Severe Covid-19 Infection via a Gene Variant

John Allen

ABSTRACT

Reportable COVID-19 infection rates vary radically from country to country. Remarkably, the reported rates in western Europe are more than an order of magnitude greater than in the tropics and sub-tropics. This suggests a common human genetic variant may be the primary culprit behind reportable COVID-19 infections.

Actual culprit identification will provide focus for efforts to identify prophylactics, treatments, and vaccines. Culprit identification will also identify pathways for relaxing the physical distancing measures that seriously constrain peoples' lives and the world's economy.

Reportable COVID-19 infection rates show a definite male bias. This suggests that the culprit variant might be located on the X chromosome.

Reportable COVID-19 infections are associated with a remarkably low expression of type I and III interferons. This suggests that the culprit variant might be associated with the detection or signaling portion of the innate immune system.

By filtering gene variants based on their frequency, geographic distribution, chromosomal location, and function, one can substantially narrow the list of candidate culprit variants.

Of nearly 900,000 candidate culprit variants analyzed as part of the multi-layer screening effort described herein, only five survive frequency, geographic, chromosomal, and functional filtering. Those five candidate culprits belong to the HDAC6 and IRAK1 genes.

The candidate variants are identified so that others may further evaluate them based on disproportional appearance among patients who have suffered serious COVID-19 infections.

WORLD-WIDE INFECTION RATES

The use of country-by-country reportable COVID-19 infection rates for data mining purposes is risky for at least two reasons. First, it is now clear that there are far more COVID-19 infections than reported. Second, countries are far from consistent in their reporting criteria and procedures. Nonetheless, the distribution of COVID-19 infections across the planet is so striking that it deserves consideration.

For this effort, I wrote a database program that allows comparison of various metrics against reportable COVID-19 infections on a country-by-country basis. I obtained the infection rates from the *worldometers* web site, including in the database every country that had at least one reported COVID-19 death. The reported infection data are current as of 11 May 2020.

I classified each country by geographic region according to the United Nations geoscheme. The non-weighted average infection rates for those regions follow immediately below, reported as infections per million residents. The numbers within parentheses are the number of countries in the sample. For comparison, the worldwide average for 171 countries was 1,040 reportable infections per million people.

> North America (3 countries): 2651 reportable infections per million residents
> Europe (44): 2,416
> Asia (39): 820
> Caribbean (18): 607
> Latin America (21): 501
> Oceania (2): 292
> Africa (44): 216

The results for European sub-regions are also of interest.

Eastern Europe (9): 909
Northern Europe (11): 2,091
Western Europe (9): 3106
Southern Europe (15): 3,145

The results suggest that countries might suffer reportable infections in proportion to the percentage of their population having European ancestry. This in turn suggests that something unique to the genetic coding of Europeans might be a major determinant of who acquires a severe, reportable COVID-19 infection.

Figure 1 shows that a country's reportable COVID-19 infection rate does indeed vary linearly with the percentage of its population belonging to a specific haplogroup. A haplogroup is a collection of people sharing a combination of genetic variants that can be traced back to a common ancestor. The haplogroup of particular interest herein, M269, is concentrated in Western Europe and North America.
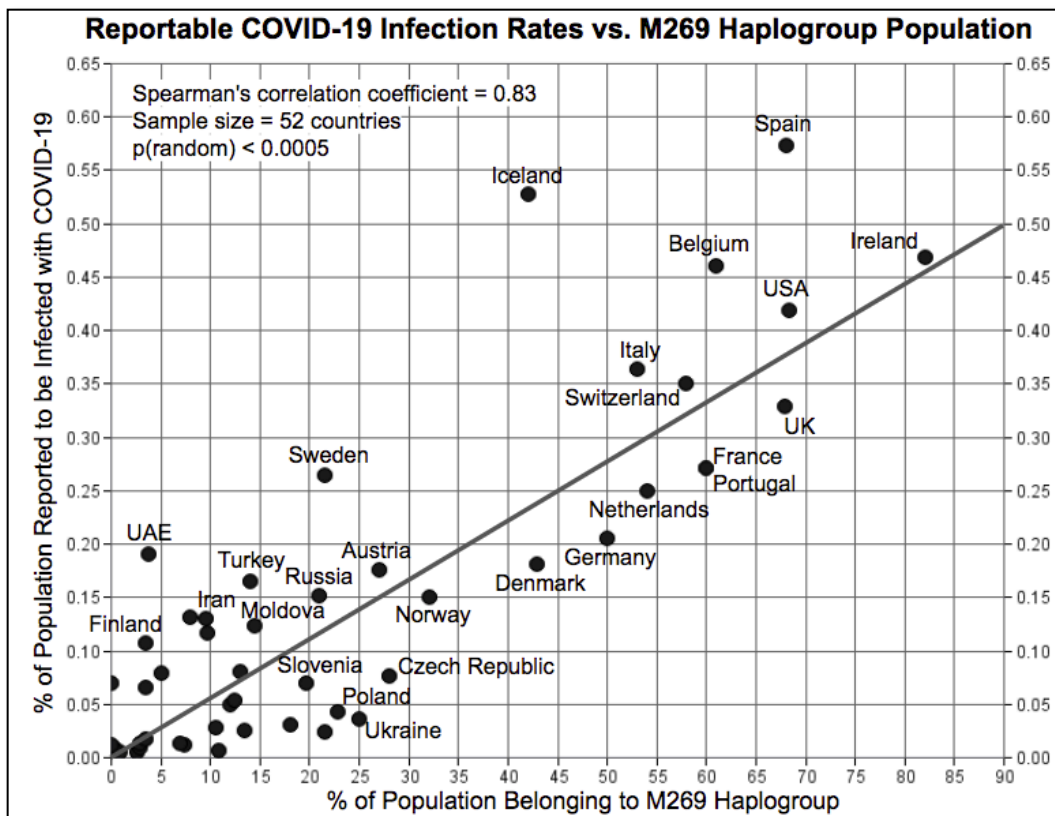

Figure 1: Reportable Infections vs. M269 Haplogroup Population

The unlabeled countries crowded near the plot origin are Albania, Algeria, Bosnia and Herzegovina, Bulgaria, China, Croatia, Egypt, Estonia, Greece, Greenland, India, Iraq, Japan, Jordan, Latvia, Lebanon, Libya, Morocco, North Macedonia, Oman, Pakistan, Romania, Serbia, Palestine, Tunisia, United Arab Emirates, and Yemen.

Spearman's correlation coefficient for the data is 0.83. That is an exceptionally large coefficient, particularly for a sample size of 52 data points. For a confidence level greater than 99.9%, the critical coefficient is only 0.45. Therefore, the possibility that a country's COVID-19

reported infection rate does not correspond with its M269 haplogroup fraction is well less than one in a thousand.

The strong correlation does not indicate that someone who does not carry the M269 gene variant collection is free from a COVID-19 infection. It is becoming increasingly clear that most COVID-19 infections are asymptomatic or paucisymptomatic, resolve without a test or hospital visit, and are therefore unreported. The work herein makes no effort to correlate genetic variants with asymptomatic or paucisymptomatic cases. Instead, the hypothesis being considered herein is that the presence of a specific gene variant causes the possessor to be more susceptible to a COVID-19 infection that develops to a case sufficiently severe to be reported.

Furthermore, the correlation of Figure 1 does not indicate that someone who does not carry the M269 genetic variant collection will not develop a severe, even fatal, COVID-19 infection. Most countries are reporting COVID-19 infections and deaths, and most countries have low M269 haplogroup populations. The correlation indicates only that a country's rate of reportable infections has, so far, correlated with the country's density of M269 haplogroup residents.

Compare, for example, the United States and China. In the US, approximately 68% of the residents carry the M269 genetic variant. In China, approximately 0.8% of the residents carry the virus. The reported COVID-19 infection rate in the US is 4,187 per million, compared to only 58 per million across all of China. The US has 72 times the number of reported infections per million. That number should no longer be surprising since the US has 86 times the percentage of M269 haplogroup residents.
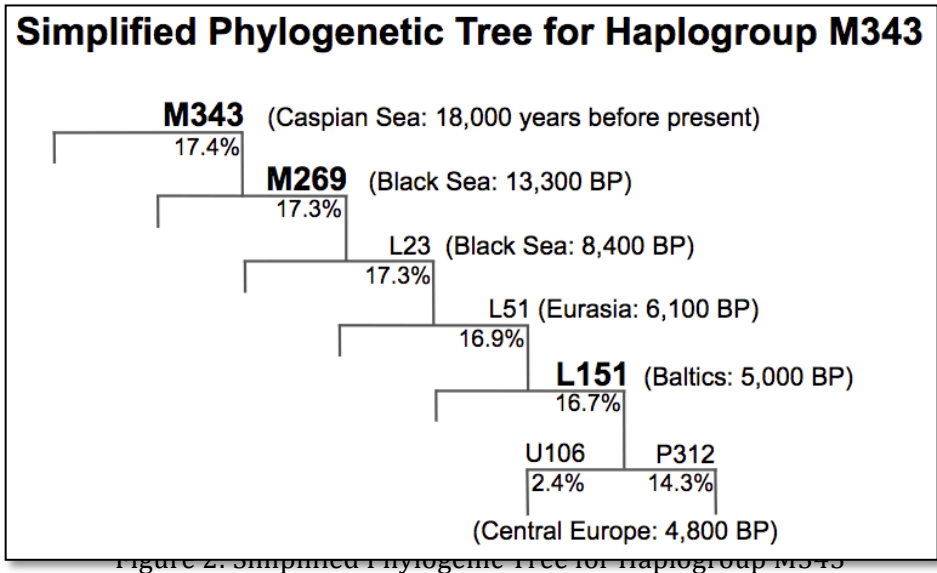
Though China has been accused of underreporting its COVID-19 infection rate, its reported rate is actually greater than the value predicted by the trend line of Figure 1. According to that trend line, valid only as of 11 May 2020, a country's infection rate, as a percentage of its population, can be estimated by multiplying its M269 population, as a percentage of its entire population, by 0.00555. Relative to the trend line, China's reported infection rate is 30% higher than expected, presuming its M269 haplogroup residents do indeed constitute 0.8% of the country's population. On the same basis, the US reported infection rate is 10% higher than expected, assuming its M269 haplogroup residents do indeed constitute 68.4% of the country's population.

Certainly crowded conditions, population comorbidities, infection control measures, infection reporting practices, and other factors can play a significant role in determining reported infection rates. The reported infection rate in the US state of New York, for example, is 22,904 per million, more than 5 times the US rate. The reported infection rate in Wuhan, the first city to suffer a major outbreak, is 4,543 per million, almost 80 times China's national rate.

Despite the many variables that contribute to reportable infections, the rate for an entire country can be approximated knowing only the percentage of the population bearing the M269 genetic marker.

THE M343 HAPLOGROUP

As shown in Figure 2, the M269 haplogroup formed as a subclade (a subgroup) of the M343 haplogroup, migrating from the Caspian Sea region to the Black Sea region approximately 13,300 years ago. The new collection of variants that identify the M269 haplogroup were genetically successful, and that haplogroup now includes almost the entirety of the surviving M343 haplogroup.

Simplified Phylogenetic Tree for Haplogroup M343

M343 (Caspian Sea: 18,000 years before present)
17.4%

M269 (Black Sea: 13,300 BP)
17.3%

L23 (Black Sea: 8,400 BP)
17.3%

L51 (Eurasia: 6,100 BP)
16.9%

L151 (Baltics: 5,000 BP)
16.7%

U106 | P312
2.4% | 14.3%

(Central Europe: 4,800 BP)

Figure 2. Simplified Phylogenetic Tree for Haplogroup M343

The M343 haplogroup is presently composed mostly, but not entirely, of individuals in the M269 haplogroup. So similar are the two groups that the correlation plot of Figure 1 includes a few countries for which only M343 percentages have been identified. For those countries, the M269 percentage was presumed to be equivalent to the M343 percentage.

Similarly, the M269 haplogroup is presently composed mostly, but not entirely, of individuals in the L151 haplogroup. Therefore, today's worldwide distribution of people bearing L151 genetic variants is similar to the distribution for people having only M269 variants, which in turn is similar to those having only M343 variants.

A heat map for the entire M343 haplogroup, including its many subclades, is presented below as Figure 3.

Figure 3. Hotspot Map of Haplogroup M343 Distribution
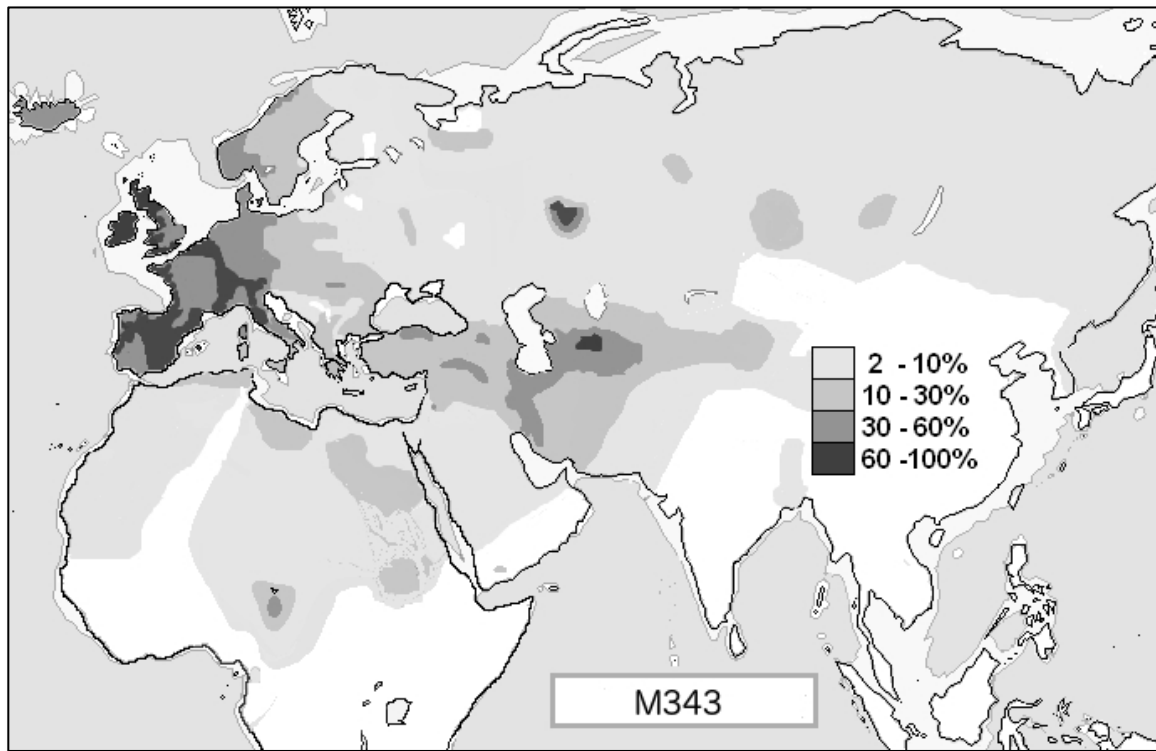
Considering Figure 3, take note of the M343 hotspot in the midst of Russia. It approximates the Republic of Bashkortostan, a federal subject of Russia, home of the Bashkirs, a Turkic ethnic group. Approximately 47.5% of the Bashkirs carry the M343 genetic marker. The reported COVID-19 infection rate of Bashkortostan is a startling 62,000 per million, compared to the worldwide average of 1,000 per million.

Considering Figure 3 further, take note of the M343 hotspot just east of the Caspian Sea. It approximates the northern portion of Turkmenistan. Approximately 37% of Turkmens carry the M343 genetic marker. Turkmenistan, one of the world's most repressive countries, has yet to officially acknowledge any COVID-19 infections within its borders. However, based on the trend line of Figure 1, we can estimate their serious infection rate to be near 2000 per million residents, approximately twice the world average. The estimate presumes an absence of serious, adverse, confounding factors such as those that seem to have driven Bashkortostan's numbers to such deadly levels.

Still considering Figure 3, take note of the M343 trends in Italy. The collection of M343 genetic variants is most heavily concentrated in the north, and more moderately concentrated in the south. A province-by-province heat map of Italy's reported COVID-19 infection rates, such as that presented in Figure 4 below, shows a corresponding distribution.
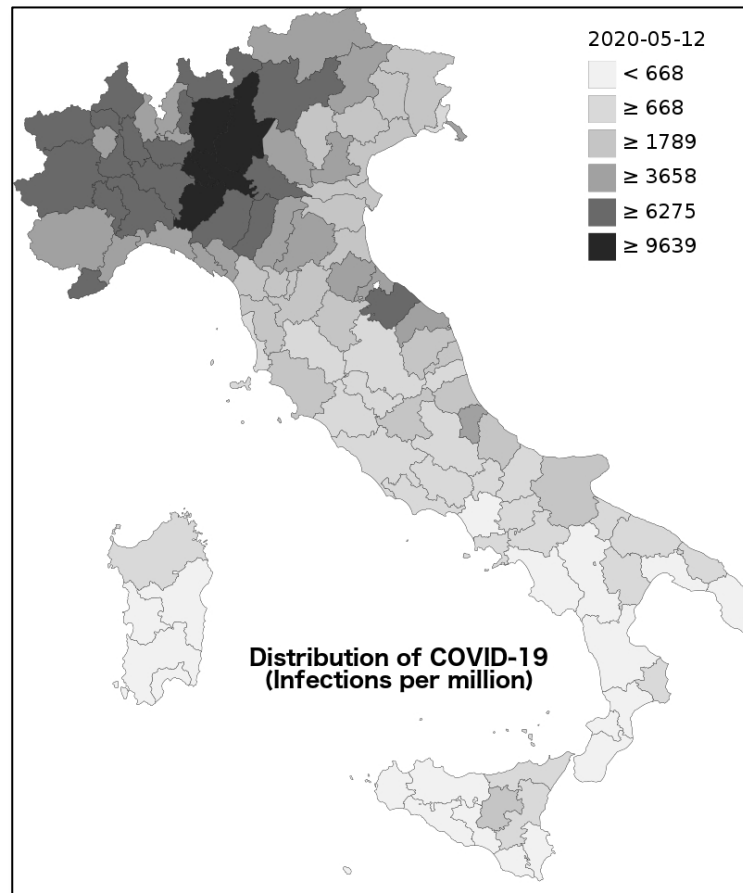
Figure 4: Hotspot Map of COVID-19 Distribution in Italy

Viewing Figure 4, take note of the small white spot near the northeast coast. That is the Republic of San Marino, a small country of 33,000 people surrounded entirely by Italy. San Mario maintains a specific ethnic profile via its widespread practice of endogamy, marrying within the limits of its own community. While the haplogroup distribution of the country has not been defined or reported, its COVID-19 infection rate has. At 18,508 reported infections per million residents, San Marino has the highest infection rate of any of the 171 countries in my database.

THE L151 HAPLOGROUP

Individuals carrying the L151 genetic variants carry also the genetic variants associated with all ancestor haplogroups, back to M343 and beyond. The reverse is not necessarily true. Individuals having the M343 genetic variants may belong to a subclade other than L151. So too with M269 and the other intermediate haplogroups.

In summary, an individual belonging to any specific haplogroup will nominally have the genetic variants of all ancestor haplogroups. That individual, however, may or may not have the genetic variants associated each descendant haplogroup.

The significance of this upward versus downward association of genetic variants is that we can, in theory, identify the specific haplogroup that introduced the genetic variant responsible for high rates of COVID-19 reportable infections. To do so, we need population percentages for the pure, unvariated haplogroup lines. In other words, we need country-by-country

percentages for those individuals who carry the M343 variants but not the M269 variants, and for those who carry the M269 variants but not the L23 variants, and so on down the line.

Fortunately, Myres *et. al.* provide the very data needed. Though their data are now ten years out of date, and though their data cover only 23 countries (but many more ethnic groups), their data are all that are apparently currently available. By comparing various sequences of their decade-old data against today's COVID-19 reported infection rates, it becomes clear that the genetic variant of interest appeared no earlier than the L151 haplogroup.

As shown in Figure 5 below, the Myres data for the L151 haplogroup, along with its two subgroups, correspond with the COVID-19 reported infection rates, on a country-by-country basis. On the other hand, the Myres data for the entire M343 haplogroup, deprived only of its L151 subgroup, do not correspond with today's COVID-19 reported infection rates.
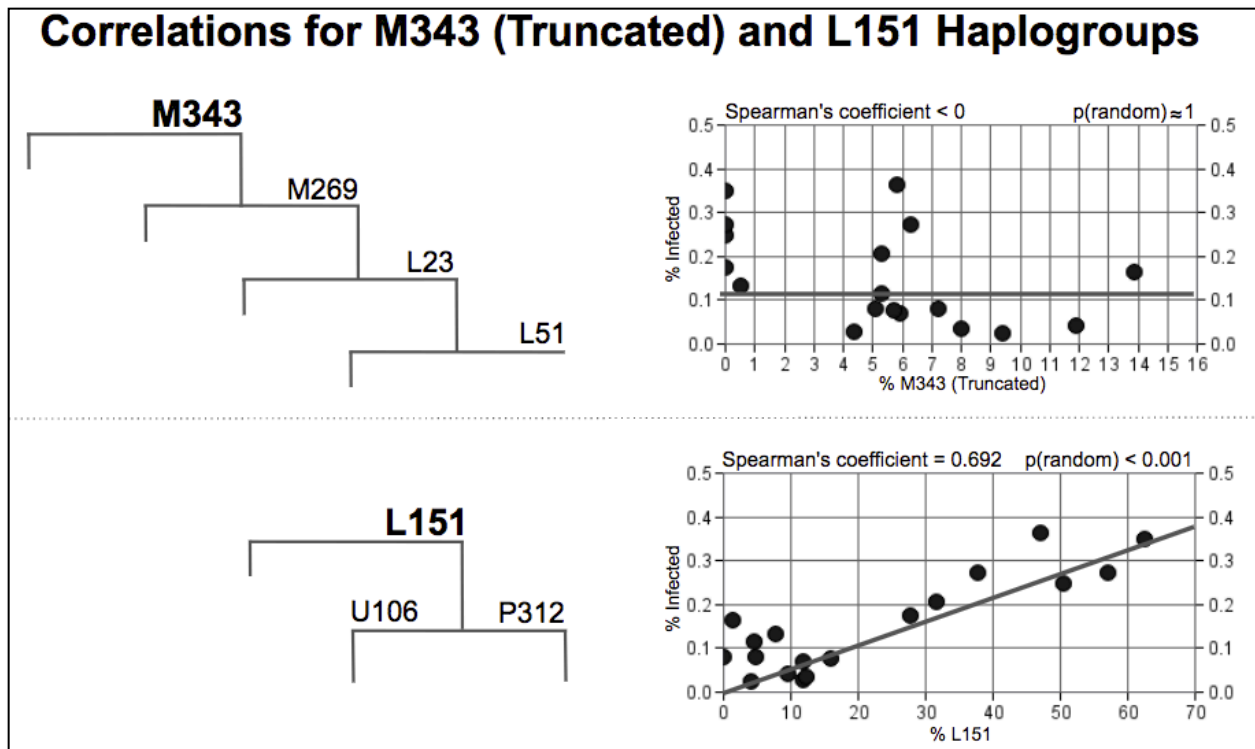

Figure 5: Correlations for M343 (Truncated) and L151 Haplogroups

It is now clear that the tight correlation between the M269 haplogroup and COVID-19 reportable infections, so prominently presented and discussed near the beginning of this paper, exists only because L151 subgroup individuals constitute an overwhelming majority within the ancestor M269 haplogroup. M269 was useful as a beginning point because its distribution is much better established than that of its L151 subgroup.

The L151 haplogroup data of Figure 5 are repeated below in Figure 6, enlarged and labeled.
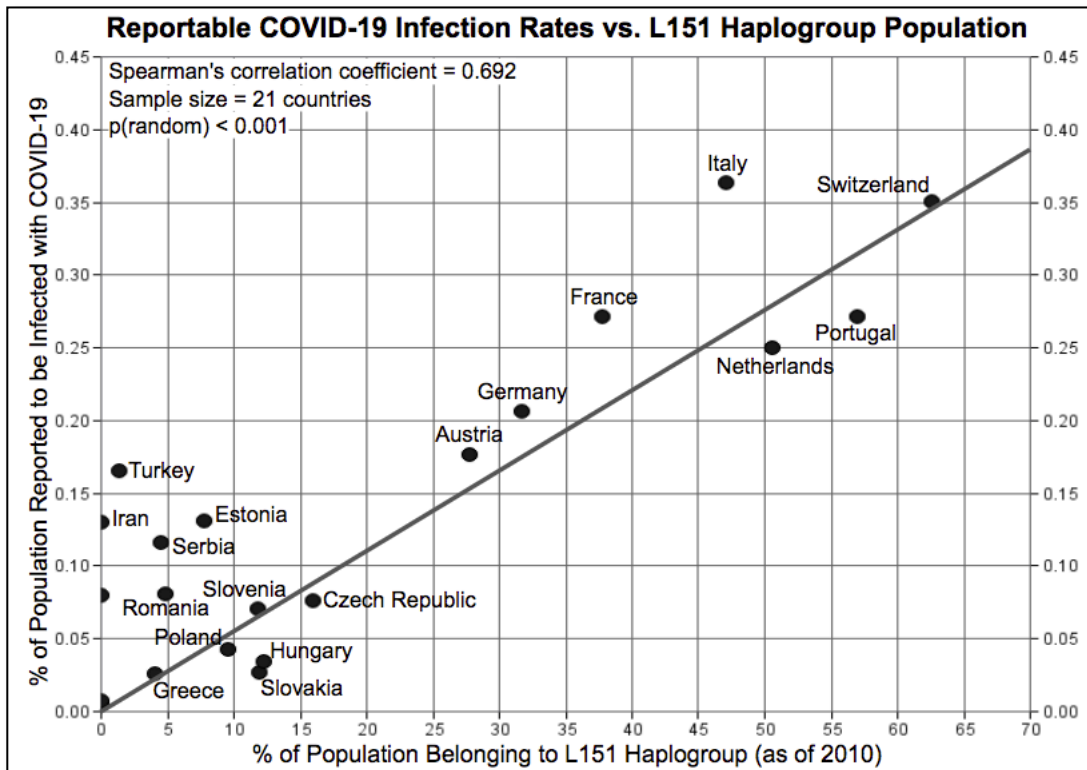
Figure 6: Reportable COVID-19 Infection Rates vs. L151 Haplogroup Population

THE L151 GENETIC MARKER

The identifying genetic variant, but not the only genetic variant, associated with the L151 haplogroup is rs9786076. The leading characters, rs, stand for reference SNP, which in turn stands for single nucleotide polymorphism.

An SNP results from a simple gene copying error, involving a single nucleotide pair, A-T or C-G, an error that successfully perpetuated itself throughout even a small percentage of the population.
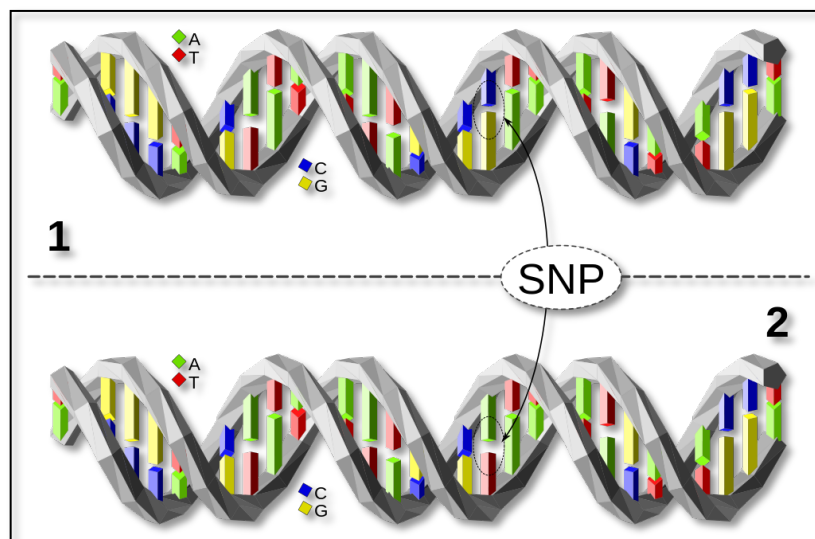

Figure 7: SNP Copying Error

The two similar but different copies of the gene are known as alleles. Since individuals within the population may carry either the reference allele and/or the variant allele, the gene is polymorphic. It exists in multiple forms.

More than 335 million SNPs have so far been discovered across multiple human populations. The National Center for Biotechnology Information (NCBI), a part of the US National Institute of Health (NIH), maintains a free public archive for information relating to SNPs and other genetic variations. The database is known as dbSNP. As information about each SNP is added to the database, the SNP is assigned an rs accession identifier. When the L151 genetic marker variant information was added to the dbSNP, the variant was assigned the identifier rs9786076. An online search for that identifier in the dbSNP system returns the following summary.

	rs9786076 *[Homo sapiens]*	
1.		
	Variant type:	SNV
	Alleles:	C>T [Show Flanks]
	Chromosome:	Y:15732138 (GRCh38)
		Y:17844018 (GRCh37)
	Validated:	by frequency,by cluster
	MAF:	C=0.1671/206 (1000Genomes)
	HGVS:	NC_000024.10:g.15732138C>T, NC_000024.9:g.17844018C>T
	PubMed LitVar	

Figure 8: dbSNP Summary for rs9786076 Gene Variant

From the dbSNP summary we learn that the L151 marker (rs9786076) is a human, single nucleotide variant, in which the minor allele has a T nucleotide (thymine) where the reference allele has an A nucleotide (adenine). The variant is located on the Y chromosome at position 15732138, at least as defined in the GRCh38 reference genome. The minor allele frequency (MAF) is 16.7% of the world's population, at least according to the 1000 Genomes project.

Since the variant is located on the Y chromosome, it is found only in males. That is as expected since the L151 haplogroup is a Y-haplogroup, a paternal haplogroup, as are the other haplogroups discussed herein. By focusing on the Y-chromosome, individual and ethnic group ancestries can, in theory, be traced along a single ancestral path all the way back to humanity's most recent common ancestor, a person known today as Y-chromosomal Adam, who lived in Africa approximately 236,000 years ago.

It is exceptionally unlikely, though, that the L151 variant rs9786076 is directly related to the significantly higher reported COVID-19 infection rates among Europeans. Though women suffer reportable COVID-19 infections at approximately half the rate of men, they do suffer the disease in non-trivial numbers, despite the complete absence of the L151 variant in their genes.

Individuals within a haplogroup, though, share not just a single gene variant. Instead they share a collection of gene variants, even though the complete collection may not be fully identified. Presuming severe cases of COVID-19 do relate to one or more genetic variants, as suggested by the correlation of COVID-19 infection and L151 geographic distributions, then we need some means of filtering the 335 million SNPs to generate a manageable list of candidate culprits for more thorough consideration.

FREQUENCY FILTERING OF GENETIC VARIANTS

Given that more than one-half percent of some European populations have already suffered reportable COVID-19 infections, whichever variant might be responsible for serious infections must exist in at least 0.5% of those European countries' populations. To determine a more realistic lower threshold, we need to consider both the percentage of the population already infected *and* the percentage of the population that must carry antibodies for herd immunity, presuming no vaccine.

Many more people have been infected than have been reported. Antibody tests in Chelsea, Massachusetts indicated that 32% of the local population carried a COVID-19 antibody. Similar antibody testing in New York suggested that 25% of New York City had been infected. For Los Angeles County, the estimate was between 2.8% and 5.6%. For Santa Clara County, 2.8%.

Indiana reported that, as of 13 May 2020, 2.8% of its residents carried COVID-19 antibodies. Sweden report that, as of 21 May 2020, 7% of the residents of Stockholm had been infected New York State reported that, as of 28 May 2020, 14.9% of its residents had been infected.

Obviously, the test results vary. Also, various antibody test methodologies have been subjected to substantial criticism. Absent a reliable percentage of the population already infected, we will only be able to deal with a wide range of possibilities.

The situation with herd immunity is only slightly better. Best estimates are that somewhere between 60% and 80% of the population must be infected or vaccinated to establish herd immunity.

If 32% of the European population already carries antibodies and if herd immunity will take hold when 60% of the European population is protected, then the infection has run approximately 50% of its natural course. That number seems highly optimistic and extremely unlikely, but it would suggest that the lower threshold for variant frequency is approximately 1% of the European Non-Finnish (EUR) population. Filtering by that threshold alone reduces the candidate culprit variants by 98%.

If, on the other hand, only 2.8% of the European population has been infected and 80% are required for herd immunity, then the viral infection has run only 3.5% of its natural course. That number, seems somewhat pessimistic but, unfortunately, more likely. It would set the lower threshold of EUR variant frequency to 14.3%. Filtering by that threshold alone reduces the candidate culprit variants by 99%.

To reduce the possibility of inadvertently excluding the actual culprit with too fine a filter, 1% will hereafter be used as the lower threshold for EUR frequency filtering. That leaves us with 17,263 candidates for further consideration.

[Note: Approximately 40% of the European population belongs to the L151 haplogroup. It is suddenly clear that whatever gene variant may be responsible for serious COVID-19, assuming one is indeed responsible, the variant does not exist within all members of the L151 haplogroup. This in turn suggests that the variant was introduced not with the L151 haplogroup, but rather with a distant sub-clave of the L151 haplogroup.]

GEOGRAPHIC FILTERING OF GENETIC VARIANTS

The worldwide distribution of reportable COVID-19 infections allows geographic filtering of genetic variants. The geographic filtering used for this investigation is based on the following observations regarding regional rates of reported COVID-19, as of 11 May 2020:

the rate among East Asian countries is 1/22nd that of European countries;
the rate among African countries is 1/11th that of European countries;
the rate among South Asian countries is 1/5th that of European countries; and
the rate among Latin American countries is 1/5th that of European countries.

To allow comparison of infection and variant distribution patterns, the database also includes worldwide distribution information for 889,030 variants most likely to be responsible for severe COVID-19 infections. These variants represent approximately 1000 genes collected from a variety of sources, mostly from the InnateDB web site.

The variant population distributions were obtained from the gnomAD database. That database, available online, includes minor allele frequencies for European non-Finnish (EUR), East Asian (EAS), South Asian (SAS), African (AFR), and Latino (LAT) populations. Since the gnomAD browser does not classify populations by geographic region, it would be merely coincidental if the collection of countries represented by the variant and infection databases matched perfectly.

The geographic filtering excludes, as candidate culprits, those variants having disproportionately large Asian, African, or Latin American frequencies relative to the European frequency, based on the following criteria:

variant rate EAS multiplied by 7 is greater than variant rate EUR;
variant rate AFR multiplied by 4 is greater than variant rate EUR;
variant rate SAS multiplied by 2 is greater than variant rate EUR; and
variant rate LAT multiplied by 2 is greater than variant rate EUR.

Note that the multipliers are approximately a third of those associated with the relative geographic distributions. Instead of 22, 11, 5, and 5, the somewhat coarse filter uses multipliers of 7, 4, 2, and 2, respectively.

Filtering by both geographic and frequency criteria reduces the candidate culprit variants by 99.9%, leaving us with 700 candidate culprits to consider.

CHROMOSOMAL FILTERING OF GENETIC VARIANTS

The SARS-CoV-2 virus shows a distinct male bias, as did the SARS-CoV virus, as do many infectious diseases.

In their 2019 paper, "*The X chromosome and sex-specific effects in infectious disease susceptibility*," Haiko Schurz *et. al.* explain, "Many diseases present with a clear sex bias, and apart from the influence of sex hormones and socioeconomic and behavioural factors, the X chromosome, X-linked genes and X chromosome inactivation mechanisms contribute to this difference. Females are functional mosaics for X-linked genes due to X chromosome inactivation and this, combined with other X chromosome inactivation mechanisms such as genes that escape silencing and skewed inactivation, could contribute to an immunological advantage for females in many infections."

X chromosome inactivation is the biological process in which one of a female's two X chromosomes becomes inactive. The inactivation keeps females from having twice as many active X chromosome related genes as do their male counterparts. For each cell within the female, it is usually random which of the two X chromosomes, either maternal or paternal, is inactivated.

If each parent carried the reference allele of a gene, then the woman would express only that one allele. She would be homozygous with respect to that gene. If, on the other hand, one

parent contributed the reference allele and the other parent contributed the variant allele, then the woman would express both the reference and variant genes, in roughly equal quantity. She would be heterozygous with respect to that gene.

Compared to a man having a specific variant on his single X chromosome, a woman having that same variant on just one of her X chromosomes would have reduced expression of that variant. If the reference allele provided viral protection that the variant did not, then the heterozygous woman would have some protection while the variant-exclusive man would not.

Regarding X chromosome inactivation, Schurz *et. al.* write, "Females carry both a maternal and paternal X chromosome, while males carry only a maternal copy. In order to regulate dosage expression of X-linked genes, one X chromosome is inactivated in females, resulting in them being functional mosaics for X-linked genes. XCI is initiated in early foetal development and either the maternal or paternal X chromosome is randomly silenced in XX cells."

Because X chromosome genes bring with them a built-in explanation for the gender bias of the SARS-CoV-2 virus, those genes deserve more attention that the others.

Filtering by geographic, frequency, and chromosomal criteria reduces the candidate culprit variants by more than 99.99%, leaving us with 38 candidate culprits to consider.

THE ACE2 GENE

A SARS-CoV-2 virus does its harm only after it has made its way into a cell. Once inside, the virus hijacks the cell's workings for its own viral replication. At some point, the replicants overwhelm the cell, kill it, and escape to infect other cells.

To gain entry into a cell, the virus must first bind itself to the cell. It is now reasonably well established and widely accepted that the SARS-CoV-2 virus initiates cell entry by binding to the ACE2 enzyme.

The *angiotensin-converting enzyme 2*, ACE2, is coded by the ACE2 gene, hence the correspondence of the names. ACE2 resides on the outer surface of cells found in virtually all internal body parts including lungs, arteries, heart, kidneys, intestines, testes, and nasopharynx.

ACE2 plays a major role in controlling blood pressure by controlling the ACE2 cousin enzyme, ACE. Some blood pressure medications, known as ACE inhibitors, also control ACE, and may increase ACE2. This increased expression of ACE2 has prompted discussion within the medical community regarding the advisability of continued use of ACE inhibitors during the pandemic. Since ACE2 also helps protect against the runaway inflammation and fibrosis associated with severe cases of COVID-19, the weight of evidence and argument so far seems to be heavily in favor of continued use of ACE inhibitors.

An enzyme that binds biologic material to a cell membrane is known as a host receptor. Viral binding occurs in large measure due to the complex, complimentary geometries of virus and receptor. It is worth considering that a specific ACE2 genetic variant encodes a superior ACE2 host receptor. If so, the worldwide distribution of the variant would correlate with the worldwide distribution of reported COVID-19 infections.

In their 24 February 2020 paper *"Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations,"* Yanan Cao *et. al.* analyzed 1700 ACE2 variants found in various genetic databases. They narrowed their found set of interest to "the 32 variants potentially affecting the amino acid sequence of ACE2." They further narrowed their focus to "seven hotspot variants." For those seven, they presented the worldwide distribution in a series of pie charts, recreated below as Figure 10.
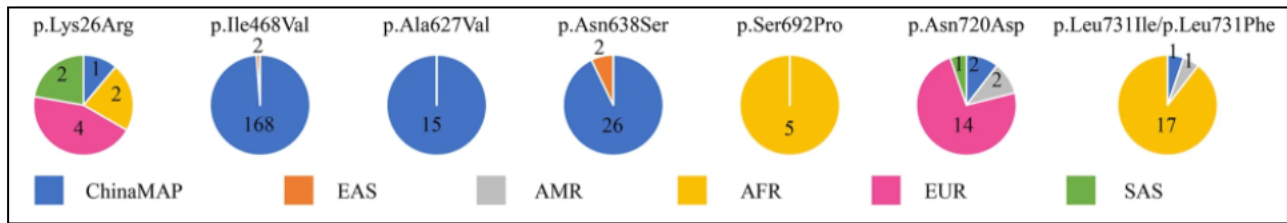
Figure 9: Population Distributions for Cao's Seven Hotspot Variants

The seven genetic variants are identified by their consequence. Considering the leftmost variant, the nomenclature p.Lys26Arg indicates that a variant at ACE2 gene location 26 coded for the arginine amino acid rather than the lysine amino acid. Working from left to right, the color coding for the pie chart is blue for China's own genetic database; orange for East Asians; gray for Americans (presumably Native Americans); yellow for Africans; red for Europeans; green for South Asians.

At the time of Cao's paper, the virus had not spread substantially beyond China's borders. The authors could not have anticipated the distinctive worldwide distribution of reportable infections that has since become obvious. The authors could do little more than report the distributions of their seven hotspot variants.

Today, now that we know the infection rates are substantially elevated among the European population, we can immediately exclude all but two of the Cao's hotspot gene variants as potential culprits. The two surviving candidates are p.Lys26Arg and p.Asn720Asp. The other five variants are not found to any significant extent within European populations, and they therefore cannot be responsible for the inordinately high infection rates of European countries.

Of the two remaining variants, p.Lys26Arg can also be excluded, since it is found half as frequently among Africans as Europeans. Africans suffer only a tenth the rate of reportable infections as Europeans, at least on a country-by-country basis.

Cao's single remaining ACE2 hotspot variant, p.Asn720Asp, is distributed in a worldwide fashion similar to the worldwide distribution of reportable COVID-19 infections, and it is therefore of considerably more interest than the others.

In their 6 April 2020 paper, *"ACE2 variants underlie interindividual variability and susceptibility to COVID-19 in Italian population,"* Elisa Benetti *et. al.* also found p.Asn720Asp to be of interest. The authors used computational molecular dynamics simulations to identify thirty-three ACE2 variants most likely to bind with the SARS-CoV-2 virus. Thirty of the thirty three variants are too rare to be the culprit. The authors focused on the remaining three: p.Asn720Asp, p.Lys26Arg, and p.Gly211Arg. The first two had been identified by Cao *et. al.* as two of their seven hotspot variants.

Regarding the worldwide distribution of their three candidate variants, Benetti *et. al.* wrote "we identified three variants [...] moderately expressed in the Italian and European-non Finnish populations and with a very low allele frequency or not occurring in the Eastern Asia population." The authors made no observation regarding the frequency of the variants within Southern Asian, African, or Latin American populations.

More refined geographic and frequency filtering further reduces the surviving trio to a single variant, p.Asn720Asp. That is the same variant that Cao *et. al.* identified as a hotspot variant. The p.Lys26Arg variant is found too frequently in Africa relative to Europe, and p.Gly211Arg is found in only 0.21% of the European non-Finnish population.

In their conclusion, Benetti *et. al.* observe that, "it is noteworthy that ACE2 is located on chromosome X and that given the low allele frequency of the identified variants the rate of homozygous women is extremely low. Therefore, the impact of X-inactivation on the alternate expression of the two alleles would guarantee a heterogeneous population of ACE2 molecules, some of which [would be] protective towards the infection [...] This hypothesis would justify the high rate of asymptomatic or paucisymptomatic patients."

In his 2 May 2020 paper, *"Coding variants in ACE2 and TMPRSS2 are not major drivers of COVID-19 severity in UK Biobank,"* David Curtis presents evidence that no ACE2 variant is responsible for reportable COVID-19 infections. After reviewing 510 ACE2 variants for 74 British patients hospitalized with COVID-19 infections, he found no significant difference in frequencies between those suffering a serious infection and everyone else in the general population. Curtis, however, did not include the specific ACE2 variants as part of his data presentation, to protect patient privacy. Unfortunately, that leaves unanswered whether or not p.Asn720Asp was one of the variants he screened. Given that he screened 510 ACE2 variants compared to 808 ACE2 variants in my database, it seems more likely than not that p.Asn720Asp was among the list of variants he excluded as possible culprits.

ACE2 ACCOMPLICES

The current understanding is that binding to the ACE2 protein is a necessary but not sufficient condition for the SARS-CoV-2 virus to make its way inside the target cell. Most researchers now agree that the SARS-CoV-2 virus, once bound to the cell's ACE2 host receptor, makes use of TMPRSS2 cathepsin proteins and/or furin, as a pathway through the cell's protective membrane.

Transmembrane protease serine 2 is an enzyme encoded by the TMPRSS2 gene. A transmembrane protein is one that that spans a cell's membrane, rather than existing merely on the outer surface of the cell wall. Many transmembrane proteins allow transport of substances across the membrane. It thus makes sense that the SARS-CoV-2 virus might bind to the ACE2 protein, a receptor on the cell's outer surface, then gain cell entry via TMPRSS2, a transmembrane protein.

As is TMPRSS2, cathepsins are proteases, meaning they are enzymes that break down proteins to effect various biological functions. Cathepsins are similar to TMPRSS2 also in that they could conceivably provide a viral pathway for the SARS-CoV-2 virus. In their 16 April 2020 paper, *"SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor"*, Markus Hoffman *et. al.* write, "SARS-CoV can use the endosomal cysteine proteases cathepsin B and L (CatB/L) and the serine protease TMPRSS2 for S protein priming in cell lines, and inhibition of both proteases is required for robust blockade of viral entry." CatB is also known as CTSB. CatL is also known as CTSL.

Some proteins are inactive when they are first synthesized. They must have segments removed in order to become active. As are TMPRSS2, CTSB, and CTSL, furin is an enzyme that activates proteins by cleaving them. In their 21 May 2020 paper, *"A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells,"* Markus Hoffman *et. al.* explain the essential role of furin. "Our results reveal commonalities between the proteolytic activation of SARS-CoV-2 and MERS-CoV. Both viruses depend on furin-mediated pre-cleavage of their S proteins at the S1/S2 site for subsequent S protein activation by TMPRSS2 in lung cells, which fail to express robust levels of cathepsin L."

My database includes 907 genetic variants for the TMPRSS2 gene, 1110 for the CTSB gene, 364 for CTSL, and 989 for furin. None of the variants survive frequency, geographic, and

chromosomal filtering. Four TMPRSS2 variants, located on chromosome 21, survive frequency and geographic filtering. Only a single furin variant, located on chromosome 15, survives frequency and geographic filtering. None of the CTSB or CTSL variants, located on chromosomes 8 and 9, survive frequency and geographic screening.

Other evidence further excludes TMPRSS2 and cathepsin variants as culprits. As mentioned during discussion of ACE2, Curtis compared the frequency of ACE2 and TMPRSS2 variants from 74 British hospitalized with COVID-19 against corresponding frequencies among the general population. He found no significant differences. Curtis considered 658 TMPRSS2 variants, compared to 907 in my database. It is more likely than not that Curtis considered one or more of the four TMPRSS2 variants that survived frequency and geographic filtering.

Furthermore, as mentioned recently, Hoffman *et. al.* report that the SARS-CoV-2 virus can use either TMPRSS2 or a cathepsin to cross the cell membrane. If correct, this eliminates the possibility that any single TMPRSS2 or cathepsin variant is responsible for serious COVID-19 infections.

In sub-summary, none of the proteins hijacked by the SARS-CoV-2 virus for cell entry (ACE2, TMPRSS2, CTSB, CTSL, and furin) likely allow surface binding or transmembrane transport via a single variant.

FUNCTIONAL FILTERING OF GENETIC VARIANTS

Type I interferons are critical components of the innate immune system, the first line of defense against invading viruses. In fact, the name "interferon" stems from the protein's ability to interfere with virus replication. The type I interferons of interest to us are IFN-α and IFN-β.

Under normal circumstances, a virus-infected cell will release type I interferons. The interferons will alert neighboring cells to heighten their viral defenses and thereby activate scores of ISGs, interferon stimulated genes. The COVID-19 pandemic is, however, an abnormal circumstance. The SARS-CoV-2 virus somehow manages to prevent cellular expression of type I interferons, at least in those hosts who suffer severe infections. In their 23 April 2020 paper, "*Impaired type I interferon activity and exacerbated inflammatory responses in severe Covid-19 patients,*" Jerome Hadjadj *et. al.* wrote, "Consistently, plasma levels of IFN-α2 protein [...] were significantly lower in critical than in mild-to-moderate patients [...] while IFN-β was undetectable in all of the patients. [...] We propose that type-I IFN deficiency in the blood is a hallmark of severe Covid-19 and could identify and define a high-risk population."

The figure below, a simplified composite of several simplified images from the InvivoGen web site, provides a crude representation of how an infected cell should detect the SARS-CoV-2 virus and how it should then generate not only type I interferons, but also pro-inflammatory cytokines. By examining the figure in detail, we might be able to narrow our search for the culprit genetic variant.
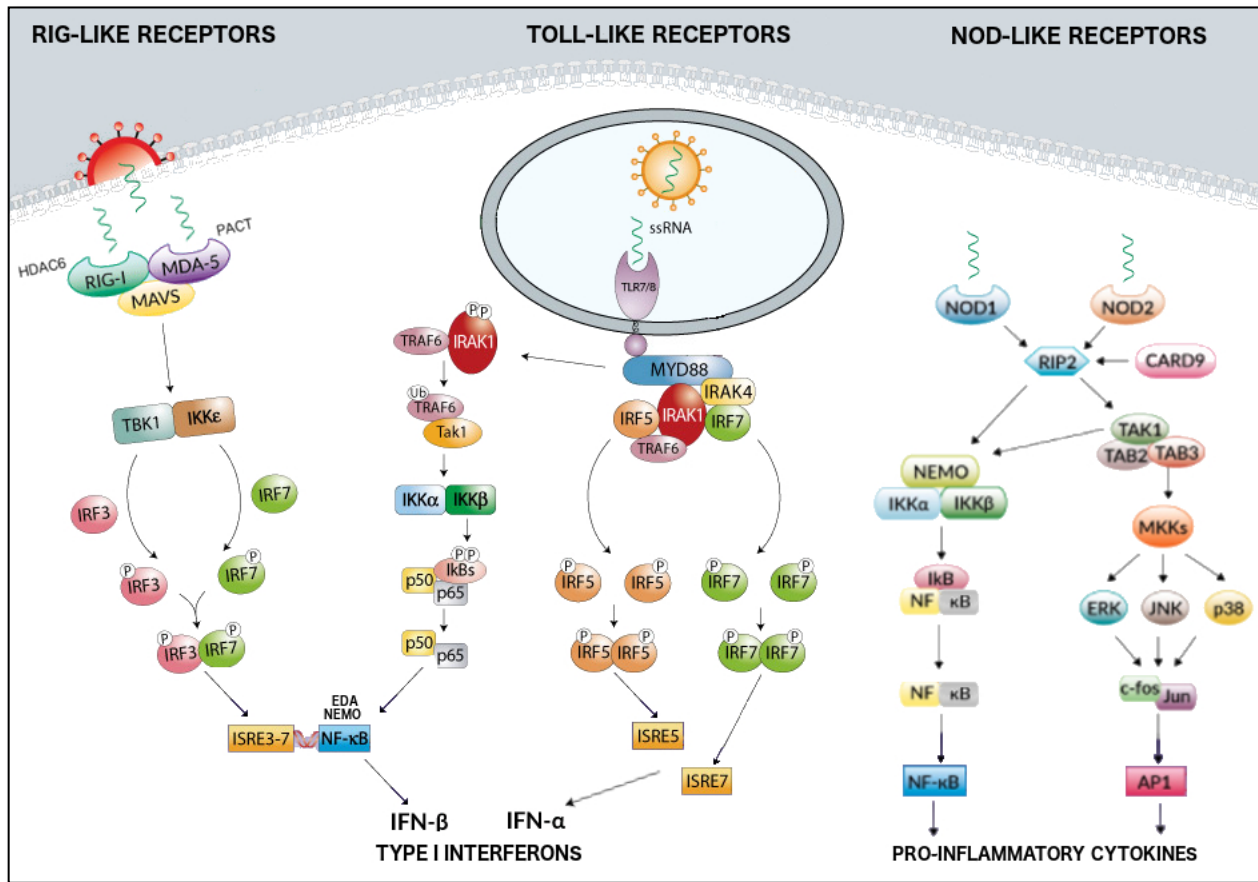
Figure 10: Viral Detection and Signaling Pathways for the Innate Immune System

At the upper left, a SARS-CoV-2 virus is shown entering the cell. RIG-like receptors (RIG-I, MDA-5) and toll-like receptors (TLR7, TLR8) detect the virus and launch multiple signaling cascades that lead to production of type 1 interferons. NOD-like receptors (NOD1, NOD2) also detect the virus and they launch multiple signaling cascades that lead to production of pro-inflammatory cytokines. Somewhere between viral entry and interferon production, the SARS-CoV-2 virus apparently manages to evade detection or disrupt signaling, at least among that portion of the population subject to severe infections.

Humans have two somewhat independent immune systems. The most ancient and faster to respond is the innate system, crudely portrayed in Figure 10. Animals, plants, fungi, even microscopic multi-cellular organisms have some form of an innate immune system. While innate systems are unable to detect and react to specific pathogens, they can detect DNA fragments, RNA fragments, and replication debris common to most pathogens. Upon detecting an unidentified pathogen, the innate system responds in a fashion that works reasonably well against most invaders.

Vertebrates have evolved a second immune system, the adaptive system. Instead of a generalized system, the adaptive system is pathogen specific. When informed by the innate immune system of a previously unknown pathogen, the adaptive immune system teaches itself to recognize the pathogen. It then builds and maintains pathogen-specific T and B memory cells that can launch custom-tailored defenses to reinforce the generalized defense of the innate system. Since the learning process requires days, the adaptive system responds slowly for the first invasion of any particular pathogen, but responds quickly to subsequent

infections. Vaccinations can be used to pre-train the adaptive system by introduction of a relatively harmless copycat pathogen.

In most COVID-19 infections, the innate immune system responds appropriately to SARS-CoV-2 infection, holding the virus at bay until the adaptive system has learned to recognize and defeat the invader. In some relatively small fraction of the human population, the SARS-CoV-2 virus manages to somehow evade or defeat the RLR and TLR signaling, leaving only NLR signaling to orchestrate an initial defense. When the RIG-like and toll-like receptors do not trigger an adequate type I interferon response, the NOD-like receptors might over-compensate with an excess of pro-inflammatory chemokines, leading to what is called the cytokine storm.

In the summary of their 26 April 2020 paper, "*Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19*," Daniel Blanco-Melo *et. al.* write, "Cell and animal models of SARS-CoV-2 infection [...] consistently revealed a unique and inappropriate inflammatory response. This response is defined by low levels of type I and III interferons juxtaposed to elevated chemokines and high expression of IL-6. We propose that reduced innate antiviral defenses coupled with exuberant inflammatory cytokine production are the defining and driving features of COVID-19."

This insight allows another level of filtering for the candidate culprit variants. Functional screening will exclude any variant not somehow involved, even remotely, with the production of type I interferons.

Application of the functional filter, along with the frequency, geographic, and chromosomal filters, reduces the list of candidate culprits by 99.999%, leaving only 8 variants for detailed consideration. The 8 remaining candidate culprits are summarized in Table 1, immediately below.

Ch	Gene	rsID	Consequence	Function	MAF EUR	MAF EAS	MAF AFR	MAF SAS	MAF LAT
X	EDA	rs140058036	c.741+47G>C	interferon production	0.0577	0.0004	0.0082	0.0175	0.0115
X	HDAC6	rs61735967	p.Arg832His	interferon production	0.0359	0.0000	0.0062	0.0035	0.0086
X	HDAC6	rs150080446	c.1623-42C>T	interferon production	0.0279	0.0000	0.0033	0.0053	0.0121
X	TLR7	rs5743780	p.Leu345Leu	Interferon production	0.0226	0.0000	0.0026	0.0065	0.0052
X	HDAC6	rs45441102	c.222+62C>T	interferon production	0.0196	0.0000	0.0022	0.0058	0.0090
X	IRAK1	rs11465836	c.1028+60C>A	interferon production	0.0129	0.0004	0.0017	0.0018	0.0021
X	IRAK1	rs192714554	c.1539+37G>C	interferon production	0.0118	0.0004	0.0015	0.0019	0.0012
X	NEMO	rs179363896	p.Asp181Asn	interferon production	0.0112	0.0000	0.0019	0.0021	0.0025

Table 1: Initial List of Variants Surviving Multi-Layer Filtering

RIG-LIKE RECEPTORS

Viruses have developed a broad range of countermeasures against innate immune systems. As large RNA viruses, coronaviruses are particularly well-suited for the task. They contain the genetic space to encode for multiple proteins that might interact with host proteins critical to detection of the virus, or signaling the presence of the virus. Both the 2002 SARS and the 2012 MERS coronaviruses, for example, manage to subvert RIG-I/MDA-5 detection and signaling component of the innate immune system.

Regarding the 2002 SARS-CoV, Jessica Chiang *et. al.* identify several of the thwarting mechanisms in their 2014 paper, "*Regulation of RIG-I-like receptor signaling by host and viral proteins.*"

> "SARS-CoV, using its M protein, blocks the RLR-induced IFN-b gene expression by interacting with RIG-I or key proteins in the RLR pathway, such as TANK, TBK1, IKK-e,

and TRAF3. [...] Two additional SARS-CoV proteins [...] have been shown to block the interaction between RIG-I and MAVS. Both proteins localize to mitochondria where they likely bind to either RIG-I or MAVS to inhibit IFN induction."

Regarding the 2012 MERS-CoV, Chiang *et. al.* write:

"Much research has been focused on understanding the pathogenicity of MERS-CoV in an attempt to limit its transmission and disease. These studies showed that while MERS-CoV replication is impaired by IFN treatment in vitro and in vivo, it fails to induce high levels of IFN during infection, indicating that this virus has effective mechanisms of IFN antagonism. The 4a, 4b, and M proteins of MERS-CoV were identified as potent antagonists of the type-I IFN response. The 4a protein specifically blocked IFN induction mediated by MDA5, but not RIG-I, an activity thought to be due to the predicted dsRNA binding ability of 4a. Subsequent studies revealed that the MERS-CoV 4a protein interacts with PACT, thereby preventing optimal ATPase and signaling activities of not only MDA5 but also RIG-I."

It has not been firmly established that the SARS-CoV-2 virus evades or corrupts the RLR detection and signaling pathway. It seems likely, though, that it does, given the low levels of type 1 interferons in patients with severe infections, given also the similarities among bat-related coronaviruses. In their 6 April 2020 paper, "*Immunological lessons from bats, pangolins and old coronaviruses; and how we can possibly apply them in a timely way for a better outcome*," Davika Dua *et. al.* explain:

"Post the 2002 SARS epidemic numerous studies have shown the presence of SARS like Coronaviruses (SL-CoV) in bats leading researchers to speculate that bats may be the ancestral host or natural reservoir for these viruses. SARS CoV 2 has been found to share 96% resemblance with RaTG1310, a bat coronavirus, which is more than the homology between it and the old SARS CoV (79%) and MERS-CoV (50%). [...] RIG-I is the most potent of RNA sensors and its function appears to be conserved between man and bat. In bats interferon induction machinery includes IRF-3 (Interferon regulatory element) and IRF7, which is also similar to humans [...] Bats also possess machinery which can limit the pro-inflammatory responses to coronavirus infection, something which humans lack [...] This cap on inflammation is believed to be one of the reasons why bats might have persistence of virus in their systems without developing signs of overt disease."

Without the RLR detection and signaling pathway, our innate immune system would need to rely on its toll-like receptors for interferon production. If the SARS-CoV-2 virus does indeed thwart RLR detection and signaling, the surviving, functional portion of our innate immune system can be approximated as in Figure 11, immediately below.
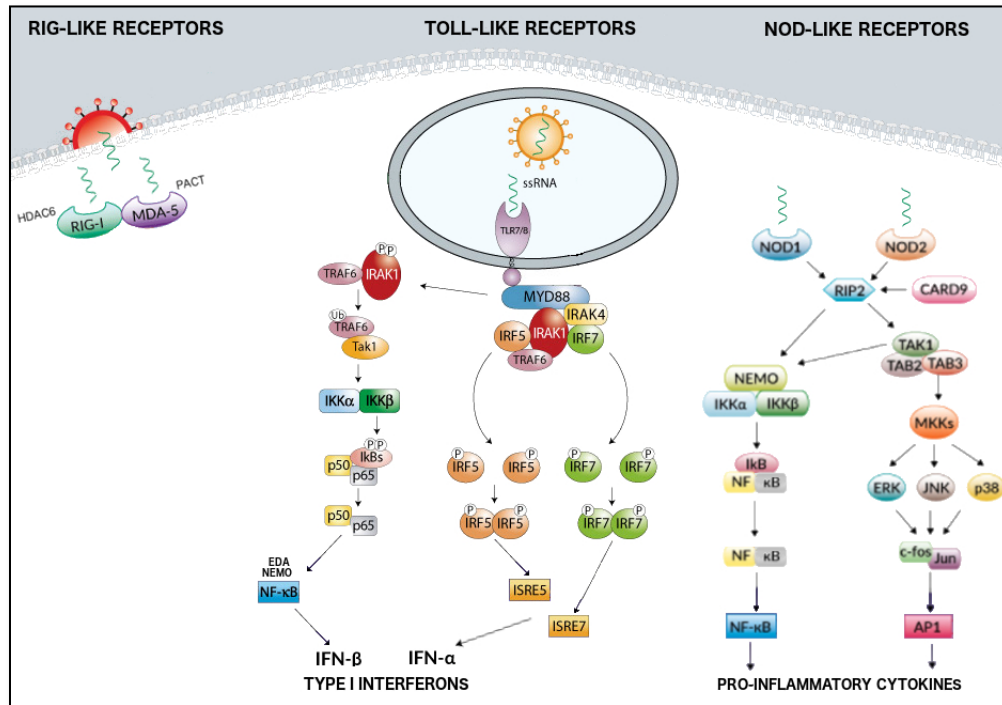
Figure 11: Innate Immune System with Compromised RLR Pathway

It seems, though, that the coronaviruses also manage to evade TLR detection and signaling, at least in some small fraction of humans, given the low levels of type I interferons associated with severe COVID-19 infections. By considering each of the variants that have survived frequency, geographic, chromosomal, and functional filtering, we narrow the list to a single candidate culprit.

TOLL-LIKE RECEPTORS

There are 10 toll-like receptor genes in the human genome. Of those ten, four are critical to viral detection for the innate immune system. TLR3, TLR7, TL8, and TLR9 detect viruses by binding to them. Upon detection, each initiates a signaling cascade that leads to production of type I interferons.

TLR9 detects DNA viruses. SARS-CoV02 is an RNA virus. Furthermore, the TLR9 gene is located on chromosome 3. No TLR9 variant, therefore, is likely to be the culprit for severe COVID-19 infections.

TLR3 detects double-stranded RNA viruses. SARS-CoV-2 is a single-stranded RNA virus. Furthermore, the TLR3 gene is located on chromosome 4. No TLR3 variant, therefore, is likely to be the culprit for severe COVID-19 infections.

TLR7 and TLR8, on the other hand, are both located on the X chromosome, and either of them should detect the SARS-CoV-2 virus. If they both fail, our type I interferon signaling system would be completely compromised, as shown in Figure 12 immediately below.
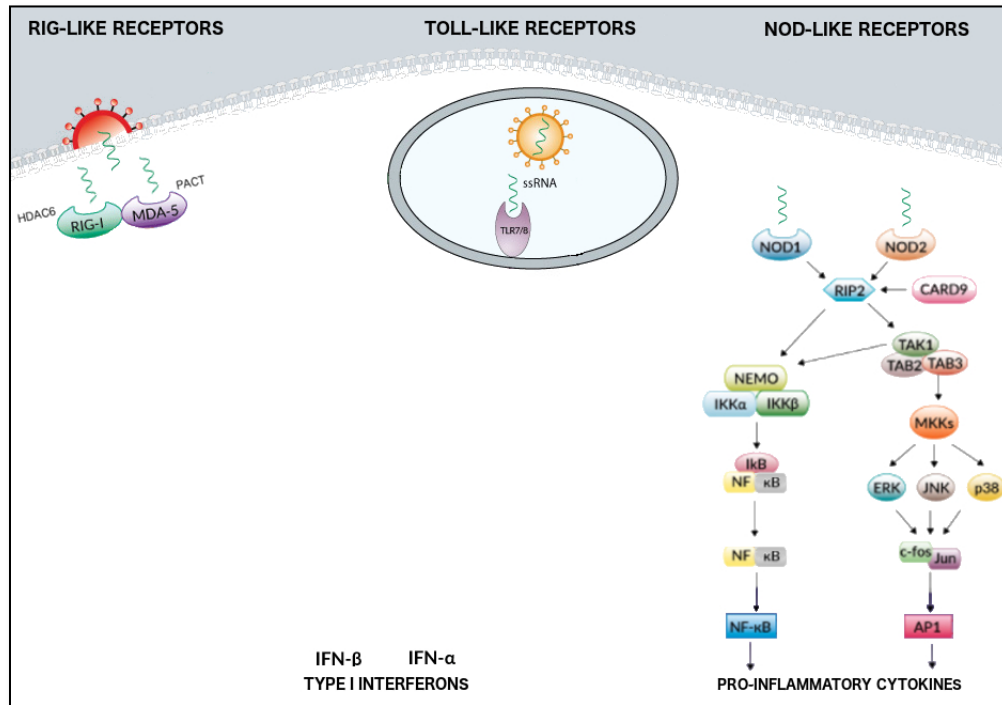
Figure 12: Innate Immune System with Compromised RLR and RLR Pathways

Of the 4693 TLR7 and TLR8 variants in the database, only one survives frequency, geographic, chromosomal and functional filtering. The single surviving candidate culprit is the TLR7 variant included in Table 1.

It is unlikely, though, that the TLR7 variant is actually responsible for severe COVID-19 infections via a complete shutdown of type I interferon production. The TLR8 receptor should still be capable of interferon signaling even in the presence of a disabled TLR7 receptor.

Mario Adan Moreno-Eutimio *et. al.* report that both TLR7 and TLR8 are capable of detecting SARS-CoV-2. In their 23 April 2020 paper, "*Bioinformatic analysis and identification of single-stranded RNA sequences recognized by TLR7/8 in the SARS-CoV-2, SARS-CoV, and MERS-CoV genomes*," they explain that "Our bioinformatic analysis showed that the SARS-CoV-2 genome contains a large number of fragments that could be recognized by TLR7/8, and it even contains more fragments than the SARS-CoV genome. This result suggests the ability to induce a rapid type I interferon response. The production of type I IFNs [...] plays a central role in the induction of antiviral responses."

Interestingly, TLR7 is inhibited by chloroquine. Chloroquine is a well-established malaria treatment and speculative COVID-19 treatment. Based only on the discussion so far, it seems unlikely that a TLR7 inhibitor, and therefore a type I interferon inhibitor, would prove to be an effective medication for SARS-CoV-2 infection.

After the world's first controlled, randomized, double-blind test of hydroxychloroquine, David Boulware *et. al.* report that administration of the TLR7 inhibitor was indeed ineffective. In their 3 June 2020 paper, "*A Randomized Trial of Hydroxychloroquine as Postexposure Prophylaxis for Covid-19*," they reported that hydroxychloroquine was ineffective as a COVID-19 prophylactic. Significantly, they reported that, of 107 study participants who contracted a COVID-19 infection, only two required hospitalization, one from the study group and one from the control group.

That finding is of considerable interest here. Presuming the SARS-CoV-2 virus already disabled the RLR interferon pathway, but not the TLR pathway, providing the chloroquine based drug would have compromised the TLR pathway via TLR7 inhibition. Those in the test group being provided the chloroquine-based drug should have suffered severe infections at a noticeably higher rate than those in the control group who were administered a placebo. The absence of differential outcome not only indicates that hydroxychloroquine is ineffective as a prophylactic, it suggests either that TLR8 can completely replace TLR7 or that the SARS-CoV-2 virus somehow thwarts the entire TLR detection and signaling pathway.

The latter possibility is particularly curious. It would leave unexplained how the virus manages to subvert the TLR interferon production pathway. More significantly, it would leave unanswered how most COVID-19 infections are asymptomatic or paucisymptomatic.

EDA / NEMO

The EDA gene, located on the X chromosome, encodes the ectodysplasin A (EDA) protein. Among its other functions, the EDA protein regulates NF-κB.

The IKBKG gene, located on the X chromosome, encodes the IKBKG protein. The protein's alias, NEMO, is derived from the protein's role as an **N**F-κB **e**ssential **mo**dulator.

NF-κB (nuclear factor kappa-light-chain-enhancer of activated B cells), among its other functions, is part of the signaling cascade of for both type I interferons and pro-inflammatory cytokines. With respect to interferon, it signals the production of IFN-β interferons.

In theory, a variant of either the EDA or NEMO genes could result in a protein that fails to properly regulate NF-κB. That would interfere with signaling both for IFN-β interferons and pro-inflammatory cytokines, as displayed below in Figure 13.
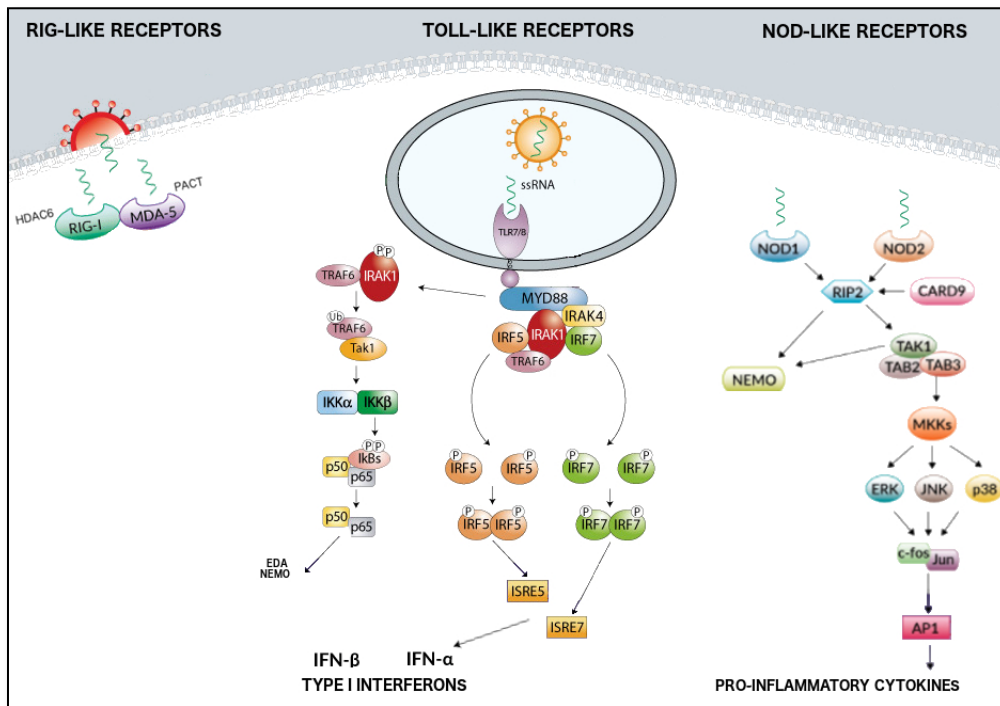

Figure 13: Possible Effect of EDA or NEMO Variant

Recall that Hadjadj *et. al.* noted that, "plasma levels of IFN-α2 protein [...] were significantly lower in critical than in mild-to-moderate patients [...] while IFN-β was undetectable in all of

the patients." While disruption of EDA and NEMO could be responsible for "undetectable" IFN-β, neither seemingly could be responsible for "significantly lower" levels of IFN-α. Furthermore, inhibition of NF-κB, either by EDA or NEMO, would disrupt one of the signaling pathways for the pro-inflammatory cytokines.

Since both IFN-α and IFN-β are seriously under-expressed in severe COVID-19 infections, and since pro-inflammatory cytokines are seriously over-expressed, it seems unlikely that either an EDA or NEMO variant is the culprit for reportable COVID-19 infections.

IRAK1

The IRAK1 gene, located on the X chromosome, codes for the interleukin-1 receptor-associated kinase 1 (IRAK1) enzyme. An enzyme is a protein that accelerates certain chemical reactions. IRAK1 is involved in both the innate and adaptive immune systems. Within the innate immune system, IRAK1 resides along each of the two signaling cascades between the toll-like receptors (TLR7 and TLR8) and type I interferon production, both IFN-α and IFN-β.

Even if the SARS-CoV-2 virus defeats RLR detection and signaling, a fully functional TLR pathway, producing both IFN-α and IFN-β interferons, could prevent most COVID-19 infections from becoming serious. If, however, some variant of IRAK1 deprived some percentage of the population of their TLR-based defense, then those unfortunate few would be subject to severe COVID-19 infection accompanied by an inflammatory cytokine storm, as indicated below in Figure 14.
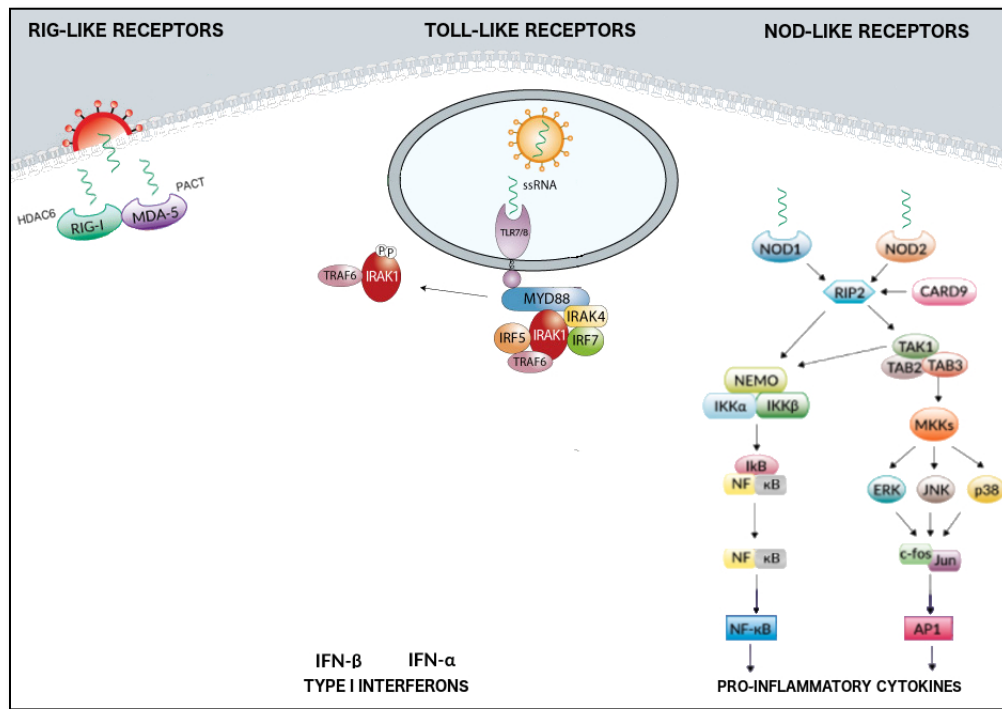

Figure 14: Possible Effect of IRAK1 Variant

Given that the SARS-CoV-2 virus likely evades RLR detection and signaling, and given that IRAK1 resides in each of the two TLR detection and signaling pathways, it seems possible that an IRAK1 variant could indeed be the culprit for severe COVID-19 infections.

There is a confusing twist, though. Within the innate immune system, IRAK1 works in conjunction with TRAF6, and the SARS-CoV-2 virus apparently subverts TRAF6.

In their 2014 paper, "*The SARS-coronavirus Papain-Like Protease: Structure, Function and Inhibition by Designed Antiviral Compounds*," Yahira Báez-Santos *et. al* forewarned us of the COVID-19 pandemic, and identified the SARS PLpro protease as an attractive target for antiviral drugs.

"Over 10 years have passed since the deadly human coronavirus that causes severe acute respiratory syndrome (SARS-CoV) emerged from the Guangdong Province of China. Despite the fact that the SARS-CoV pandemic infected over 8500 individuals, claimed over 800 lives and cost billions of dollars in economic loss worldwide, there still are no clinically approved antiviral drugs, vaccines or monoclonal antibody therapies to treat SARS-CoV infections. The recent emergence of the deadly human coronavirus that causes Middle East respiratory syndrome (MERS-CoV) is a sobering reminder that new and deadly coronaviruses can emerge at any time with the potential to become pandemics. Therefore, the continued development of therapeutic and prophylactic countermeasures to potentially deadly coronaviruses is warranted. The coronaviral proteases, papain-like protease (PLpro) and 3C-like protease (3CLpro), are attractive antiviral drug targets because they are essential for coronaviral replication. [...] PLpro has the additional function of stripping [...] host-cell proteins to aid coronaviruses in their evasion of the host innate immune responses."

In their 2016 paper, "*SARS Coronavirus Papain-Like Protease Inhibits the TLR7 Signaling Pathway through Removing Lys63-Linked Polyubiquitination of TRAF3 and TRAF6*," Shih-Wen Li *et. al.* identified TRAF6 as being among the host-cell proteins disabled by SARS PLpro. "In summary, SARS-CoV PLPro inhibits TLR7-mediated signaling [...] PLPro diminishes Lys63-linked ubiquitination of TRAF3 and TRAF6 and then inactivates their downstream molecules [...] The results let us conclude that SARS-CoV PLPro negatively regulates the [...] TLR7-mediated antiviral and inflammatory responses."

If the SARS-CoV-2 virus does indeed disable TLR signaling downstream of the TRAF6 molecule, then the virus is able to disrupt type I interferon production completely, whether signaled by a RIG-like receptor or a toll-like receptor. No IRAK1 variant would be required to expose the host to a severe COVID-19 infection. Nearly everyone contracting the virus would suffer a severe infection, which is not the case.

The logic chain constructed so far cannot be both accurate and complete. Something is wrong or something is missing.

HDAC6

Histone deacetylase 6 is an enzyme encoded by the HDAC6 gene, which is located on the X chromosome. HDACs are enzymes that remove acetyl groups from various proteins. An acetyl group is a specific arrangement of 1 oxygen, 2 carbon, and 3 hydrogen atoms.

When a HDAC deacetylizes a protein, it changes that protein's behavior. Of particular interest here is the interaction between HDAC6 and the RIG-I receptor. When HDAC6 detects a virus, it deacetylizes the RIG-I receptor, essentially waking it from slumber. The RIG-I receptor then too detects the virus. The RIG-I receptor then signals for type I interferon production.

With respect to the SARS-CoV-2 virus, this interaction between HDAC6 and RIG-I is of no immunity benefit if the virus interrupts the RLR signaling cascade at TANK, TBK1, IKK-e, or TRAF3, as Jessica Chiang *et. al.* explain happens with SARS-CoV-1.

Significantly, Kai Zeng *et. al.* indicate that HDAC6 itself might be able to more directly signal type I interferon production via IRF3, bypassing RIG-I, TANK, TBK1, and IKK-e in the process.

In their 2016 paper, "*Cellular defence or viral assist: the dilemma of HDAC6*," they write, "More specifically, HDAC6 functions as a coactivator of virus-induced IRF-3-dependent transcription." They present a figure showing dual signaling pathways for HDAC6, one through RIG-I and the other independent of RIG-I. In the graphical representation being used herein, the alternate HDAC6 pathway is portrayed in Figure 15, immediately below.
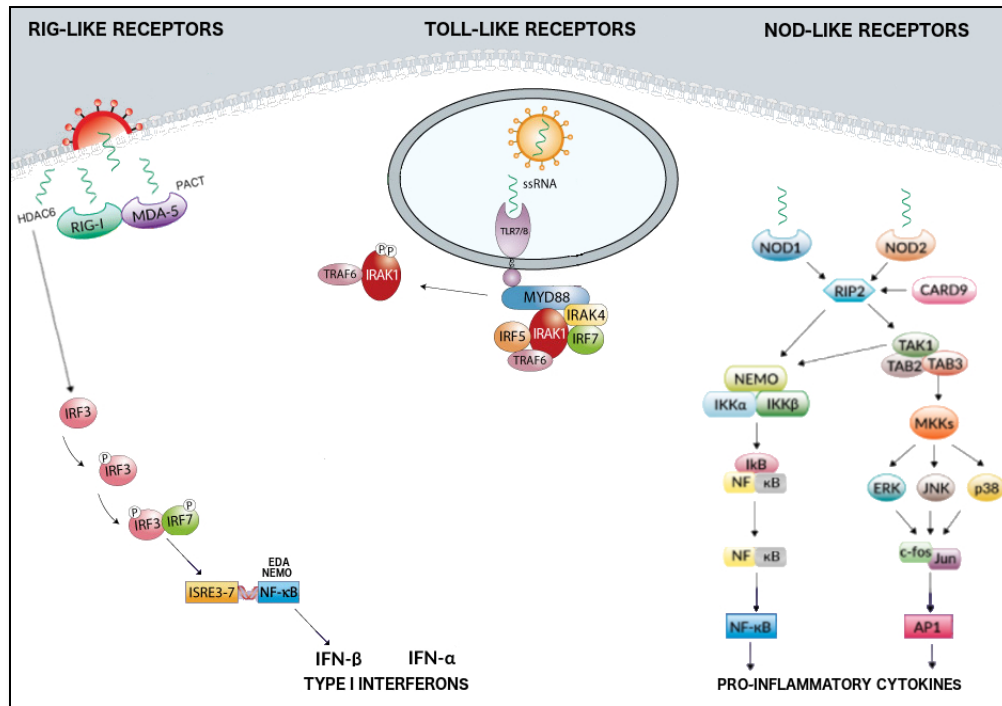

Figure 15: Alternative HDAC6 Detection and Signaling Pathway

Not only might HDAC6 combat the SARS-CoV-2 virus by signaling type I interferon production, it might combat the virus more directly by impairing the virus's ability to fuse with the host cell prior to entry. Regarding such a role in combating HIV, Zeng *et. al.* write, "In this respect, the fusion step acts as the main target for inhibition of [HIV] infection, in which HDAC6 functions as a restriction factor"

HDAC6 might also impair viral entry. Regarding such a role in combating the bird flu, Zeng writes, "Furthermore, the antiviral effect of HDAC6 on virus entry and infection has been examined in vivo, which shows that HDAC6 overexpression significantly enhances resistance to avian H5N1 virus infection and extends the survival rate."

HDAC6 might also impair viral replication. Regarding such a role in combating HIV, Zeng writes, "HDAC6 is capable of regulating the transactivation activity of Tat, a […] protein essential for HIV-1 transcription and replication."

Once a virus replicates itself inside a host cell, the copies construct protective shells, exit the cell, and proceed to infect other cells. HDAC6 may interfere also with these final intracellular actions. Zeng writes, simply, "HDAC6 may impair viral assembly and release."

Not only might HDAC6 play a vital, life-saving role as part of the innate immune system, it seems to play a valuable role in the adaptive immune system as well. Zeng writes, "Besides viral recognition in innate immunity, HDAC6 is also essential for antigen presentation and cytotoxic T lymphocyte (CTL)-mediated lysis of virus-infected cells in adaptive immune response, in which viruses have exploited various strategies to [evade] T cell immunity."

It now seems possible that HDAC6 might provide considerable innate immune protection even if the SARS-CoV-2 virus disrupts both the RLR and TLR detection and signaling pathways. It now seems possible that the relatively few people who carry an HDAC6 gene variant lack the HDAC6 proteins necessary to hold the SARS-CoV-2 virus at bay. Those unfortunate few would be subject to severe, even fatal infection.

CONCLUSION

The worldwide distribution of reportable COVID-19 infections suggests that severe infections are related to a genetic variant.

Through a combination of frequency, geographic, chromosomal, and functional filtering, the list of candidate culprit variants can be narrowed to three variants of the HDAC6 gene and two variants of the IRAK1 gene.

Ch	Gene	rsID	Consequence	Function	MAF EUR	MAF EAS	MAF AFR	MAF SAS	MAF LAT
X	HDAC6	rs61735967	p.Arg832His	interferon production	0.0359	0.0000	0.0062	0.0035	0.0086
X	HDAC6	rs150080446	c.1623-42C>T	interferon production	0.0279	0.0000	0.0033	0.0053	0.0121
X	HDAC6	rs45441102	c.222+62C>T	interferon production	0.0196	0.0000	0.0022	0.0058	0.0090
X	IRAK1	rs11465836	c.1028+60C>A	interferon production	0.0129	0.0004	0.0017	0.0018	0.0021
X	IRAK1	rs192714554	c.1539+37G>C	interferon production	0.0118	0.0004	0.0015	0.0019	0.0012

Table 2: Final List of Variants Surviving Multi-Layer Filtering

This conclusion can be tested by searching for disproportionally high frequencies of the five variants among individuals who have suffered severe COVID-19 infections, relative to a control group of individuals who have suffered asymptomatic or paucisymptomatic infections, or showed no infection whatsoever.

REFERENCES

1. *worldometers*; "COVID-19 Coronavirus Pandemic"
   Infections rates and country populations
   https://www.worldometers.info/coronavirus/

2. *Wikipedia*; "United Nations geoscheme"
   Categorization of countries by geographic region
   https://en.wikipedia.org/wiki/United_Nations_geoscheme

3. *Eupedia*; "Distribution of European Y-chromosome DNA (Y-DNA) haplogroups by country in percentage"
   Population numbers for haplogroups M269 and M343
   https://www.eupedia.com/europe/european_y-dna_haplogroups.shtml

4. *Wikipedia*; "Haplogroup R-M269"
   Population numbers for haplogroup M269
   https://en.wikipedia.org/wiki/Haplogroup_R-M269

5. Myres, Natalie M., *et al*; "A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe", Supplementary Table S4; *European Journal of Human Genetics*, published online 25 August 2010
   Comprehensive collection of population numbers for most R1b haplogroups
   https://www.nature.com/articles/ejhg2010146

6. *newgeography*; "Race, Ancestry and Genetic Composition of the U.S."
    State by state haplogroup population from which I calculated M269 percentage for US
    https://www.newgeography.com/content/005051-race-ancestry-and-genetic-composition-us

7. *Nature*; "Population structure of Y chromosome SNP haplogroups in the United States and forensic implications for constructing Y chromosome STR databases"
    https://www.sciencedirect.com/science/article/pii/S0379073805006158

8. *NDTV*; "Wuhan Starts Mass Testing Campaign After New COVID-19 Cases Emerge",
    Total COVID-19 cases in Wuhan = 50,339 as of 13 May 2020
    https://www.ndtv.com/world-news/coronavirus-china-wuhan-starts-mass-testing-campaign-after-new-covid-19-cases-emerge-2228311

9. *Intelligencer*; "Coronavirus in New York: Latest Updates"
    Total COVID-19 cases in New York City = 192,394 as of 14 May 2020
    https://nymag.com/intelligencer/article/new-york-coronavirus-cases-updates.html

10. *Wikipedia*; "Haplogroup R1b"
    Hotspot image for M343 haplogroup; converted to grayscale, label modified
    https://en.wikipedia.org/wiki/Haplogroup_R1b

11. *The Weather Channel*; "Coronavirus, Covid-19: Ufa, Republic of Bashkortostan, Russia"
    Total COVID-19 cases in Bashkortostan = 252,245 as of 14 May 2020
    https://weather.com/coronavirus/l/ecc562d6ac455484d9f5ceb6bfc963ba60bf38338297dc46e06130a48cac575d

12. *Wikipedia*; "Y-DNA haplogroups in populations of Central and North Asia"
    M343 (R1b) haplogroup population for Turkmens
    https://en.wikipedia.org/wiki/
    Y-DNA_haplogroups_in_populations_of_Central_and_North_Asia

13. *Wikipedia*; "COVID-19 pandemic in Italy"
    Hotspot image for Italy's COVID-19 infections; converted to grayscale; label added
    https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Italy

14. *Wikipedia*; "Human genetic variation"
    Source for SNP image
    https://en.wikipedia.org/wiki/Human_genetic_variation

15. *Wikipedia*: "dbSNP"
    Background on the NCBI genetic database
    https://en.wikipedia.org/wiki/DbSNP

16. *NCBI*; "dbSNP"
    Homepage for the NCBI genetic database
    https://www.ncbi.nlm.nih.gov/snp/

17. *InnateDB*, "InnateDB Innate Immunity Genes"
    List of genes involved with the innate immune system
    https://www.innatedb.com/annotatedGenes.do?type=innatedb

18. Schurz, Haiko; "The X chromosome and sex-specific effects in infectious disease susceptibility," *Human Genomics*, 8 January 2019

X-chromosome inactivation
https://doi.org/10.1186/s40246-018-0185-z

19. Cao, Yanan *et. al.*; "Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations"; *Cell Discovery,* 24 Feb 2020
ACE2 variants among East Asian and other populations; COVID-19
Source for pie chart distributions for Cao's seven hotspot variants
https://doi.org/10.1038/s41421-020-0147-1

20. Beneti, Elisa, *et. al.*; "ACE2 variants underlie interindividual variability and susceptibility to COVID-19 in Italian population," *medRxiv* preprint, 6 April 2020
ACE2 variants among the Italian population; COVID-19
https://doi.org/10.1101/2020.04.03.20047977

21. Hoffman, Markus, *et. al.*; "A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells," *Molecular Cell*, 21 May 2020
FURIN essential for SARS-CoV-2 entry into host cell
https://doi.org/10.1016/j.molcel.2020.04.022

22. Curtis, David; "Coding variants in ACE2 and TMPRSS2 are not major drivers of COVID-19 severity in UK Biobank subjects," posted online by *medRxiv*, 2 May 2020
ACE2 and TMPRSS variants not responsible for serious COVID-19 infections
https://doi.org/10.1101/2020.05.01.20085860

23. Hoffman, Markus; "SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor", *Cell*, 16 April 2020
Both TMPRSS2 and cathepsin B/L can effect SARS-CoV-2 viral entry into cells
https://doi.org/10.1016/j.cell.2020.02.052

24. Hadjadj, Jerome, *et. al.*; "Impaired type I interferon activity and exacerbated inflammatory responses in severe Covid-19 patients," *medRxiv*, 23 April 2020
SAVS-CoV-2 associated with under-expression of interferons
https://doi.org/10.1101/2020.04.19.20068015

25. *InvivoGen*
Source for images showing type I interferon and pro-inflammatory cytokine production. (Note: the modified images presented herein have been substantially altered from the originals; *InvivoGen* is not responsible for any errors introduced in the process.)
https://www.invivogen.com/review-type1-ifn-production
https://www.invivogen.com/spotlight-covid-19-predicted-immune-responses
https://www.invivogen.com/review-nlr

26. Blanco-Melo, Daniel, *et. al.*; "Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19," *Cell*, 26 April 2020
Severe infections reflect imbalanced response of innate and adaptive immune systems
https://doi.org/10.1016/j.cell.2020.04.026

27. Chiang, Jessica J., *et. al.*; "Regulation of RIG-I-like receptor signaling by host and viral proteins," *Elsevier*, 21 June 2014
SARS-CoV thwarts RIG-I detection and signaling pathway
http://dx.doi.org/10.1016/j.cytogfr.2014.06.005

28. Dua, Davika, *et. al.*; "Immunological lessons from bats, pangolins and old coronaviruses; and how we can possibly apply them in a timely way for a better outcome," *Preprints*, 6 April 2020
Similarity of SARS-CoV-2 and innate immune systems between bats and humans
https://doi.org/10.20944/preprints202004.0071.v1

29. Moreno-Eutimio, Mario Adan, *et. al.*; "Bioinformatic analysis and identification of single-stranded RNA sequences recognized by TLR7/8 in the SARS-CoV-2, SARS-CoV, and MERS-CoV genomes," *Elsevier*, 23 April 2020
TLR7 and TLR8 detect SARS-CoV-2
https://doi.org/10.1016/j.micinf.2020.04.009

30. de Groot, Natasja G. and Brontop, Ronald E.; "COVID-19 pandemic: is a gender-defined dosage effect responsible for the high mortality rate among males?" *Immunogenetics*, 28 April 2020
Identification of TLR7 and TLR8 as candidate culprits for COVID-19 infections
https://dx.doi.org/10.1007%2Fs00251-020-01165-7

31. Boulware, David R., *et. al.*; "A Randomized Trial of Hydroxychloroquine as Postexposure Prophylaxis for Covid-19," *New England Journal of Medicine*, 3 June 2020
Hydroxychloroquine ineffective as prophylactic; may indicates SAS-CoV-2 inhibits TLR7
Only of 2 of 107 participants who contracted COVID-19 were hospitalized
DOI: 10.1056/NEJMoa2016638

32. Báez-Santos Yahira M., *et. al.*; "The SARS-coronavirus Papain-Like Protease: Structure, Function and Inhibition by Designed Antiviral Compounds," *Elsevier*, 29 December 2014
SARS PLpro critical for viral replication
https://doi.org/10.1016/j.antiviral.2014.12.015

33. Li, Shih-Wen, *et. al.*; "SARS Coronavirus Papain-Like Protease Inhibits the TLR7 Signaling Pathway through Removing Lys63-Linked Polyubiquitination of TRAF3 and TRAF6," *International Journal of Molecular Sciences*, 5 May 2016
SARS viruses inhibit TRAF6
https://www.mdpi.com/1422-0067/17/5/678/html

34. Zeng, Kai, *et. al.*; "Cellular defence or viral assist: the dilemma of HDAC6," *Journal of General Virology*, 8 December 2016
HDAC6 might provide an alternate signaling pathway for type I interferons
https://doi.org/10.1099/jgv.0.000679