Determinants of Genetic Drift Rate and their Relevance to the *Out—of—Africa* Hypothesis

Data Analysis Project Sample

Known Author

Introduction

Context of the project

The major histocompatibility complex (MHC) in humans plays a key role in triggering an immune response (i.e. would lead to the destruction of pathogen-virus-infected cells through presentation of antigenic peptides to T-cells; Cambier, Littman, & Weiss, 2001). This complex is located on chromosome 6 in humans, and the genes that compose it are commonly referred to as human leucocyte antigen (HLA). Among these genes, some (HLA-A, -B, -C for class I, and HLA-DPB1, -DRB1, -DQB1 for class II) are known to be particularly polymorphic, and taking into consideration the HLA variation existing in a given human population thus results in recognising a vast set of antigenic peptides.

It is now well accepted that different kind of mechanisms (i.e. stochastic factors related to geographical and demographic expansion of modern humans, and natural selection) influence the evolution of HLA polymorphism (see Sanchez-Mazas, Lemaitre, & Currat, 2012). In fact, geographical factors have been found to influence genetic diversity in humans: in particular, heterozygosity has been shown to decrease as a function of increasing distance from East Africa, thus supporting the "Out-of-Africa" theory of modern humans. Another common theory known as the "pathogen-driven-balancing selection" model posits that heterozygosity on HLA loci allows for higher fitness in pathogen-rich environments. However, the same evolution mechanisms might not be at play for the different classes of HLA molecules since they have different roles in immunity.

In this analysis project, the main focus was on exploring the relation between demographic history of populations (more specifically, their rate of genetic drift) and other variables for loci HLA-DPB1 and HLA-DQB1. These two specific genes are known to bind to other MHC class II genes (namely HLA-DPA1 and HLA-DQA1) to form functional protein complexes critical to elicit an immune response from the body.

The purpose of exploring these relations was to observe to which extent the investigated models supported one of the two major hypotheses of the evolution of HLA polymorphism presented above. More specifically, the "Out-of-Africa" hypothesis would predict that slower genetic drift is associated with smaller distance from Addis-Abeda; while rapid genetic drift would be associated with less heterozygosity, which would relate to less fitness (operationalised as allelic richness) in pathogen-rich envionments according to the "pathogen-driven-balancing selection" model.

Data description

For each population in the data set, the following information is available:

- **Locus**: HLA gene for which the population was sampled
- **Population**: Common name of the sampled population
- **Country**: Country where lives the sampled population

- **Region** (added after pre-processing): Geographical region (EUR: Europe, NAF: North Africa, SSA: Sub-Saharan Africa, WAS: West Asia, SAS: South Asia, CAS: Central Asia, NEA: Northeast Asia, NAM: North America, SAM: South America, SEA: Southeast Asia, OCE: Oceania, AUS: Australia)
- **Demography**: Assumed demographic history of the population, through either rapid genetic drift (RGD), for small-sized and isolated ones, or slow genetic drift (SGD) for the others (large outbred populations)
- **Sample_size**: Numbers of individuals collected to make inference about the population
- **H_expected**: Expected heterozygosity (H) within a sampled population at Hardy-Weinberg equilibrium
- **Num_Diff_Alleles**: Number of different alleles observed in the population
- **Allele_Richness**: Number of different alleles observed in the population with the sample size taken into account
- **Distance_Addis.Abeda**: Distance (in km) for the population from East Africa (taking Addis-Abeda, 9.03 N, 38.74 E as the reference) across landmass, assuming that human populations did not cross large bodies of water during their migration history
- **Pathogen_richness**: Information on pathogen richness of the country, extracted from the GIDEON database which provides information on the presence and the prevalence of infectious diseases by country
- **Virus_pathogens**: Numbers of viruses considered as pathogens
- **Non-Virus_pathogens**: Numbers of non-viruses considered as pathogens

```
str(HLA)
```

```
'data.frame':   481 obs. of  12 variables:
 $ Locus              : Factor w/ 6 levels "A","B","C","DPB1",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Population         : Factor w/ 158 levels "Aka Pygmies",..: 87 32 157 27 27 51 70 70 70 38 ...
 $ Country            : Factor w/ 77 levels "Algeria","Australia",..: 45 45 60 9 9 26 36 36 36 41 ...
 $ Demography         : Factor w/ 2 levels "RGD","SGD": 2 2 2 2 2 2 2 2 2 2 ...
 $ Sample_size        : int  144 134 372 124 124 130 286 530 482 276 ...
 $ H_expected         : num  0.913 0.922 0.936 0.921 0.929 ...
 $ Num_Diff_Alleles   : int  21 27 28 22 27 20 39 30 28 21 ...
 $ Allelic_Richness   : num  16.6 19.3 19 17.5 19.5 ...
 $ Distance_Addi.Abeda: num  5218 5416 4416 6881 6775 ...
 $ Pathogen_richness  : int  203 203 225 195 195 206 232 232 232 218 ...
 $ Virus_pathogens    : int  40 40 45 38 38 39 46 46 46 42 ...
 $ Non.Virus_pathogens: int  156 156 173 150 150 160 178 178 178 168 ...
```

Table 1: Summary of HLA data (unprocessed)

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Sample_size	481	219.422	223.264	3	102	218	2,000
H_expected	481	0.822	0.133	0.001	0.782	0.908	1.021
Num_Diff_Alleles	481	18.744	12.820	3	10	24	143
Allelic_Richness	481	13.065	5.608	1.997	8.817	16.823	31.585
Distance_Addi.Abeda	481	10,887.770	7,499.213	164.490	5,175.900	13,251.570	31,235.340
Pathogen_richness	481	218.767	16.901	183	205	233	250
Virus_pathogens	481	42.975	4.339	34	40	46	51
Non.Virus_pathogens	481	168.530	12.774	141	159	182	193

Data pre-processing

Populations with problematic data were excluded (typically, expected heterozygosity smaller than 0 or greater than 1, sample size smaller than 50). A subset of the data was then extracted containing only the two loci of interest (HLA-DPB1 and HLA-DQB1). Populations were subsequently grouped by geographical region (see Figure 1) to synthesize information. Subsequent exploratory analysis of the distribution of expected heterozygosity and pathogen richness investigated the genetic validity of that categorisation.

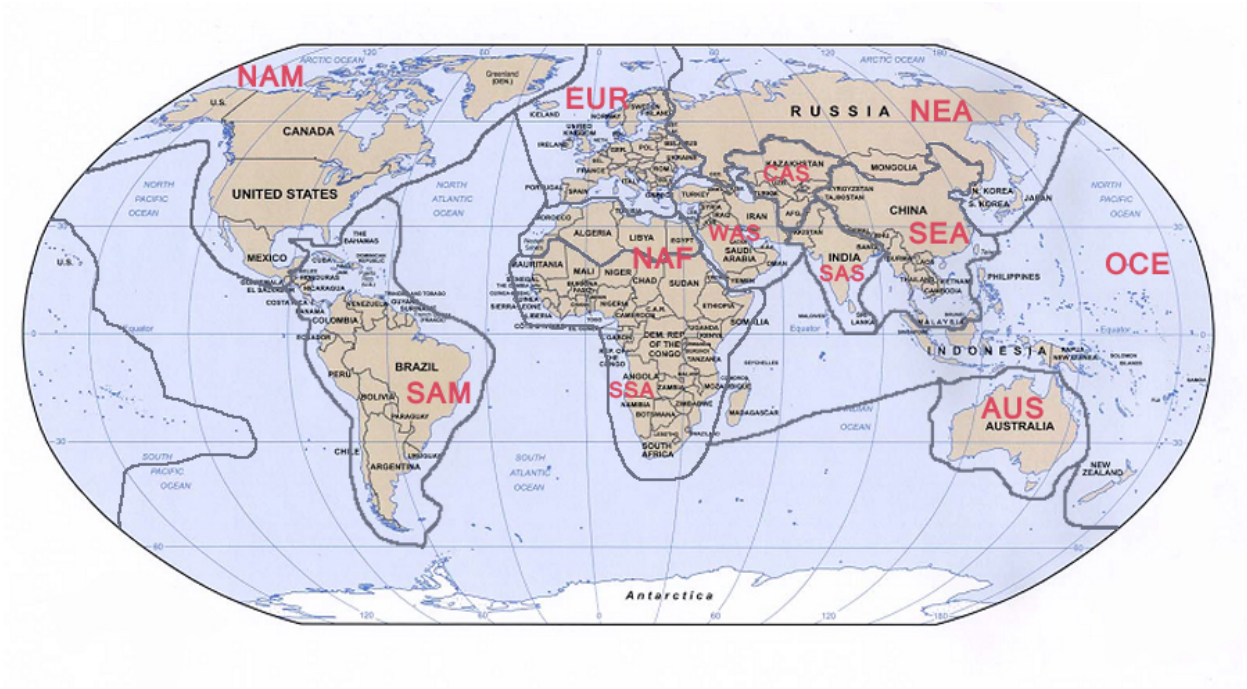
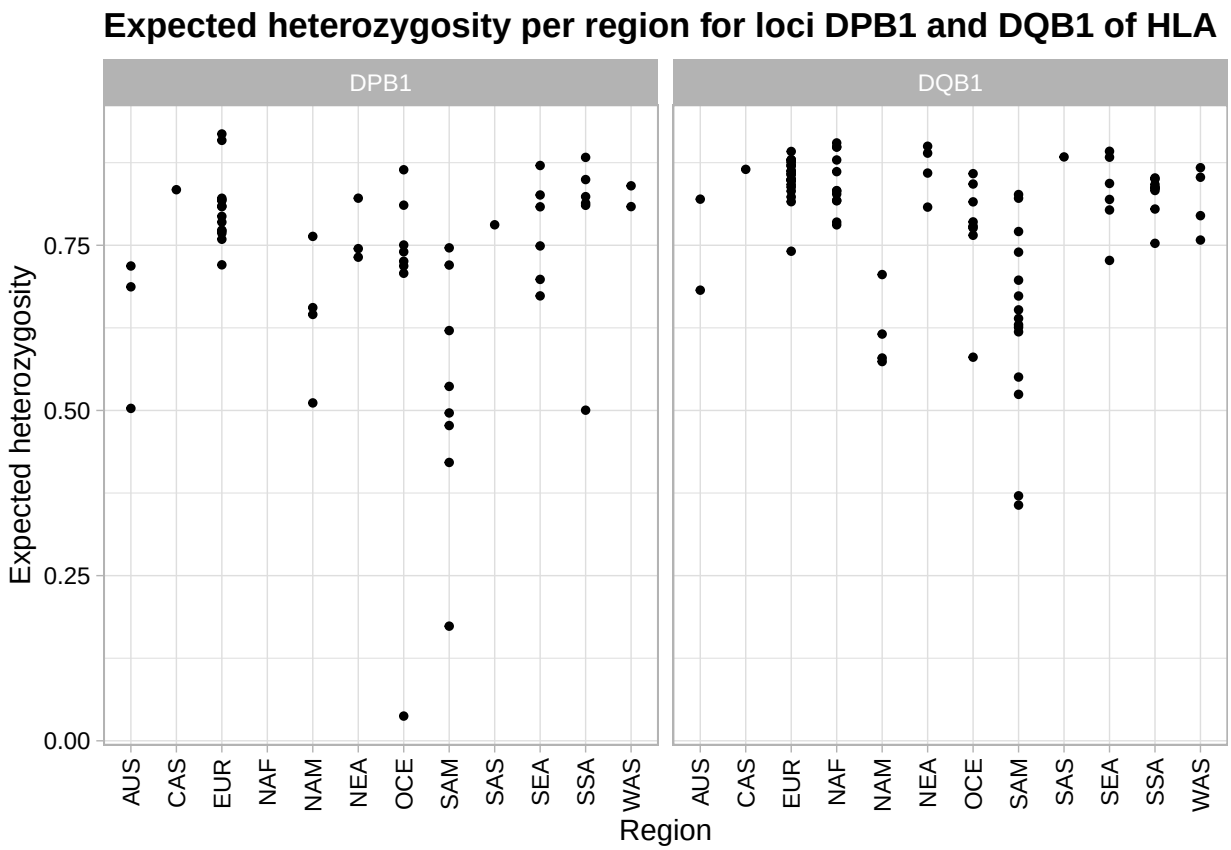


Figure 1: Categorisation map for samples in the HLA data set.

```
str(HLA_sub)
```

```
'data.frame':   144 obs. of  13 variables:
 $ Locus               : Factor w/ 6 levels "A","B","C","DPB1",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ Population          : Factor w/ 158 levels "Aka Pygmies",..: 157 70 86 118 33 80 126 92 156 88 ...
 $ Country             : Factor w/ 77 levels "Algeria","Australia",..: 60 36 39 10 16 57 77 43 43 43 ...
 $ Region              : Factor w/ 12 levels "AUS","CAS","EUR",..: 11 11 7 11 11 11 11 8 8 8 ...
 $ Demography          : Factor w/ 2 levels "RGD","SGD": 2 2 2 1 2 2 2 1 1 1 ...
 $ Sample_Size         : int  174 244 326 162 168 186 456 206 182 108 ...
 $ H_Expected          : num  0.824 0.883 0.864 0.5 0.849 ...
 $ Num_Diff_Alleles    : int  14 34 15 11 16 16 20 10 12 5 ...
 $ Allelic_Richness    : num  10.82 18.27 11.7 8.62 12.38 ...
 $ Distance_Addis.Abeda: num  4416 1168 3182 2408 2995 ...
 $ Pathogen_Richness   : int  225 232 213 227 230 227 217 233 233 233 ...
 $ Virus_Pathogens     : int  45 46 45 50 49 49 44 43 43 43 ...
 $ Non.Virus_Pathogens : int  173 178 160 170 173 170 165 182 182 182 ...
```

```
ggplot(data=HLA_sub, aes(x = Region, y = H_Expected)) +
  geom_point(size = 1) +
  facet_grid(~Locus) +
  labs(title = "Expected heterozygosity per region for loci DPB1 and DQB1 of HLA",
       x = "Region", y = "Expected heterozygosity") +
  theme_light() +
  theme(plot.title = element_text(face = "bold"), axis.text =
          element_text(colour = "black"), axis.text.x =
          element_text(angle = 90, vjust = 0.5, hjust = 1))
```

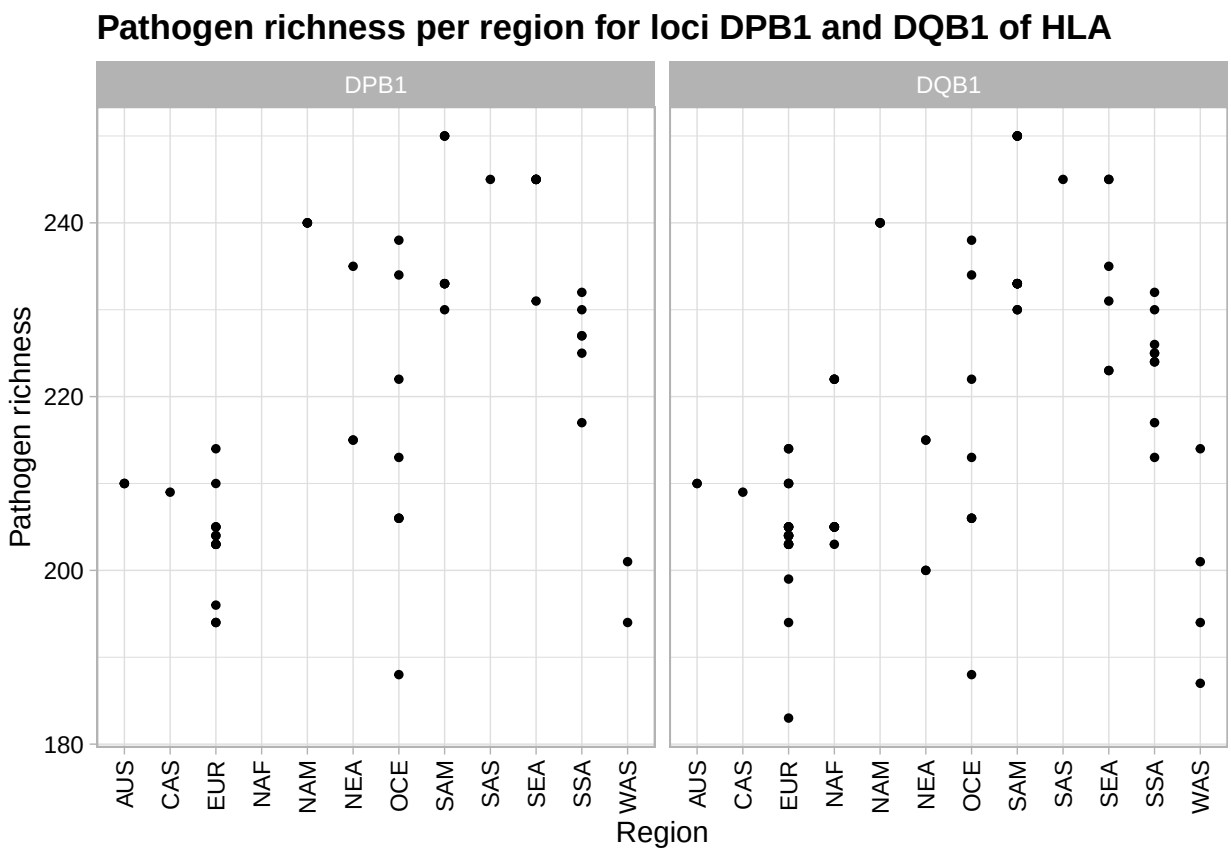


The distribution of expected heterozygosity per region seems to display more variability for locus DPB1 than for DQB1. When looking at the data per region, South America seems to display much more variability than the other regions with a greater range of possible expected heterozygosity. Furthermore, the smallest expected heterozygosity is found in an Oceanian sample (Trobriand Islanders) for locus DPB1, with a value (0.037) that is clearly different from the general tendency in that region. In fact, this population has remained secluded for a very long time, and has underwent rapid genetic drift, which explains such low expected heterozygosity.

Table 2: Smallest expected heterozygosity sample

	Locus	Population	Country	Region	Demography	H_Expected
287	DPB1	Trobriand Islanders	Papua New Guinea	OCE	RGD	0.037

```r
ggplot(data=HLA_sub, aes(x = Region, y = Pathogen_Richness)) +
  geom_point(size = 1) +
  facet_grid(~Locus) +
  labs(title = "Pathogen richness per region for loci DPB1 and DQB1 of HLA",
       x = "Region", y = "Pathogen richness") +
  theme_light() +
  theme(plot.title = element_text(face = "bold"), axis.text =
          element_text(colour = "black"), axis.text.x =
          element_text(angle = 90, vjust = 0.5, hjust = 1))
```

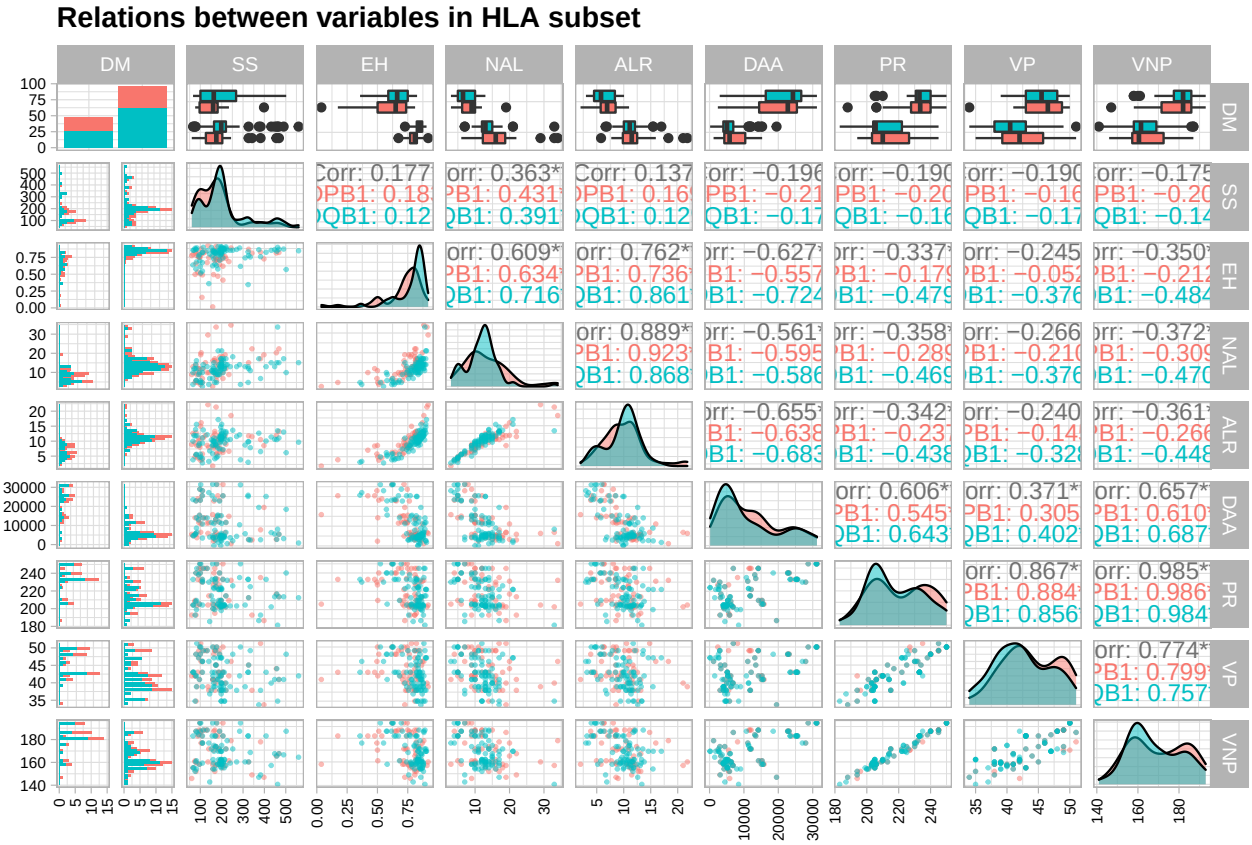


The distribution of pathogen richness per region seems to be approximately similar for both loci of interest, with the greatest values found in South America (for DPB1 and DQB1), and the smallest one found in Europe (for DQB1). Moreover, variability seems to be the greatest in Oceania with a greater range of possible values. Pathogen richness was found to be especially high in Brazilian populations (250, the maximum of the index) which is coherent with the persistent challenge of infectious diseases with new one emerging where old ones were beginning to be controlled and effectively treated (Waldman & Sato, 2016). On the other hand, the small pathogen richness observed in the Azores population might be related to its small size and short history: in fact, the first human inhabitants of the Azores have come in the 15th century.

Some general observations can also be made: both pathogen richness and expected heterozygosity seem to behave similarly for the two different loci of human leucocyte antigen with approximately the same distribution per region. However, some regions that are represented for one locus of HLA have no available data for the other locus, such as North Africa which is missing for DPB1, and Northeast Asia which is missing for DQB1. It is also worthy of note that the categorisation per region does not seem to be very informative for some specific regions (e.g. South America or Oceania) since data are very variable (maybe due to the greater geographical distance between populations within certain regions).

Table 3: Greatest and smallest pathogen richness samples

	Locus	Population	Country	Region	Demography	Pathogen_Richness
248	DPB1	Guarani	Brazil	SAM	RGD	250
249	DPB1	Kaingang	Brazil	SAM	RGD	250
250	DPB1	Ticuna	Brazil	SAM	RGD	250
323	DQB1	Guarani-Kaiowa	Brazil	SAM	RGD	250
324	DQB1	Guarani-Nandewa	Brazil	SAM	RGD	250
325	DQB1	Guarani	Brazil	SAM	RGD	250
326	DQB1	Kaingang	Brazil	SAM	RGD	250
327	DQB1	Ticuna	Brazil	SAM	RGD	250
348	DQB1	Azoreans	Azores	EUR	SGD	183



Legend: Blue: Locus DPB1, Red: Locus DQB1
Variable labels: DM: Demography; SS: Sample size; HE: Expected heterozygosity; NAL: Number of different alleles; ALR: Allelic richness; DAA: Distance to Addis−Abeda; PR: Pathogen Richness; VP: Virus pathogen; VNP: Virus non −pathogen

When looking at the relations between all other variables for loci DPB1 and DQB1, there seems to be no major difference between both loci regarding how the variables are distributed, and how they relate to each other. In most of the diagrams, the graphical representations overlap.

Overall, rapid-genetic-drift samples (upper row, upper boxes) display more variability in their size, have smaller allelic richness, fewer different alleles, a greater distance from Addis-Abeda, and more pathogen richness than slow-genetic-drift samples (upper row, lower boxes).

Some obvious links are observed between number of different alleles and allelic richness, and between pathogen

richness and virus pathogen/non-pathogen, since, for both cases, the second variables are functions of the first. In the case of allelic richness, given that this variable is a corrected version of the number of different alleles, it shall be retained for the rest of the analyses.

More worthy of note is the non-linear relation between expected heterozygosity and allelic richness ($r = .76$), which is sensible given that heterozygosity is defined as difference of alleles for a given locus.

Data Analysis

Logistic Regression

According to our initial hypotheses and on our exploratory analysis, generalised linear models were computed to test if HLA Locus, distance to Addis-Abeda, expected heterozygosity, allelic richness, and pathogen richness could account for variability in the rate of genetic drift (**Demography** variable). The results are displayed in Table 4. The first two tested models contained (1) all the variables mentioned above plus the interaction of allelic richness with pathogen richness, (2) the same model without that interaction. New models were then iteratively computed using only the significant effects of the greater model they are nested in. Even by doing this, the Akaike Information Criterion kept mildly increasing for smaller models (apart from model 1 to model 2).

Table 4: Comparison of generalised linear models

	Dependent variable:			
	Demography			
	(1)	(2)	(3)	(4)
Locus DQB1	−1.477	−1.590		
	(1.060)	(1.040)		
Distance to Addis-Abeda	−0.0003***	−0.0003***	−0.0003***	−0.0003***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Expected heterozygosity	18.000*	19.257**	21.886***	21.545***
	(9.309)	(9.312)	(6.266)	(5.908)
Allelic richness	−2.184	0.546		
	(3.547)	(0.343)		
Pathogen richness	−0.179	−0.058**	−0.035	
	(0.163)	(0.028)	(0.023)	
Allelic richness:Pathogen richness	0.013			
	(0.016)			
Constant	24.767	−2.299	−4.876	−12.210***
	(36.362)	(7.681)	(6.394)	(4.570)
Observations	144	144	144	144
Log Likelihood	−20.229	−20.534	−23.600	−24.832
Akaike Inf. Crit.	54.457	53.067	55.201	55.664
Note:				*p<0.1; **p<0.05; ***p<0.01

A formal ANOVA test was conducted to compare the four models (see Table 5). Model 1 and 2 did not significantly differ, which is coherent with the fact that the pathogen richness by allelic richness interaction effect was found to be not significant. Similarily, Model 3 and 4 did not differ. However, a significant reduction in residual deviance was found when comparing Model 2 to Model 3. For that reason, and for the fact that this model displayed the smallest Akaike Information Criterium, the greater model without the interaction of allelic richness with pathogen richness (Model 2) was retained. Diagnostic plots were produced to validate it

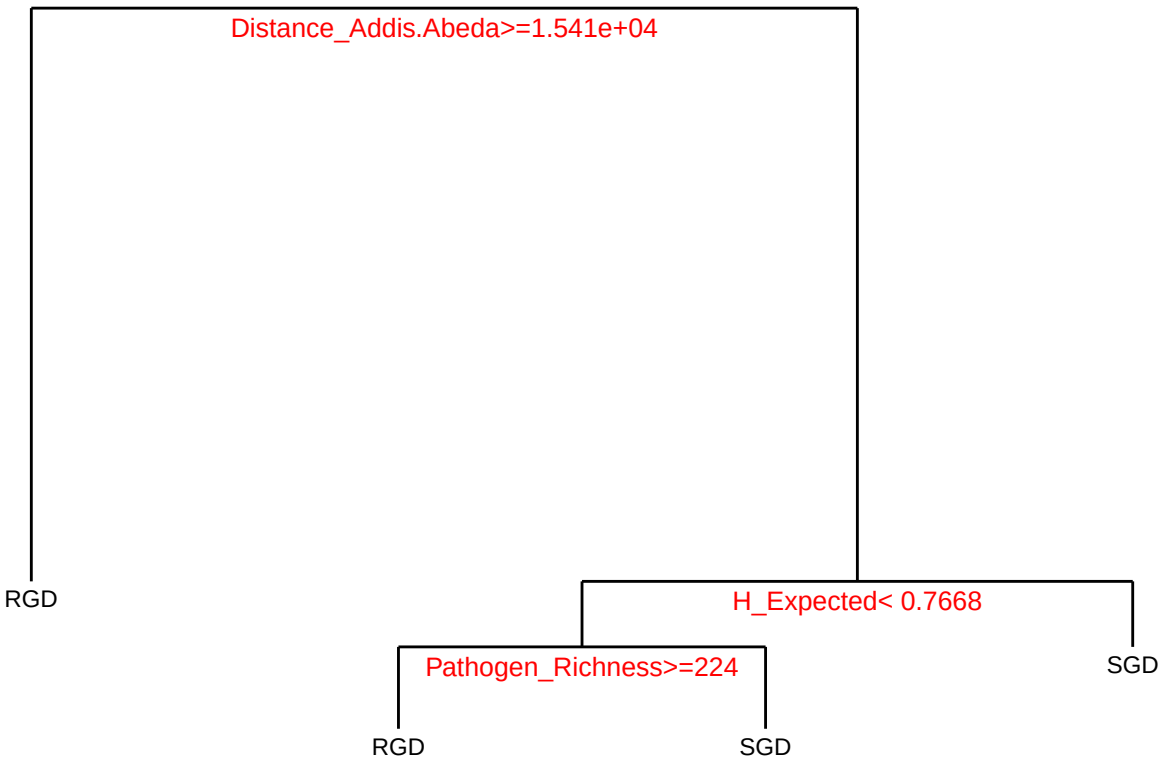
(see Appendix 1).

Table 5: Formal comparison of regression models through ANOVA

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	137	40.457			
2	138	41.067	-1	-0.610	0.435
3	140	47.201	-2	-6.134	0.047
4	141	49.664	-1	-2.463	0.117

Partitioning Method

In addition to a regression modeling approach, the data were partioned using the regression model selected in the previous part with the aim to investigate how well could HLA Locus, distance to Addis-Abeda, expected heterozygosity, allelic richness, and pathogen richness differentiate fast from slow genetic drift. The results displayed here are consistent with the rest of our analyses: the three significant main predictors (Distance to Addis-Abeda, Expected heterozygosity, and Pathogen richness) are the most useful in partioning the data.



Conclusion

The present data analysis was aimed at exploring the relation between rate of genetic drift and other genetic characteristics of populations with the more general objective to better understand the evolution mechanisms of human leucocyte antigen. The data analysed in fact supported one specific model: it was observed that the probability that a sample underwent a slow genetic drift decreased as a function of distance to Addis-Abeda (Appendix 2, Plot 1), supporting the "Out-of-Africa" model. Even though this might be true, some specific samples (such as the Pygmies Aka, discussed in Appendix 2) did not follow this relationship. In fact, genetic

drift rate is not strictly under dependence of geographical effects: populations close to the "origin point" might have rapidly been re-sampled and secluded in their history, having them undergo a fast genetic drift. However, geographical information seems to provide an excellent partioning point to differentiate the rate of genetic drift of a given population.

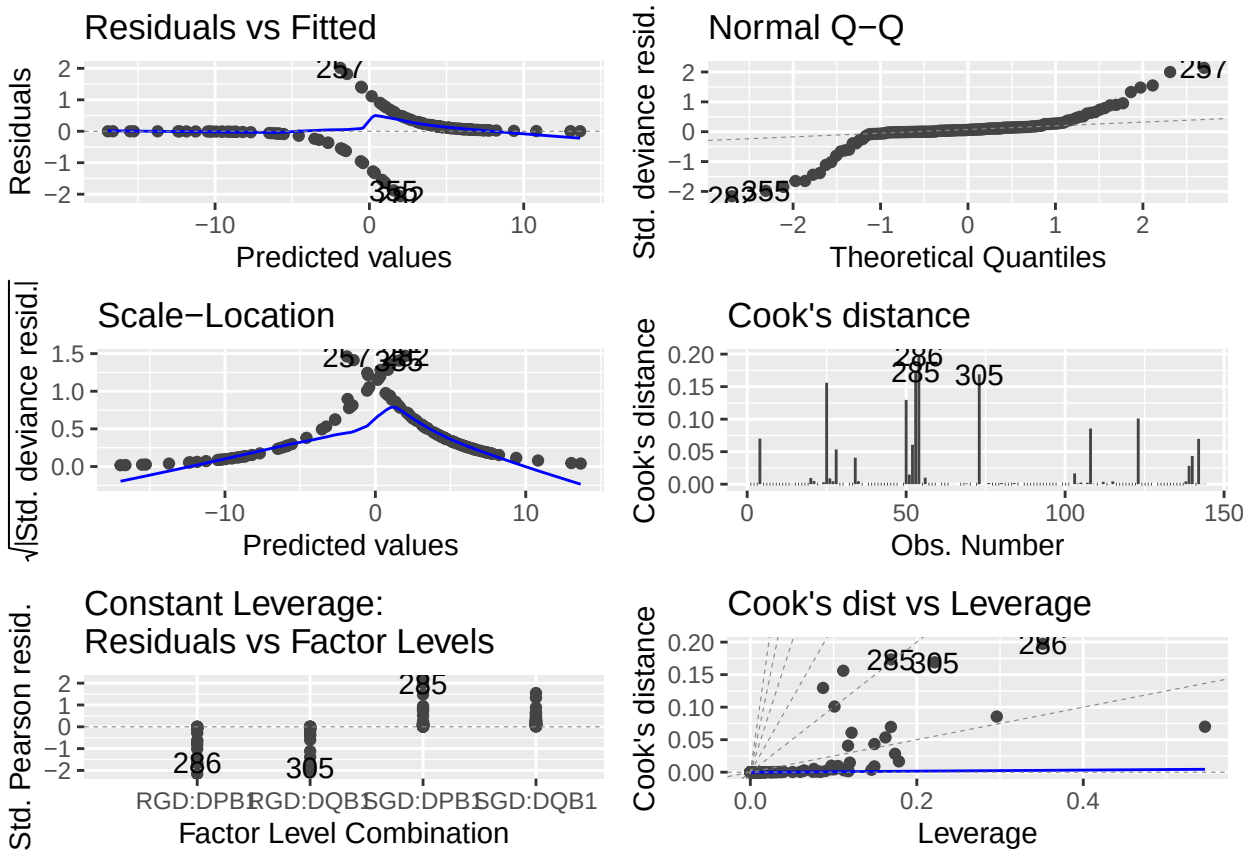
The present project also sought to ascertain the relevance of the "pathogen-driven-balancing selection" model to understand evolution mechanisms of HLA genes. It was also observed that the probability of slow genetic drift increased as a function of expected heterozygosity (Appendix 2, Plot 2), allelic richness (Appendix 2, Plot 3), but decreased as a function of pathogen richness (Appendix 2, Plot 4). Allelic richness and pathogen richness were not found to interact, which would have been expected given the hypothesis that higher allelic richness would be associated with slower genetic drift only in rich pathogen environments. However, the partioning method yielded intersting information: slow genetic drift did happen more frequently for low expected heterozygosity - low pathogen richness samples. The lack of information about allelic richness (operalisation of fitness) does not permit to decide if these observations are relevant to the "pathogen-driven-balancing selection" model tested here. Overall, the data reported here could not support this model of the evolution mechanisms of HLA genes.

In summary, it seems that genes HLA-DPB1 and HLA-DQB1 of the major histocompatibility complex in humans follow a stochastic evolution process under the main influence of geographical migration, supporting the "Out-of-Africa" model of the evolution of anatomically modern humans.

Appendices

Appendix 1: Diagnostic plots

```
autoplot(glm1.2, 1:6)
```

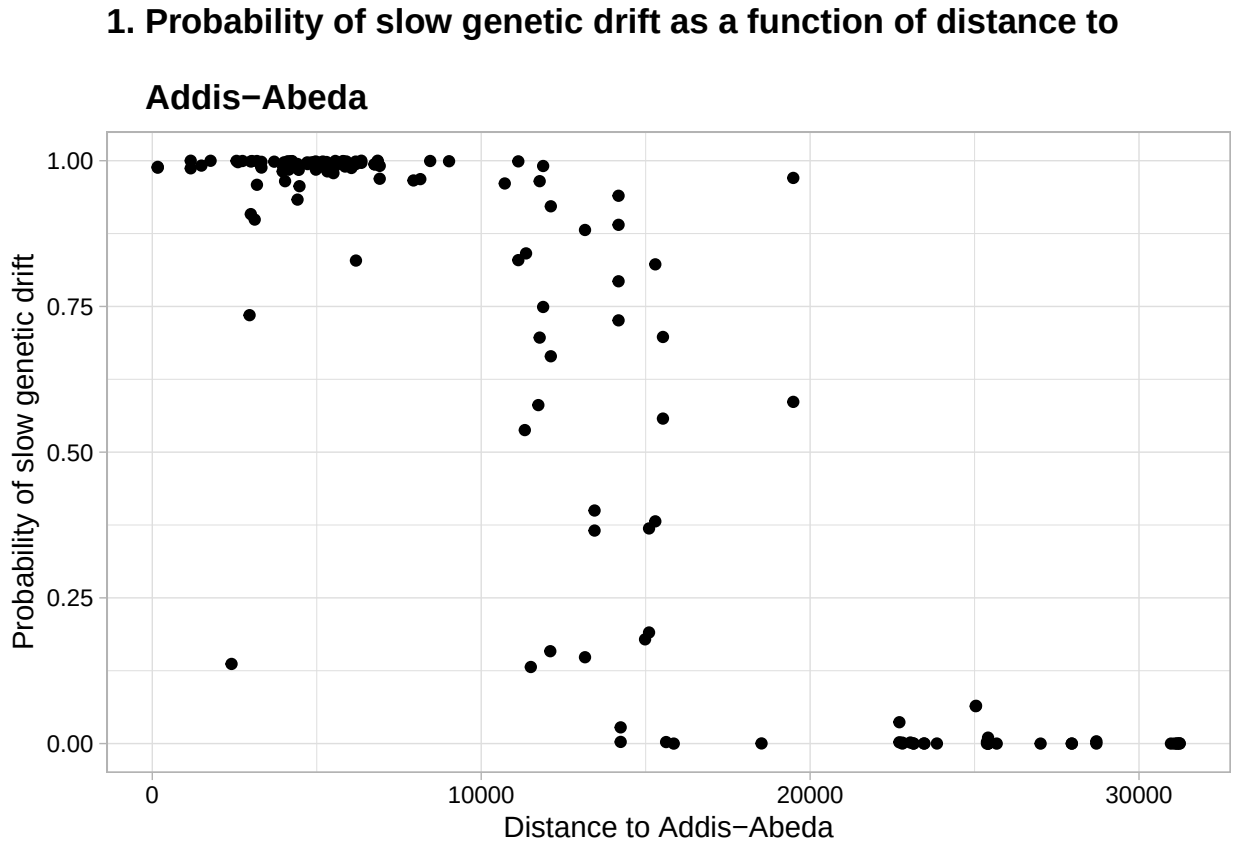


According to the diagnostic plots, there seems to be no extreme values (Cook's distances are well below 1). Note that it is not possible to check for homoscedasticity since we do not know the expected distribution of residuals.

Appendix 2: Fitted probability of rapid genetic drift

Note that the following representations only express the link between pairs of variables and do not account for the full model explored in the present data analysis project. However, they remain a useful tool to visually represent how the different variables explored influence the probability that a population underwent a rapid or slow genetic drift. See comments under plots 1 and 4. Note that the model curve is not displayed in the following graphs since they hinder the readability of the plots in some cases (e.g. negative probabilities).

```
ggplot(data = HLA_sub,
       aes(x = Distance_Addis.Abeda, y = fitted(glm1.2))) +
  geom_point() +
  labs(title = "1. Probability of slow genetic drift as a function of distance to \n
    Addis-Abeda",
       x = "Distance to Addis-Abeda", y = "Probability of slow genetic drift") +
  theme_light() +
  theme(plot.title = element_text(face = "bold"), axis.text =
          element_text(colour = "black"))
```



A noticeable feature of the graphical representation of the probability of slow genetic drift as a function of distance to Addis-Abeda is the presence of values that do not follow the logistic s-shaped curve of the regression model, which is less the fact for the two following graphical representations: in fact, at least one data point clearly falls outside the model's shape with a distance to Addis-Abeda of approximately 2'500 kilometers and a SGD probability of .13.

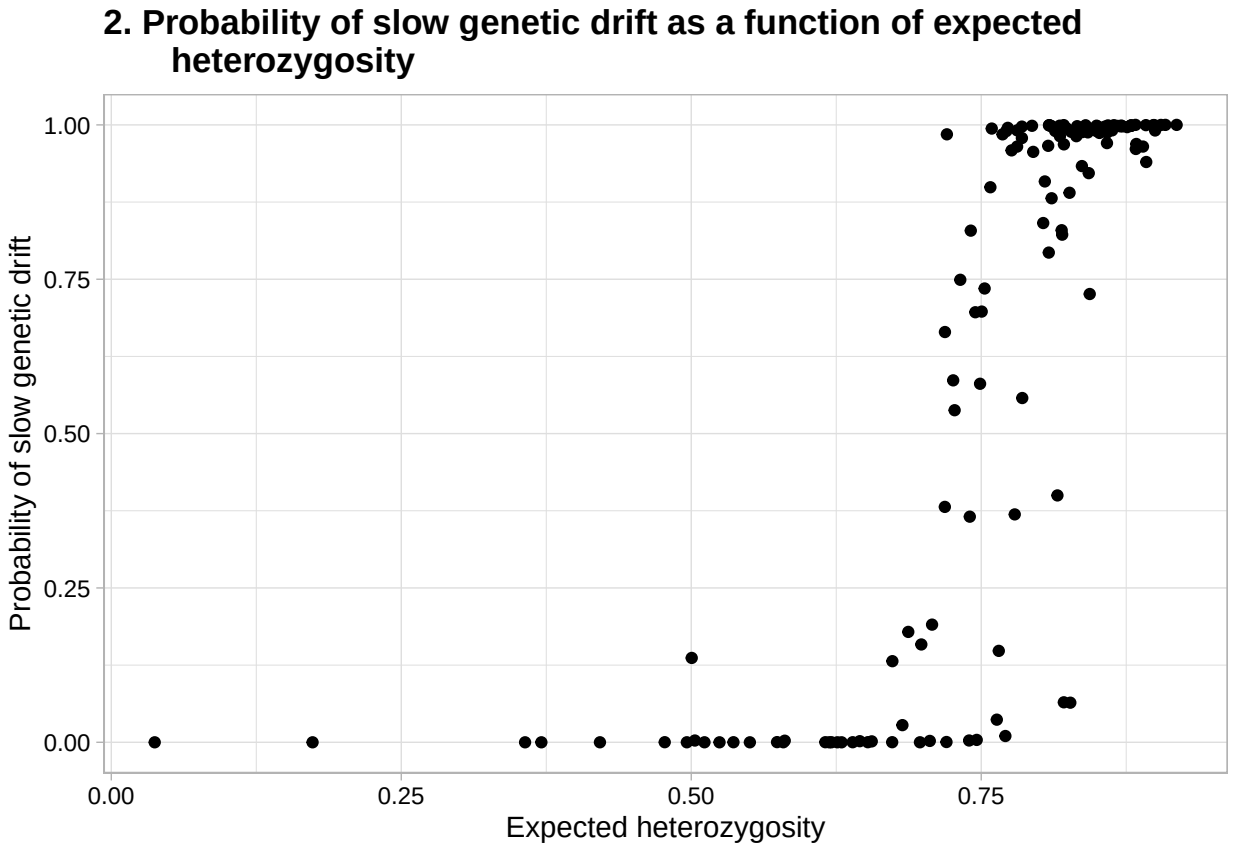
The sample in question was extracted from the Pygmy Aka population in Central African Republic (Sub-Saharian Africa) which underwent a remarkably rapid genetic drift despite its proximity to Addis-Abeda. Indeed, the particularity of its phenotype (Pygmies are relatively small in height compared to other populations)

Table 6: Outlier data point for selected regression model

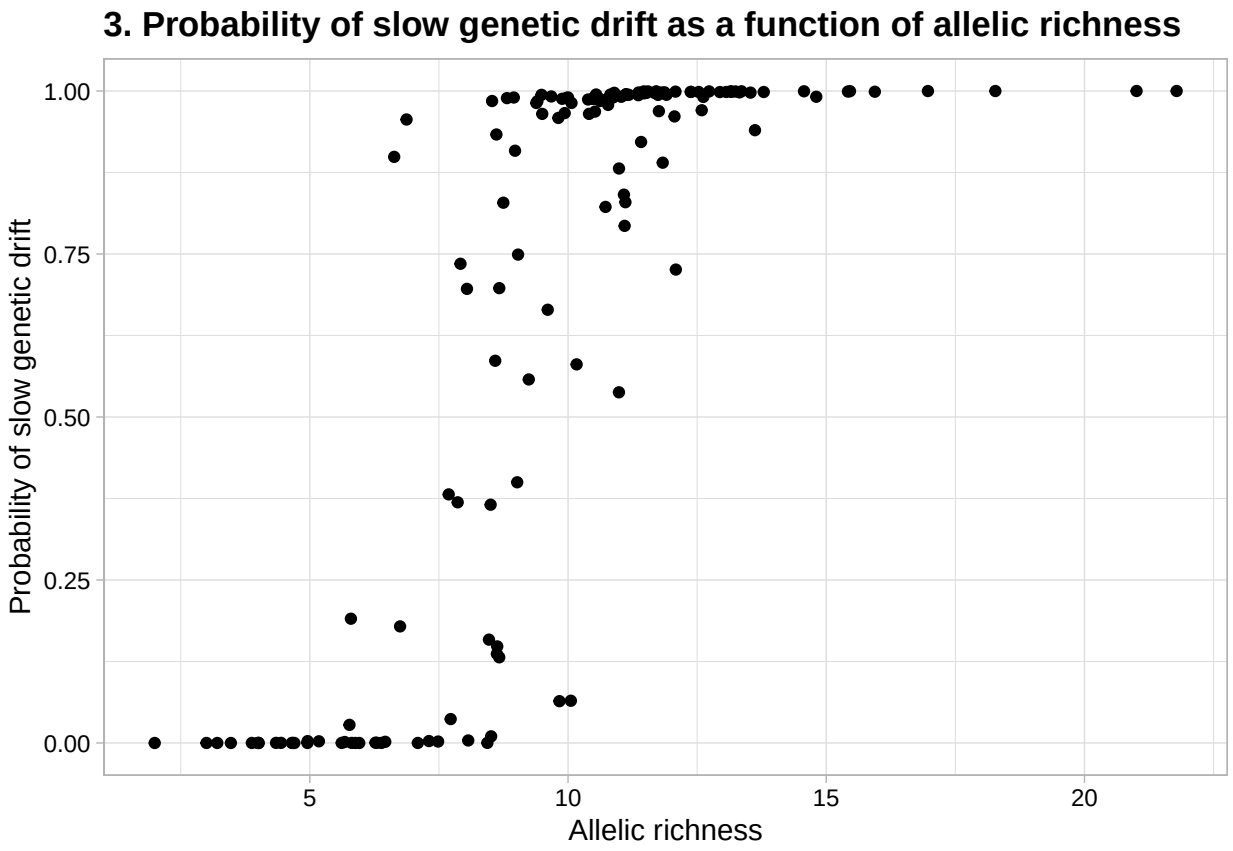
	Locus	Population	Country	Distance_Addis.Abeda	Fitted
236	DPB1	Pygmy Aka	Central African Republic	2,408.500	0.137

may be a consequence and proof of the fast genetic drift they underwent (with a possible "Founder effect" that selected a very specific phenotype that quickly differentiated from other populations). In fact, the logistic regression model did not fail to categorise this sample as having a low probability of slow genetic drift.

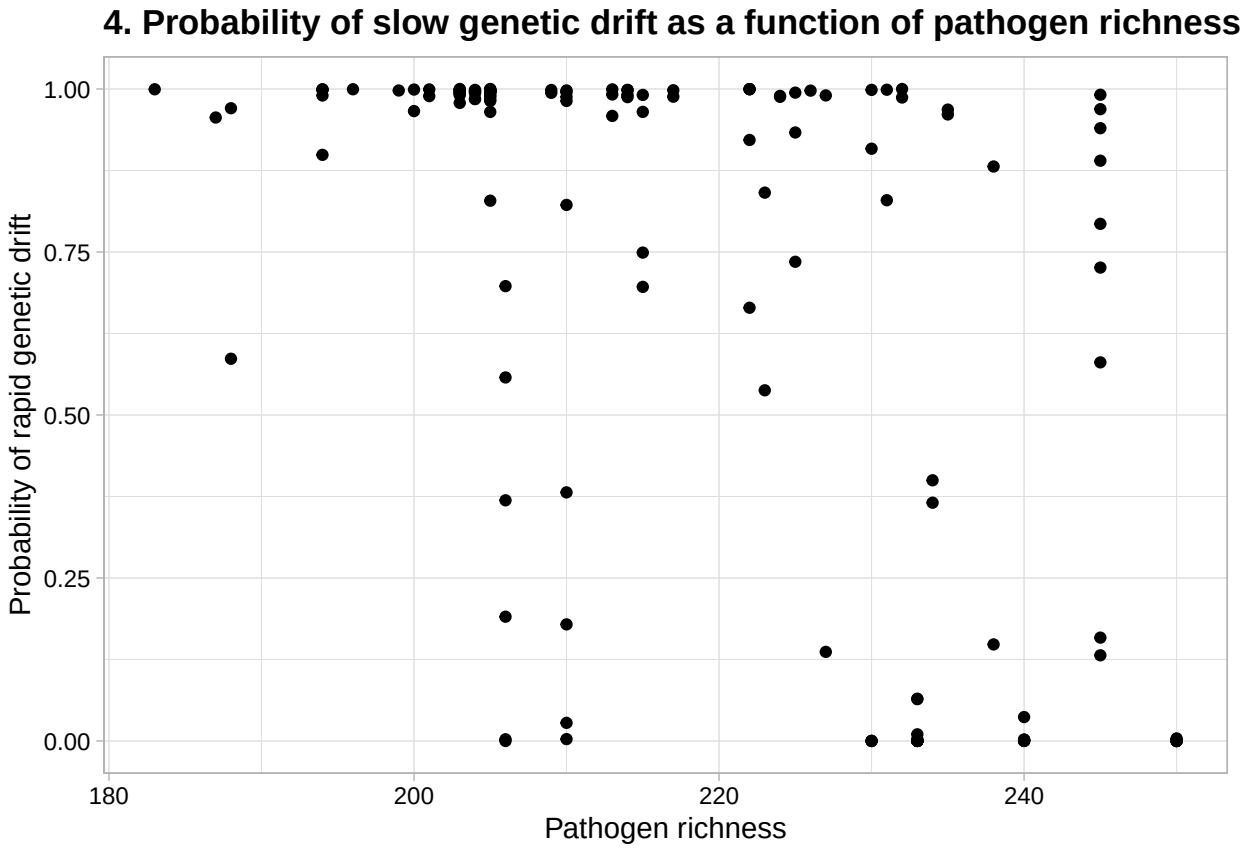
```
ggplot(data = HLA_sub,
       aes(x = H_Expected, y = fitted(glm1.2))) +
  geom_point() +
  labs(title = "2. Probability of slow genetic drift as a function of expected
       heterozygosity",
       x = "Expected heterozygosity", y = "Probability of slow genetic drift") +
  theme_light() +
  theme(plot.title = element_text(face = "bold"), axis.text =
          element_text(colour = "black"))
```



```r
ggplot(data = HLA_sub,
       aes(x = Allelic_Richness, y = fitted(glm1.2))) +
  geom_point() +
  labs(title = "3. Probability of slow genetic drift as a function of allelic richness",
       x = "Allelic richness", y = "Probability of slow genetic drift") +
  theme_light() +
  theme(plot.title = element_text(face = "bold"), axis.text =
          element_text(colour = "black"))
```



```r
ggplot(data = HLA_sub,
       aes(x = Pathogen_Richness, y = fitted(glm1.2))) +
  geom_point() +
  labs(title = "4. Probability of slow genetic drift as a function of pathogen richness",
       x = "Pathogen richness", y = "Probability of rapid genetic drift") +
  theme_light() +
  theme(plot.title = element_text(face = "bold"), axis.text =
          element_text(colour = "black"))
```



The lack of information provided by pathogen richness in discriminating between slow and genetic drift is not very surprising since this predictor was only found to be significant in the greater model without interaction. However, it is puzzling that the partioning analysis chose this variable over allelic richness which offers a much better discrimination between slow and rapid genetic drift.

References

- Cambier, J. C., Littman, D. R., Weiss, A. (2001). Antigen presentation to T lymphocytes. In C. A. Janeway, P. Travers, M. Walport & M. J. Shlomchik (Eds.), *Immunobiology: The Immune System in Health and Disease* (pp.186-218). New York: Garland Publishing.
- Genetics Home References (2018, January 18). Genes [Section name]. Retrieved from https://ghr.nlm.nih.gov/gene.
- Sanchez-Mazas, A., Lemaitre, J.-F., & Currat, M. (2012). Distinct evolutionary strategies of human leucocyte antigen loci in pathogen-rich environments. *Philosophical Transactions of the Royal Society*, *367*, 830-839.
- Waldman, E. A., & Sato, A. P. S. (2016). Path of infectious diseases in Brazil in the last 50 years: An ongoing challenge. *Revista de Saúde Pública*, 50-68.

Session Information

```
R version 4.0.4 (2021-02-15)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 21.04

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.9.0
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.9.0

locale:
 [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_DK.UTF-8        LC_COLLATE=en_GB.UTF-8
 [5] LC_MONETARY=en_DK.UTF-8    LC_MESSAGES=en_GB.UTF-8
 [7] LC_PAPER=en_DK.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_DK.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] ggdendro_0.1.22  rpart_4.1-15     ggfortify_0.4.12 stargazer_5.2.2
[5] GGally_2.1.0     ggplot2_3.3.3

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.6         highr_0.8          pillar_1.4.7       compiler_4.0.4
 [5] RColorBrewer_1.1-2 plyr_1.8.6         tools_4.0.4        digest_0.6.27
 [9] evaluate_0.14      lifecycle_0.2.0    tibble_3.0.6       gtable_0.3.0
[13] pkgconfig_2.0.3    rlang_0.4.10       DBI_1.1.1          yaml_2.2.1
[17] xfun_0.20          gridExtra_2.3      withr_2.4.1        dplyr_1.0.4
[21] stringr_1.4.0      knitr_1.31         generics_0.1.0     vctrs_0.3.6
[25] grid_4.0.4         tidyselect_1.1.0   reshape_0.8.8      glue_1.4.2
[29] R6_2.5.0           rmarkdown_2.6      farver_2.0.3       tidyr_1.1.2
[33] purrr_0.3.4        magrittr_2.0.1     MASS_7.3-53.1      scales_1.1.1
[37] ellipsis_0.3.1     htmltools_0.5.1.1  assertthat_0.2.1   colorspace_2.0-0
[41] labeling_0.4.2     stringi_1.5.3      munsell_0.5.0      crayon_1.4.0
```