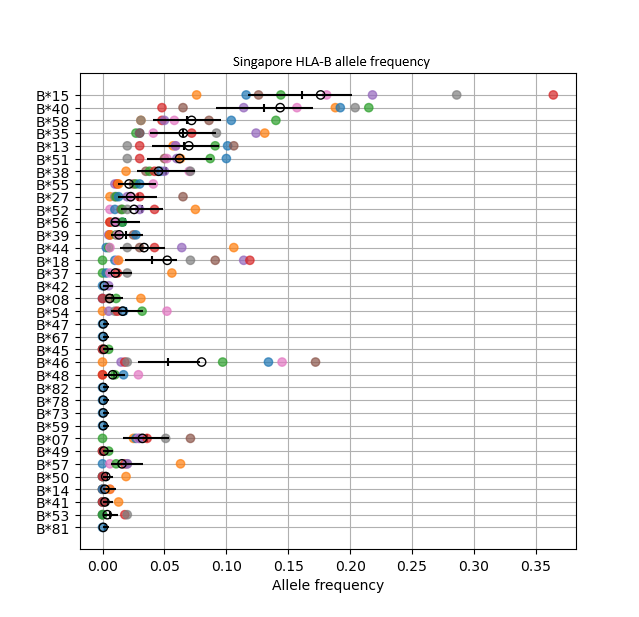
# Immune gene frequencies and rates of autoimmune disease

## Estimating national disease rates without diagnosis

The HLA immune genes are largely responsible for triggering the immune system. This responsibility means it is possible to prediction 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. 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 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 spondyloarthritis.

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 study larger regions or even single countries that have been involved in multiple studies. Luckily `HLAfreq` 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 spondyloarthritis is closely associated with 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 version (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 mean 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")

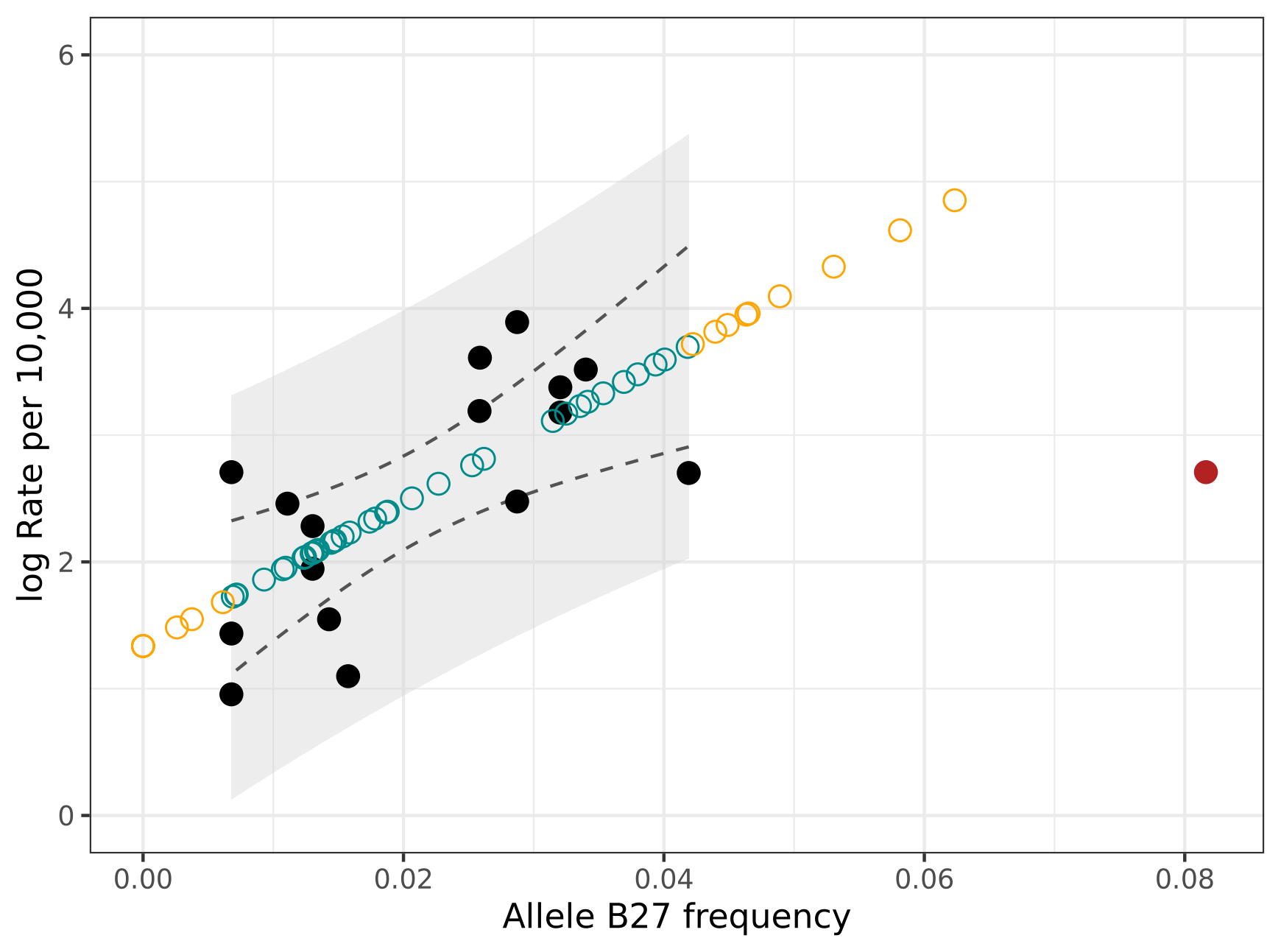
# Plot gene frequencies

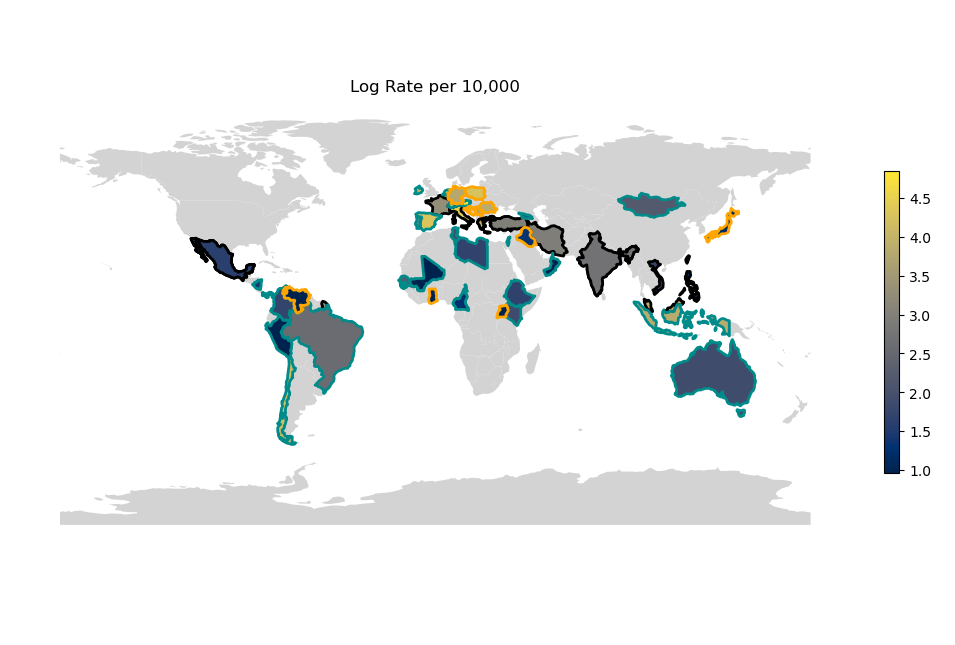
HLAfreq.plotAF(caf, aftab.sort\_values("allele\_freq"), hdi=hdi, compound\_mean=hdi)

```

The HLA-B\*27 allele is strongly associated with ankylosing spondyloarthritis, 90% of all patients have this version of the gene. Now we have the national allele frequencies we can pair them with the national disease rates to study the correlation. After log transformation, the disease rate was normally distributed so I could fit an ordinary least squares linear regression. As expected, there was a significant positive correlation. Finland had unusually high frequency of HLA-B\*27 and was identified as an outlier based on statistical leverage, so it was removed from the regression model.

We can use our linear regression model to predict rates of ankylosing spondyloarthritis in countries where we know the HLA-B\*27 frequency. This greatly increases the number of countries with disease rate estimates. However, we should be cautious about predicting disease rates for countries with HLA-B\*27 rates outside of the range of our model. This is particularly true of high HLA-B\*27 rates as Finland (with it’s high HLA-\*27 rate) 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 ommited from the model is plotted in red. The predicted disease rates are plotted as open circles, cyan for countries in the models 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. Resources like allelefrequency.net and HLAfreq 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. Here I have investigated the relationship between HLA allele frequency and ankylosing spondyloarthritis rates but any other variable of interest such as severity of infection