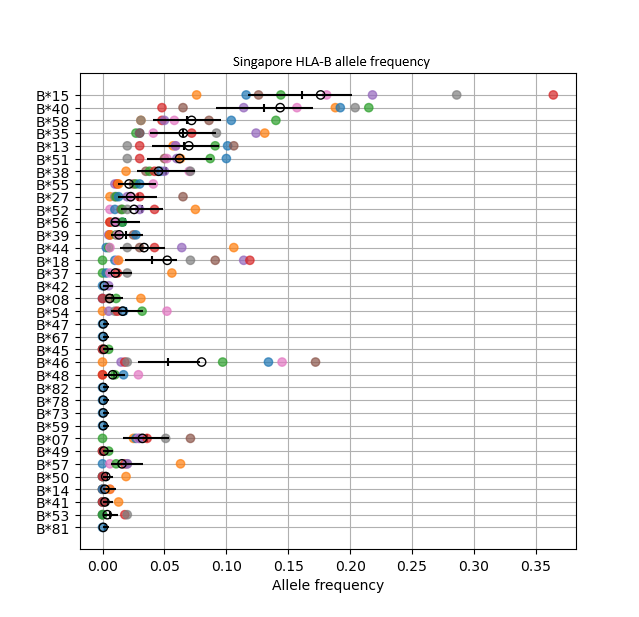
# Estimating disease rates without diagnosis

## Immune gene frequencies and rates of autoimmune disease

The HLA immune genes are largely responsible for triggering the immune system. This responsibility means we can predict a person’s immune response based on their HLA genes. HLA genes are associated with a person’s response to infection, vaccination, and often very strongly linked to autoimmune diseases. So strongly linked infact, that in large groups we can predict disease rates from HLA gene frequencies. HLA frequencies are widely studied and so often available, allowing us to estimate rates of autoimmune conditions which may be missing or inaccurate due to the challenges of diagnosis. In this post we’ll combine studies to generate accurate estimates of immune gene frequencies and use that to predict national rates of ankylosing spondylitis.

[allelefrequencies.net](http://allelefrequencies.net/) is a public database of human immune gene frequency data from across the world. However, it can be difficult to download and combine data from multiple projects; this makes it hard to take advantage of all this data. Luckily `[HLAfreq](https://barinthusbio.github.io/HLAfreq/HLAfreq.html)` is a python package which makes it easy to get the latest data from allelefrequencies.net and prepare them for our analysis. (Full disclosure, I am one of the authors of HLAfreq!).

Ankylosing spondylitis is a form of arthritis, and 90% of patients have a specific version of the HLA B gene. To get the frequency of this version in different countries I downloaded all available frequency for this gene and combined studies of the same country, weighting by sample size. In brief the combination is based on the Dirichlet distribution and we can use a Bayesian approach to estimate uncertainty too. Singapore is used as an example in the figure below. Different HLA-B gene versions (also known as alleles) are shown on the y axis with their frequency in Singapore on the x axis. Data from the original Singapore studies are shown in colour, and combined estimates in black. I focused on the weighted average in this analysis which is shown by the black circles. HLAfreq also calculates a Bayesian measure with uncertainty which is indicated by the black bars.



The code used to download, combine, and plot the HLA-B allele frequency data for Singapore.

```

# Download raw data

base\_url = HLAfreq.makeURL(“Singapore”, standard="g", locus="B")

aftab = HLAfreq.getAFdata(base\_url)

# Prepare data

aftab = HLAfreq.only\_complete(aftab)

aftab = HLAfreq.decrease\_resolution(aftab, 1)

# Combine data from multiple studies

caf = HLAfreq.combineAF(aftab)

hdi = HLAhdi.AFhdi(aftab, credible\_interval=0.95)

caf = pd.merge(caf, hdi, how="left", on="allele")

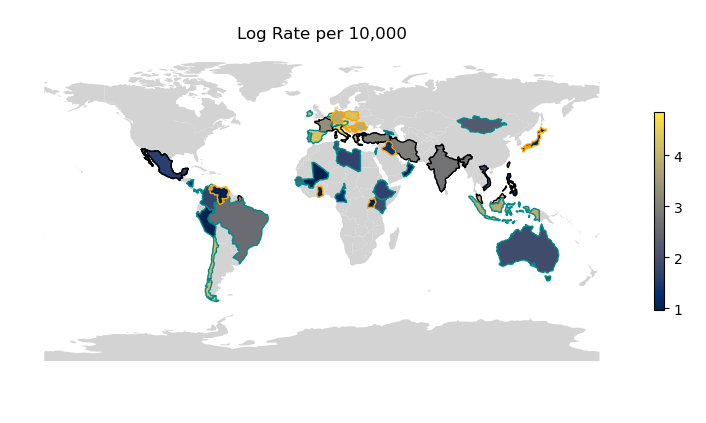
# Plot gene frequencies

HLAfreq.plotAF(caf, aftab.sort\_values("allele\_freq"), hdi=hdi, compound\_mean=hdi)

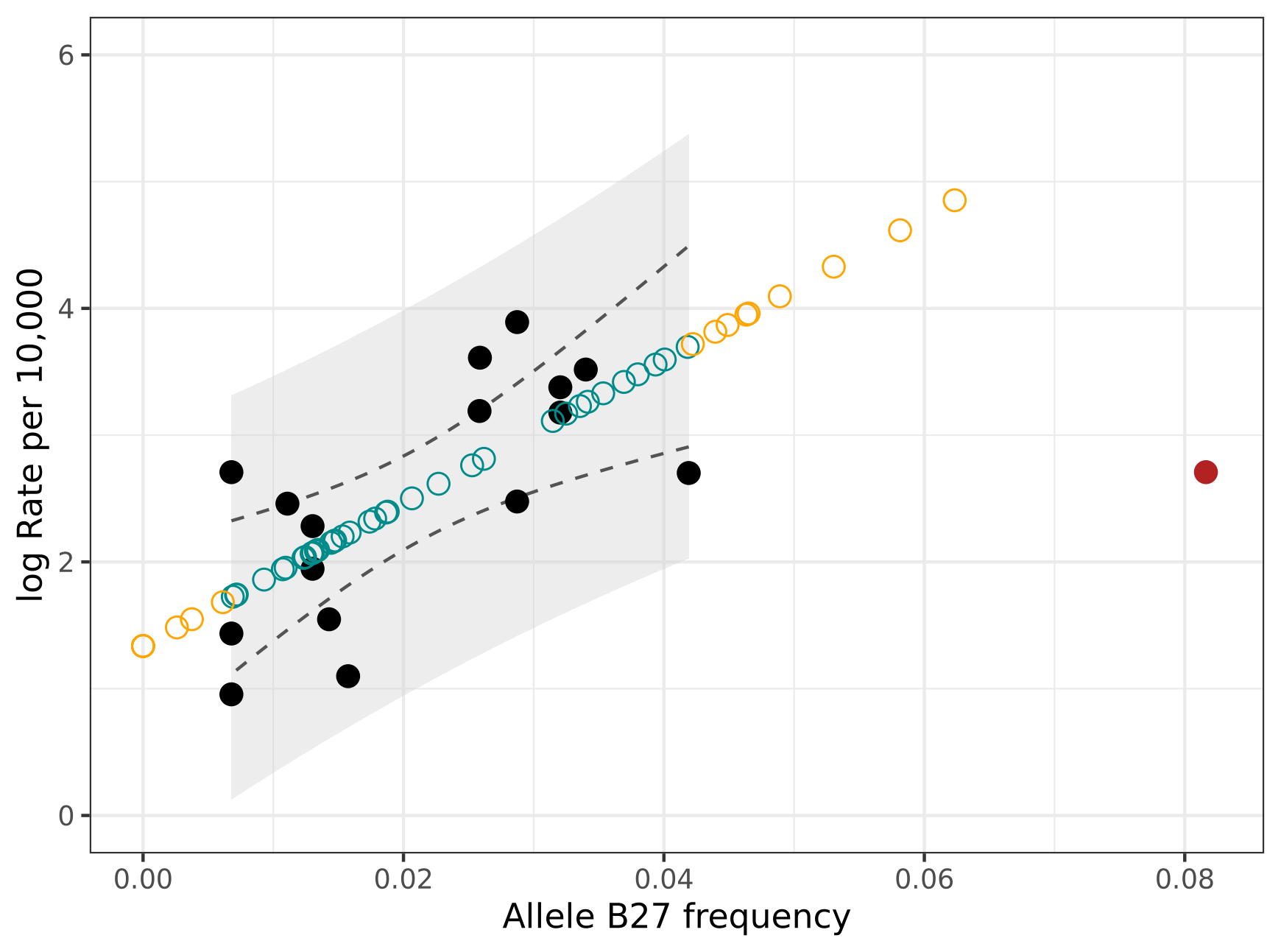
```

Now we have the national allele frequencies we can pair them with the national disease rates to study the correlation. I have used the disease rates reported in Dean et al 2014. I log transformed the disease rate to make it normally distributed so I could fit an ordinary least squares linear regression. As expected, there was a significant positive correlation; countries with higher frequencies of HLA-B\*27 had higher rates of ankylosing spondylitis. The exception to this was Finland which had an unusually high frequency of HLA-B\*27 but a middling rate of disease. I removed Finland from the model as an outlier, a decision which was supported by “statistical leverage”. Leverage means this one point had too large an influence on the overall model, we want the model to tell us about countries in general not any one country in particular.

We can use our linear regression model to predict rates of ankylosing spondylitis in countries where we know the HLA-B\*27 frequency. This tells us that countries like Austria and Croatia have high predicted ankylosing spondylitis rates. Using these predictions increases the number of countries with disease rate estimates from 16 to 52 and can help identify countries that could benefit from additional surveillance. In the world map below, countries with low known or predicted rates of ankylosing spondylitis are plotted in blue and high rates in yellow. Countries with known rates are outlined in black and those with predicted rates are outlined in cyan or orange. Cyan is used for countries in the range of our model and orange is used for countries outside our model’s range, see below for why this is important.



We should be cautious about predicting disease rates for countries with HLA-B\*27 rates outside of the range of our model. Of the 36 countries we have predicted disease rates for, 10 have HLA-B\*27 frequencies higher or lower than any country we used in our model. As these countries fall outside the range of our model we can’t be sure the model will give accurate predictions for these countries. In particular, predictions may be unreliable for countries with high HLA-B\*27 rates, we already know that Finland did not fit our model. This could be because of a non-linear trend but we do not have enough data to explore these high frequencies.

The countries with known disease rates are plotted with filled points, Finland which was omitted from the model is plotted in red. The predicted disease rates are plotted as open circles, cyan for countries in the model’s range and orange outside of it. The confidence intervals of the model are shown as dashed lines, and the prediction intervals are shown as a grey ribbon. A quick reminder about the difference: we expect the true relationship to fall within the credible intervals 95% of the time, and we expect 95% of data points to fall within the prediction intervals.

# Conclusion

HLA genes have a strong impact on human health through infection, vaccination, autoimmune diseases, and organ transplants. Because of these strong relationships, we can use widely available HLA frequency data to study these health traits indirectly. Resources like [allelefrequency.net](http://allelefrequencies.net/) and [HLAfreq](https://barinthusbio.github.io/HLAfreq/HLAfreq.html) make it easier to study these relationships, either by looking at these correlations directly or using allele frequencies as a proxy when other data is missing. I hope this post has got you thinking about questions to ask using HLA frequency data.

# References

Dean, L. E., Jones, G. T., MacDonald, A. G., Downham, C., Sturrock, R. D., & Macfarlane, G. J. (2014). Global prevalence of ankylosing spondylitis. *Rheumatology*, *53*(4), 650-657.