the high genomic proximity between South Asians of ATB ancestry and East Asians.

The genomic proximity between East Asians and South Asians mostly from Northeast India with prominent ATB ancestry, is likely reflected through fewer numbers of COVID-19 cases from this region, so far (one case each in Arunachal Pradesh and Mizoram, and two in Manipur). Notably the death to recovery ratio in the largely ATB dominated region of Ladakh, India is zero, indicating complete recovery so far of COVID-19 infected individuals from this region,while the same in its neighboring state of Jammu and Kashmir with predominant ANI ancestral fractions is discernibly higher (0.5). Further, Indian states with large number of indigenous tribal population with prevalence of AAA ancestry and/or with large fractions of AASI-related ancestry, eg. Chhattisgarh, Jharkhand and Orissa have not registered any COVID related deaths till date, indicating that South Asians with dominant AAA and AASI related ancestry may also likely be less severely affected against the most detrimental effects of COVID-19 infection.

Nevertheless, the severity of COVID-19 manifestation is likely to vary appreciably across the South Asian populations. Approximately 12% people of ANI and ASI ancestries belong to mitochondrial haplogroup U, which, as described above, is found at high frequency among Mesolithic European hunter-gatherers (WHGs) [21]. Our interrogation of worldwide populations suggests that the WHG ancestral fraction is likely associated with acute COVID-19 manifestation and is predictive of debilitating effects of COVID-19 infection among South Asian populations with substantial fractions of WHG ancestry.Overall COVID-19 manifestation in South Asia is likely to be somewhat intermediate to that observed in Europe and East Asia. This is underscored bythe recent finding that AFs of *ACE2* variants among South Asians are intermediate between East Asians and Europeans [5].

medRxiv preprint doi: https://doi.org/10.1101/2020.04.05.20054627; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Limitations and conclusion

The present study shines light of underlying genetic signatures that may be associated with disparate COVID-19 severity and manifestation in worldwide populations. Nevertheless we note that the current work has been performed using publicly available genomic data and a more robust understanding in this regard will emanate from sequencing/genotyping endeavours for COVID-19 patients across the spectrum of varied nationalities/ancestries and geographical locationsincluding individuals with mild to moderate symptoms, severe manifestations and death. Our results strongly advocate the adoption of a rigorous worldwide population genetics driven approach to expand and substantiate our current knowledge, as well as facilitate thedevelopment of population specific therapeutics to mitigate this worldwide challenge.

References

1. Laboratory testing of 2019 novel coronavirus (2019-nCoV) in suspected human cases: interim guidance, 17 January 2020 [Internet]. [cited 2020 Apr 9]. Available from: https://apps.who.int/iris/handle/10665/330676

2. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020 Feb;S1473309920301201. PMID:32087114

3. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020 Mar;579(7798):270–273. PMID:32015507

4. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet 2020 Feb;395(10224):565–574. PMID: 32007145

5. Cao Y, Li L, Feng Z, Wan S, Huang P, Sun X, et al. Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations. Cell Discov 2020 Dec;6(1):11. PMID: 32133153

6. Li G, Fan Y, Lai Y, Han T, Li Z, Zhou P, et al. Coronavirus infections and immune responses. J Med Virol 2020 Apr;92(4):424–432. [doi: 10.1002/jmv.25685]

7. Quach H, Rotival M, Pothlichet J, Loh Y-HE, Dannemann M, Zidane N, et al. Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. Cell 2016 Oct 20;167(3):643-656.e17. PMID:27768888

medRxiv preprint doi: https://doi.org/10.1101/2020.04.05.20054627; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

8. Mathieson I, Alpaslan-Roodenberg S, Posth C, Szécsényi-Nagy A, Rohland N, Mallick S, et al. The genomic history of southeastern Europe. Nature 2018 Mar;555(7695):197–203. PMID: 29466330

9. Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, Booth T, et al. The Beaker phenomenon and the genomic transformation of northwest Europe. Nature 2018 Mar;555(7695):190–196. PMID: 29466337

10. Olalde I, Mallick S, Patterson N, Rohland N, Villalba-Mouco V, Silva M, et al. The genomic history of the Iberian Peninsula over the past 8000 years. Science 2019 Mar 15;363(6432):1230–1234. PMID: 30872528

11. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006 Aug;38(8):904–909. PMID: 16862161

12. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 2007 Sep;81(3):559–575. PMID: 17701901

13. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 2015 Jun;522(7555):207–211. PMID: 25731166

14. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient Admixture in Human History. Genetics 2012 Nov;192(3):1065–1093. PMID: 22960212

15. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 2009 Sep 1;19(9):1655–1664. PMID: 19648217

16. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005 Jan 15;21(2):263–265. PMID: 15297300

17. Dayem Ullah AZ, Oscanoa J, Wang J, Nagano A, Lemoine NR, Chelala C. SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. Nucleic Acids Res 2018 Jul 2;46(W1):W109–W113. PMID: 29757393

18. Lipson M, Szécsényi-Nagy A, Mallick S, Pósa A, Stégmár B, Keerl V, et al. Parallel palaeogenomic transects reveal complex genetic history of early European farmers. Nature 2017 Nov;551(7680):368–372. PMID: 29144465

19. Lamnidis TC, Majander K, Jeong C, Salmela E, Wessman A, Moiseyev V, et al. Ancient Fennoscandian genomes reveal origin and spread of Siberian ancestry in Europe. Nat Commun 2018 Dec;9(1):5018. PMID: 30479341

20. Flegontov P, Altınışık NE, Changmai P, Rohland N, Mallick S, Adamski N, et al. Palaeo-Eskimo genetic ancestry and the peopling of Chukotka and North America. Nature 2019 Jun;570(7760):236–240. PMID: 31168094

medRxiv preprint doi: https://doi.org/10.1101/2020.04.05.20054627; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

21. Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 2014 Jan;505(7481):87–91. PMID: 24256729

22. Wang C-C, Yeh H-Y, Popov AN, Zhang H-Q, Matsumura H, Sirak K, et al. The Genomic Formation of Human Populations in East Asia [Internet]. Genomics; 2020 Mar. [doi: 10.1101/2020.03.25.004606]

23. Lipson M, Cheronet O, Mallick S, Rohland N, Oxenham M, Pietrusewsky M, et al. Ancient genomes document multiple waves of migration in Southeast Asian prehistory. Science 2018 Jul 6;361(6397):92–95. PMID: 29773666

24. Shinde V, Narasimhan VM, Rohland N, Mallick S, Mah M, Lipson M, et al. An Ancient Harappan Genome Lacks Ancestry from Steppe Pastoralists or Iranian Farmers. Cell 2019 Oct;179(3):729-735.e10. PMID: 31495572

25. Narasimhan VM, Patterson N, Moorjani P, Rohland N, Bernardos R, Mallick S, et al. The formation of human populations in South and Central Asia. Science 2019 Sep 6;365(6457):eaat7487. PMID: 31488661

26. Basu A, Sarkar-Roy N, Majumder PP. Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. Proc Natl Acad Sci 2016 Feb 9;113(6):1594–1599. PMID: 26811443

27. Tätte K, Pagani L, Pathak AK, Kõks S, Ho Duy B, Ho XD, et al. The genetic legacy of continental scale admixture in Indian Austroasiatic speakers. Sci Rep 2019 Apr;9(1):6104. PMID:30967570

28. Gnecchi-Ruscone GA, Jeong C, De Fanti S, Sarno S, Trancucci M, Gentilini D, et al. The genomic landscape of Nepalese Tibeto-Burmans reveals new insights into the recent peopling of Southern Himalayas. Sci Rep 2017 Dec;7(1):15512. PMID: 29138459

medRxiv preprint doi: https://doi.org/10.1101/2020.04.05.20054627; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table1: European Mesolithic hunter gatherers (WHGs) ancestry fractions, mean temperature (15 days), mean humidity (15 days), and COVID-19 death to recovery ratios of 10 European countries employed in the study

Countries	**Mean Temperature (°C, 15 days')**	**Mean Humidity (°C, 15 days')**	**WHG ancestry fraction (%)**	**Death/Recovery ratio 05.04.2020**	**Death/Recovery ratio 08.04.2020**
Norway	4.09	75.82	18.2	1.94	2.8
UnitedKingdom	8.18	69.45	26.1	31.94	45.62
France	12.64	74.09	20.9	0.49	0.53
Spain	10	58.64	21.1	0.33	0.33
Italy	10	62	23.1	0.73	0.7
Greece	11.91	74.45	14.6	0.87	0.3
CzechRepublic	4.91	52.36	23.8	0.79	0.5
Hungary	7.1	49.9	17.2	0.51	0.62
Bulgaria	3.9	71	16.6	0.49	0.55
Ukraine	6.8	45.4	17	1.28	1.49

medRxiv preprint doi: https://doi.org/10.1101/2020.04.05.20054627; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 2: Neolithic Iranian and Mesolithic hunter-gatherers of Eastern Europe (EHGs) ancestry fractions among 10 European countries employed in the study

Countries	**Neolithic Iranian ancestry fraction**	**Mesolithic hunter-gatherers of Eastern Europe (EHGs) ancestry fraction**
Norway	56.4	25.4
United Kingdom	51.5	22.4
France	61.2	17.9
Spain	68.7	10.2
Italy	69.5	8.3
Greece	73.6	11.7
CzechRepublic	54	22.2
Hungary	60.8	22
Bulgaria	68.5	14.8
Ukraine	57.7	25.3

medRxiv preprint doi: https://doi.org/10.1101/2020.04.05.20054627; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

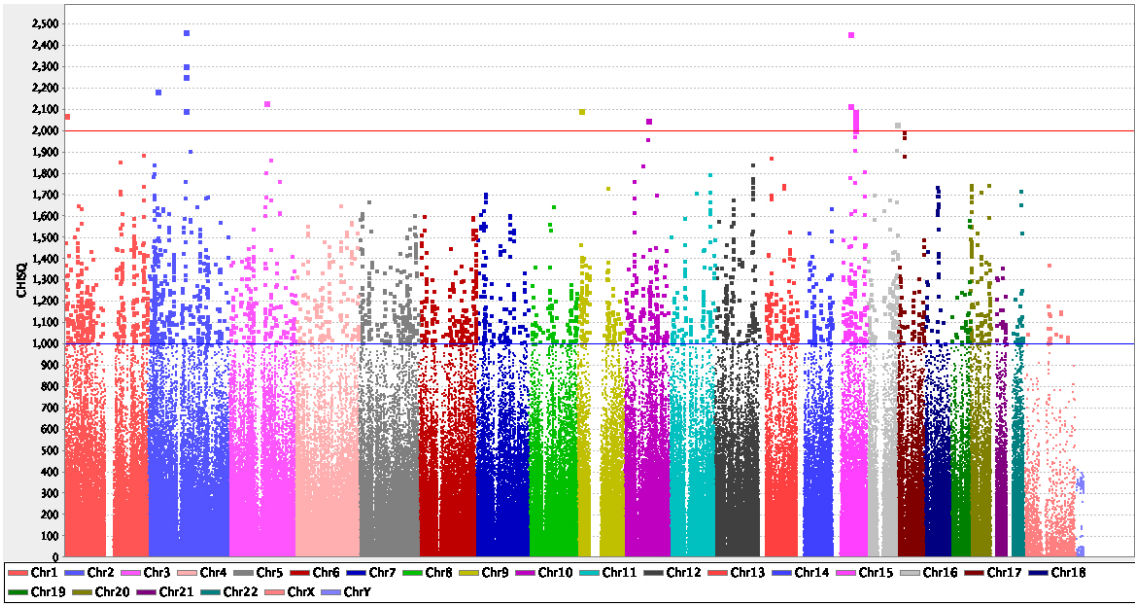
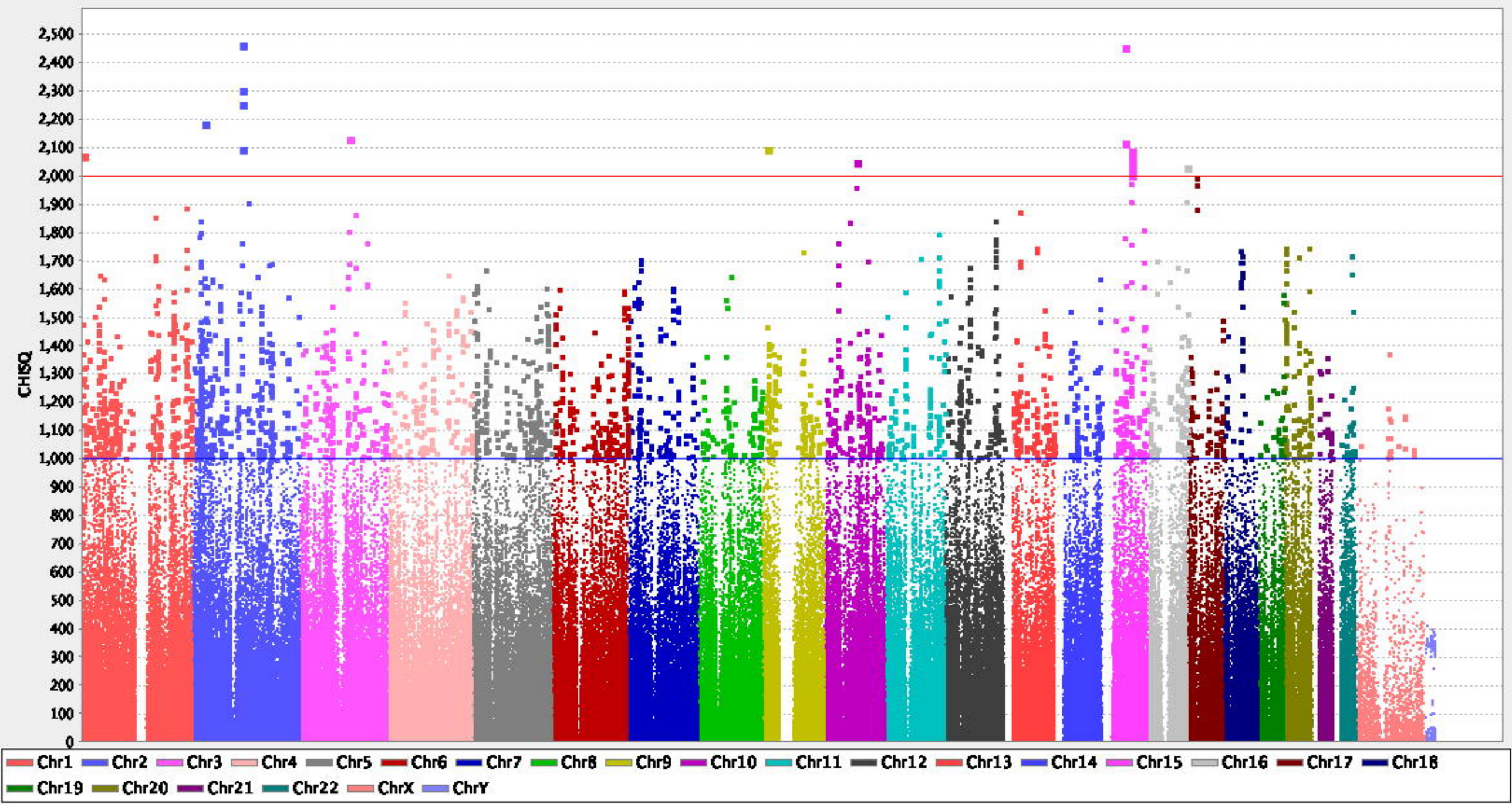


Fig.1: Manhattan Plot. GWAS with 753 European genomes were compared to 838 Eastern Asian genomes. Out of 597,573 SNPs employed, 385,450 (64.5%) markers revealed highly significant variation between Europeans and East Asians. The chi square values are plotted in Y-axis. The SNPs are designated with dots. The SNPs with chi square values >1000 are indicated with the blue line and those with chi square value >2000 are indicated with the red line. While 18 SNPs showed chi square values >2000, 2,992 SNPs had chi square values >1000.


CHISQ
2,500
2,400
2,300
2,200
2,100
2,000
1,900
1,800
1,700
1,600
1,500
1,400
1,300
1,200
1,100
1,000
900
800
700
600
500
400
300
200
100
0
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
Chr22
ChrX
ChrY