snplinkage: Single Nucleotide Polymorphism Linkage Disequilibrium Visualizations

Thomas Charlon

Abstract

Single nucleotide polymorphisms are the most common genetic variations in humans and genome-wide microarrays measure up to several millions. Physically close polymorphisms often exhibit correlation structures and are said to be in linkage disequilibrium when they are more correlated than randomly expected. The **snplinkage** package provides linkage disequilibrium visualizations by displaying correlation matrices annotated with chromosomic positions and gene names. Two types of displays are provided to focus on small or large regions, and both can be extended to combine associations results or investigate feature selection methods. This article introduces the package and illustrates the variety of correlation structures found genome-wide by focusing on 3 regions associated with autoimmune diseases: the chromosome 1 region 1p13.2 (111-115 Mbp), the chromosome 8 region 8p23.1 (7-12 Mbp) and the chromosome 6 region 6p21.3 (29-35 Mbp).

Keywords: Genetic variability, microarray, SNP.

1. Introduction

Single nucleotide polymorphisms (SNPs) are the most common genetic variations in humans (Group *et al.* 2001) and several million have been identified in worldwide populations. Nucleotides are a set of 4 organic molecules (A, C, T, G) that bind by pairs and chromosomes are long chains of millions of nucleotide pairs. The human genome consists of 23 pairs of chromosomes totalling more than 3 billion nucleotide pairs and 99.9% of the sequences are identical between humans. Most variations are SNPs, in which a single nucleotide varies within a population and is surrounded by unique non-varying sequences.

Physically close SNPs exhibit correlation structures due to biological, hereditary and evolutionary factors. When they are more correlated than randomly expected they are said to be in linkage disequilibrium and groups of SNPs in strong linkage disequilibrium are called haplotypes. Genetic association studies often investigate correlation structures to explore specific regions associated with a trait or disease (Awadalla, Thapa, Burdon, Hewitt, and Craig 2011), and genome-wide association studies usually remove highly correlated SNPs to increase statistical power using TagSNP selection, a method that computes the correlations between SNPs among a sliding window, usually 500,000 base pairs (bp), and removes one of two SNPs that have a pair-wise correlation higher than a threshold, usually using the Pearson correlation $r^2 = 0.8$.

The **snplinkage** package provides linkage disequilibrium visualizations that combine the corre-

lation matrix of SNPs with their chromosomic positions and gene names in order to investigate correlation structures of specific regions. Two types of displays are available to focus on small or large regions, and both can be extended to combine association results or to investigate feature selection methods. The correlations are computed using the **SNPRelate** package and the plots are customizable **ggplot2** and **gtable** objects and are annotated using the **biomaRt** package.

This article introduces the package using a subset of the human genome diversity project dataset of 319 individuals. First, the variety of correlation structures found genome-wide are illustrated using 3 regions that include genes associated with autoimmune diseases: the 1p13.2 region (chromosome 1, 111-115Mbp), where small haplotypes of tens of SNPs are found, similarly as in most genomic regions; the 8p23.1 region (chromosome 8, 7-12Mbp) where a large well-defined haplotype of several hundreds of SNPs is found; and the *MHC* region (chromosome 6, 29-35Mbp), where several large but diffuse haplotypes are found, indicative of the complexity of weakly correlated signals in this region. Then, to demonstrate combining association results with linkage disequilibrium visualization, a simple geographical association study is performed on SNPs from the *MHC* region and significantly associated SNPs are visualized. Finally, TagSNP feature selection biplots are demonstrated.

2. Genotype data

The **snplinkage** includes a subset of the human genome diversity project dataset of 7,656 SNPs from 319 individuals: 157 Europeans and 162 Middle Easterners and North Africans. 5,000 genome-wide SNPs were regularly sampled from all chromosomes by position rank, and 2,656 SNPs were selected from 4 regions with genes associated with autoimmune diseases: the chromosome 1 region 1p13.2 from 113 to 115 Mbp, chromosome 6 region *MHC* between 29 and 33 Mbp, the chromosome 8 region 8p23.1 selected selected between 11 and 12 Mbp, and the chromsome 11 epitope between 49 and 56 Mbp. The data, originally 3 text files corresponding to the genotype, the samples annotations and the SNPs annotations, is first converted to the snpgds format (cf. **SNPRelate** package).

Quality control is performed to remove samples and SNPs with many missing values ($>$ 3% and $>$ 1%, respectively) and feature selection is performed by removing SNPs with low variance using minor allele frequency filtering (MAF, $<$ 5%) and physically close duplicated SNPs using TagSNP selection (500 kbp, $r^2 = 0.99$). The `snprelate_qc` function returns a list of two objects: the new subsetted Genotype Data object, slot `gdata`, and a data frame with quality control details of each filtering step, slot `df_info` (Table 1).

```
R>   library('snplinkage')
R>   gds_path <- save_hgdp_as_gds()
R>   gdata <- load_gds_as_genotype_data(gds_path)
R>   qc <- snprelate_qc(gdata, tagsnp = .99)
R>   print_qc_as_tex_table(qc)
```

Step	Parameter	Samples	SNPs
Raw	NA	319	7656
Samples NAs	0.03	318	7656
Identity by state - Twins	0.99	318	7656
SNPs NAs	0.01	318	7268
MAF	0.05	318	6639
TagSNP	0.99	318	6311

Table 1: Quality control and feature selection of the subset of the human genome diversity project dataset.

3. Linkage disequilibrium

3.1. Small haplotypes in the 1p13.2 region

The chromosome 1 region 1p13.2 illustrates the type of correlations and haplotypes most encountered genome-wide: small groups of 5-20 physically close SNPs more correlated with nearby SNPs than with others. The region includes the PTPN22 gene (Begovich, Carlton, Honigberg, Schrodi, Chokkalingam, Alexander, Ardlie, Huang, Smith, Spoerke *et al.* 2004) which is associated with rheumatoid arthritis and influences T and B cell receptors.

The `select_region_idxs` function is first called to select 20 SNPs based on their chromosome and position. The `gtable_ld_gdata` function is then called, using the Genotype Data object and the SNP indexes, and returns a gtable object.

The figure shows a set of 6 strongly correlated SNPs forming a well-defined haplotype, next to a larger set with weak and diffuse correlations (Figure 1).

```
R>   snp_idxs_1p13 <- select_region_idxs(qc$gdata,
+      chromosome = 1, position_min = 114.4e6, n_snps = 20, offset = 12)
R>   plt <- gtable_ld_gdata(qc$gdata, snp_idxs_1p13, labels_colname = 'probe_id')
R>   grid::grid.draw(plt)
```

Using a larger set of 100 SNPs enables to observe the patterns of weak and diffuse correlations. The `select_region_idxs` function is now called to select 100 SNPs from the 1p13.2 region. The `gtable_ld_gdata` function switches to a point-based visualization when the number of SNPs exceeds 40 or by setting the `diamonds` parameter to FALSE, and the point size can be adjusted using the `point_size` parameter.

The figure shows that even though most SNPs do not form well-defined haplotypes as the previous one (here on the right side), groups of 10-20 successive SNPs are weakly correlated and form diffuse haplotypes (Figure 2).

```
R>   snp_idxs_1p13_large <- select_region_idxs(qc$gdata, chromosome = 1,
+      position_min = 114e6, n_snps = 100)
R>   plt <- gtable_ld_gdata(qc$gdata, snp_idxs_1p13_large)
R>   grid::grid.draw(plt)
```

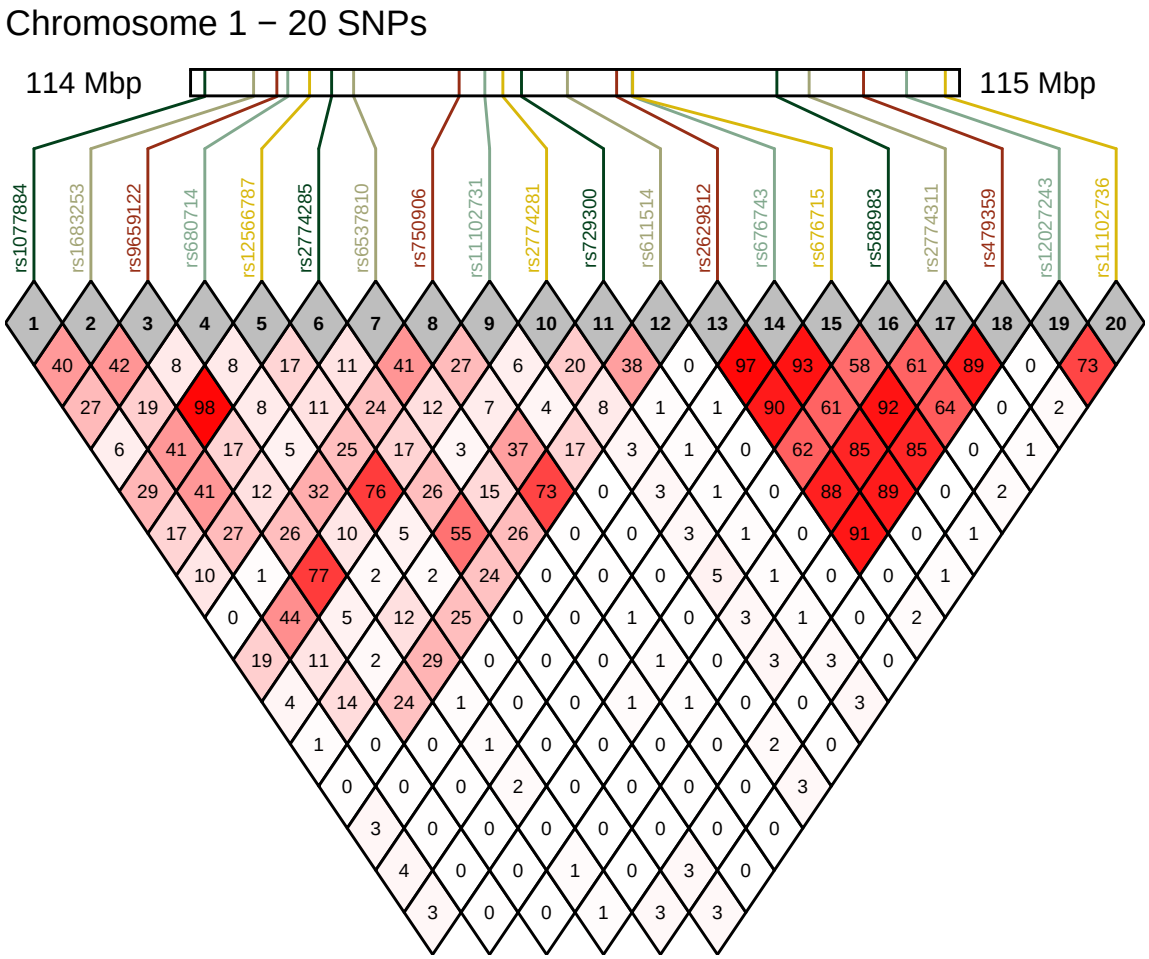

Figure 1: Linkage disequilibrium of 20 SNPs in the 1p13.2 region shows a well-defined haplotype with strong correlations from SNPs 13 to 18, and weak and diffuse correlations from SNPs 1 to 12.

3.2. Large well-defined haplotype in the 8p23.1 region

The chromosome 8 region 8p23.1 is the largest inversion haplotype in humans and includes hundreds of strongly correlated SNPs from 7 Mbp to 13 Mbp. Several other smaller inversions haplotypes with large and strong correlations can be found genome-wide. The 8p23.1 region includes the BLK gene (Namjou, Ni, Harley, Chepelev, Cobb, Kottyan, Gaffney, Guthridge, Kaufman, and Harley 2014) which is associated with systemic lupus erythematosus and encodes an enzyme involved in cell processes regulation called a kinase.

Here, the functions `snprelate_ld` and `gtable_ld` are called separately to demonstrate how to visualize a pre-computed linkage disequilibrium data frame, `df_ld`, using a SNPs annotations data frame, `df_snp`.

Visualizing all 305 SNPs between 11 Mbp and 12 Mbp shows that hundreds of SNPs located up to 1 Mbp from each other are strongly correlated and form a large haplotype (Figure 3).

```
R>   snp_idxs_8p23 <- select_region_idxs(qc$gdata, chromosome = 8,
```

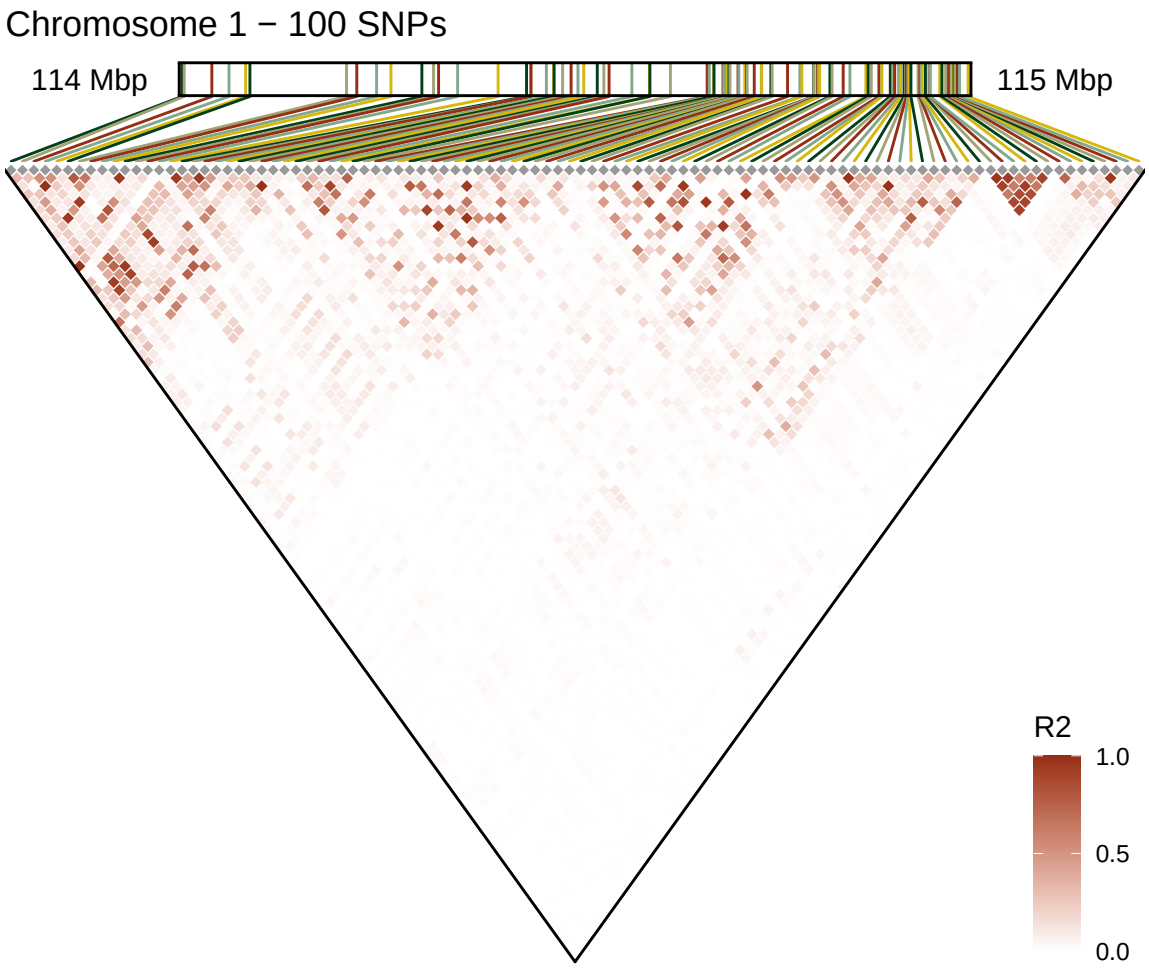

Figure 2: Linkage disequilibrium of 100 SNPs in the 1p13.2 region shows weakly correlated groups of 10-20 successive SNPs forming diffuse haplotypes, and a well-defined haplotype of 6 SNPs on the right side.

```
+      position_min = 11e6, position_max = 12e6)
R>   df_ld <- snprelate_ld(qc$gdata, snps_idx = snp_idxs_8p23, quiet = TRUE)
R>   plt <- gtable_ld(df_ld, df_snp = gdata_snps_annots(qc$gdata))
R>   grid::grid.draw(plt)
```

3.3. Large and diffuse haplotypes in the MHC region

The chromosome 6 *MHC* region between 29 and 35 Mbp is a complex region known to encode key components of the immune system. It includes the *HLA* genes (Simmonds and Gough 2007), known to be the most associated with several autoimmune diseases, as *HLA-B* and *HLA-DR*, and has a particularly high SNP density compared to other genomic regions. It is also known to be strongly associated with human ancestry.

Visualizing 187 SNPs between 32,2 and 32,8 Mbp shows large but diffuse correlations between tens of SNPs: the haplotypes are larger than in the 1p13.2 region, although they are not as

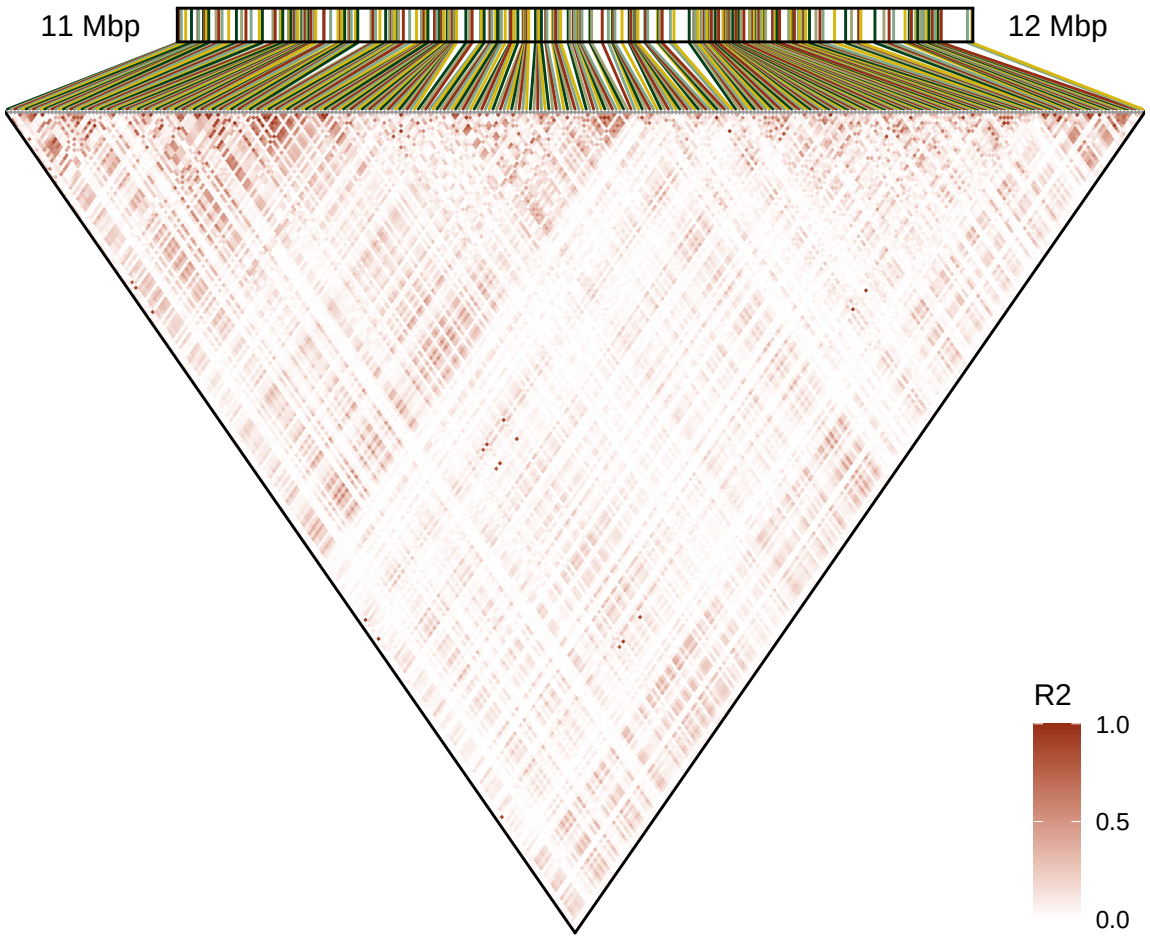

Figure 3: Linkage disequilibrium of 305 SNPs in the 8p23.1 region, the largest inversion haplotype, with hundreds of strongly correlated SNPs forming one large haplotype.

strongly correlated and well-defined as in the 8p23.1 region, thus indicating a high variety of weakly correlated signals (Figure 4).

```
R>   snp_idxs_hla <- select_region_idxs(qc$gdata,
+      chromosome = 6, position_min = 32.2e6, position_max = 32.8e6)
R>   plt <- gtable_ld_gdata(qc$gdata, snp_idxs_hla)
R>   grid::grid.draw(plt)
```

3.4. Gene names annotations with biomaRt

The `gdata_add_gene_annots` function enables to download the gene names of SNPs using the **biomaRt** R package to further investigate specific regions. In the code below, the commented line is the standard way to add gene names annotations to the Genotype Data object using SNP indexes, but is replaced by the `gdata_add_gene_annots_hladr_example` function to avoid downloading the annotations each time this article is generated.

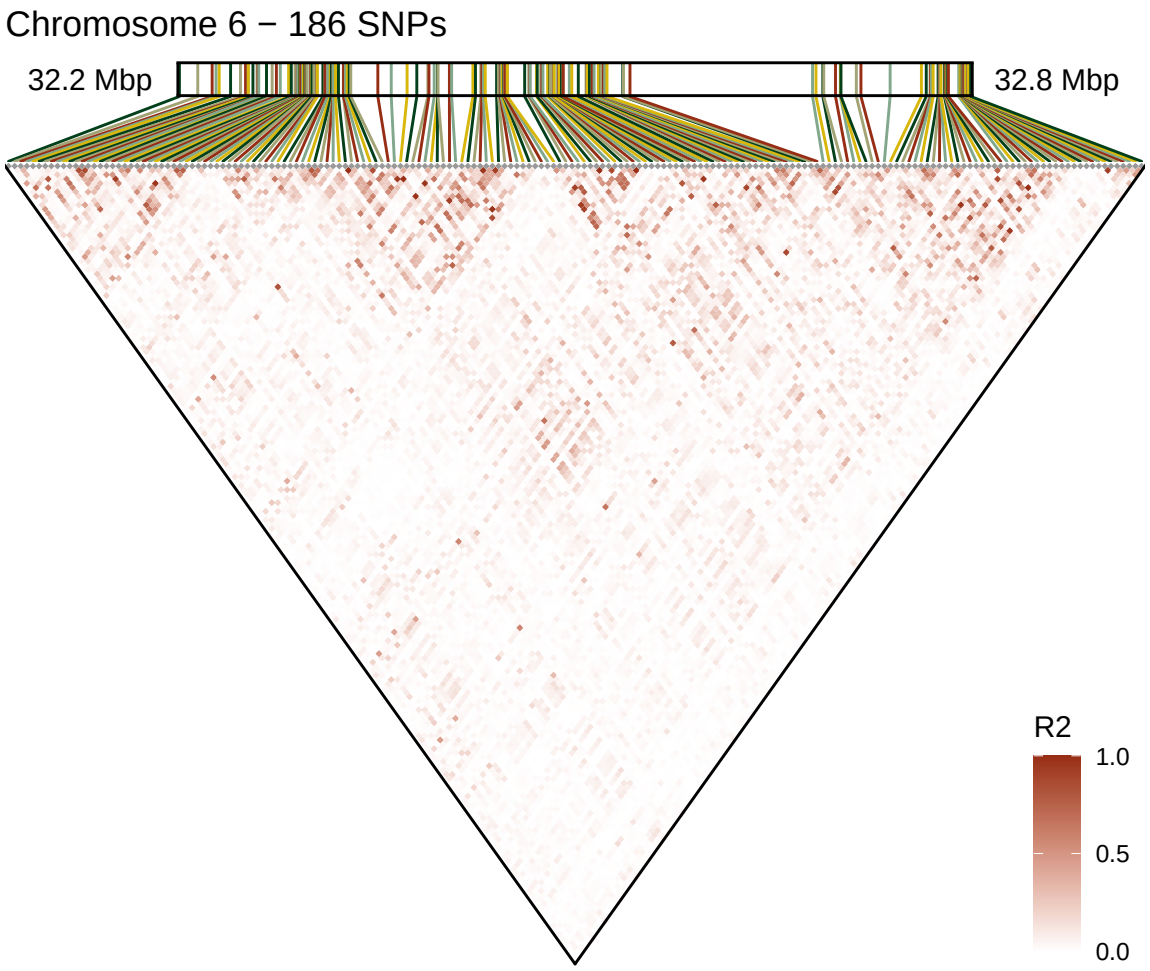

Figure 4: Linkage disequilibrium of 187 SNPs in the *MHC* region reveals large but diffuse correlations between tens of SNPs, indicating a high variety of weakly correlated signals.

Results show that the chromosome 6 region between 32.5 and 32.8 Mbp includes the genes *HLA-DRA*, *HLA-DRB1*, *HLA-DQA1* and *HLA-DQB1* (Figure 5). SNPs with identical gene names are more likely to form haplotypes, although not in this example.

```
R>   snp_idxs_hladr <- select_region_idxs(qc$gdata,
+      chromosome = 6, position_min = 32.5e6, n_snps = 20, offset = 9)
R>   # qc$gdata <- gdata_add_gene_annots(qc$gdata, snp_idxs_hladr)
R>   qc$gdata <- gdata_add_gene_annots_hladr_example(qc$gdata, snp_idxs_hladr)
R>   plt <- gtable_ld_gdata(qc$gdata, snp_idxs_hladr, labels_colname = 'gene')
R>   grid::grid.draw(plt)
```

4. Combining association studies results

Genome-wide association studies compare the frequency of each SNP between groups of in-

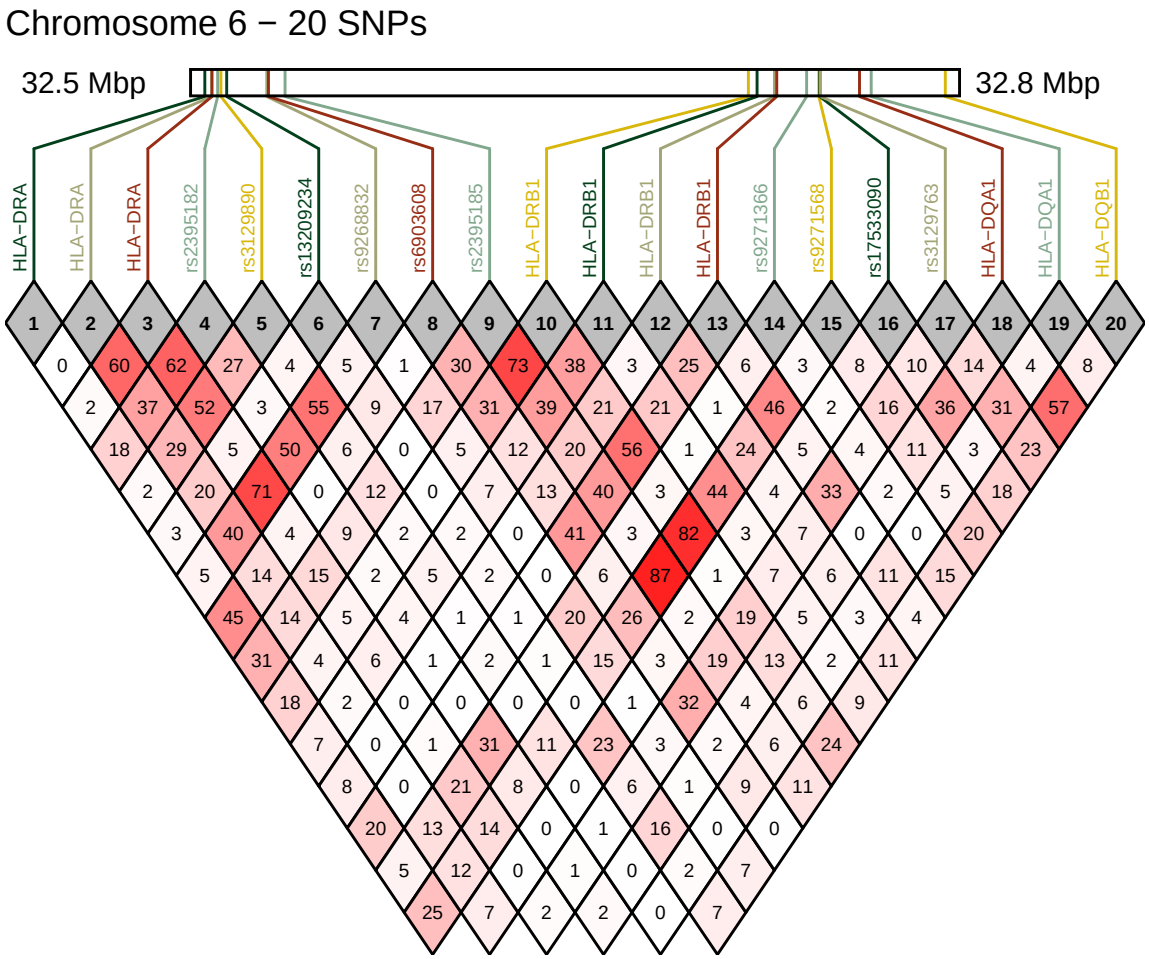

Figure 5: Chromosome 6 region between 32.5 and 32.8 Mbp, with genes *HLA-DRA*, *HLA-DRB1*, *HLA-DQA1* and *HLA-DQB1*.

dividuals sharing the same trait or disease. Due to linkage disequilibrium, SNPs significantly associated with disease are usually localized in specific regions and are surrounded by moderately associated SNPs, leading visualizations of association scores by chromosomic position to form skyline-like images, called Manhattan plots. The most associated regions can be further investigated by visualizing their linkage disequilibrium to identify which groups of SNPs are correlated and form haplotypes.

Here, an association study is performed based on the geographical region of individuals, using 1,218 SNPs from the chromosome 6 *MHC* region between 29 and 33 Mbp from 157 Europeans and 162 Middle Easterners and North Africans. After applying a chi-squared test, p-values are corrected using the false discovery rate (FDR) method (Benjamini and Hochberg 1995) and the gene names of the 20 most associated SNPs are fetched using the **biomaRt** package (pre-downloaded as previously).

Linkage disequilibrium visualization of the 20 most associated SNPs reveals that SNPs from the *CCHCR1* and *HLA-B* genes are weakly correlated (Figure 6).

```
R>   snp_idxs_mhc <- select_region_idxs(qc$gdata,
+      chromosome = 6, position_min = 29e6, position_max = 33e6)
R>   df_assocs <- chisq_pvalues_gdata(qc$gdata, snp_idxs_mhc)
R>   df_top_aim <- subset(df_assocs, rank(-pvalues, ties.method = 'first') <= 20)
R>   #qc$gdata <- gdata_add_gene_annots(qc$gdata, rownames(df_top_aim))
R>   qc$gdata <- gdata_add_gene_annots_aim_example(qc$gdata, rownames(df_top_aim))
R>   plt <- gtable_ld_associations_gdata(df_top_aim, qc$gdata,
+      labels_colname = 'gene')
R>   grid::grid.draw(plt)
```

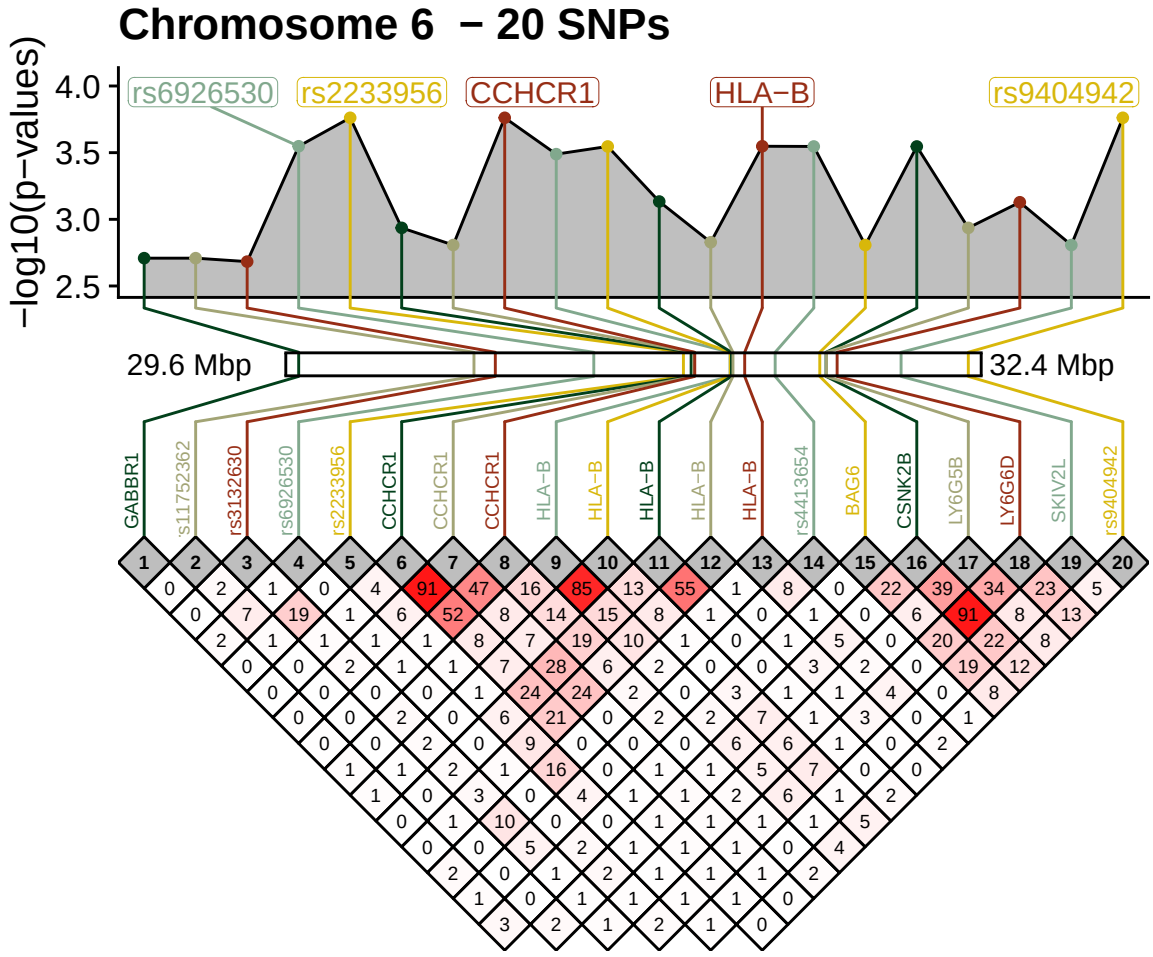

Figure 6: Linkage disequilibrium visualization of the 30 most associated SNPs reveals that SNPs from the *CCHCR1* and *HLA-B* genes are weakly correlated.

Linkage disequilibrium visualization of all the 68 SNPs significantly associated (FDR p-value < 0.05) shows the diffuse haplotype formed by SNPs from the *CCHCR1* and *HLA-B* genes (Figure 7).

```
R>   plt <- gtable_ld_associations_gdata(df_assocs, qc$gdata,
```

```
+       labels_colname = 'gene')
R>    grid::grid.draw(plt)
```

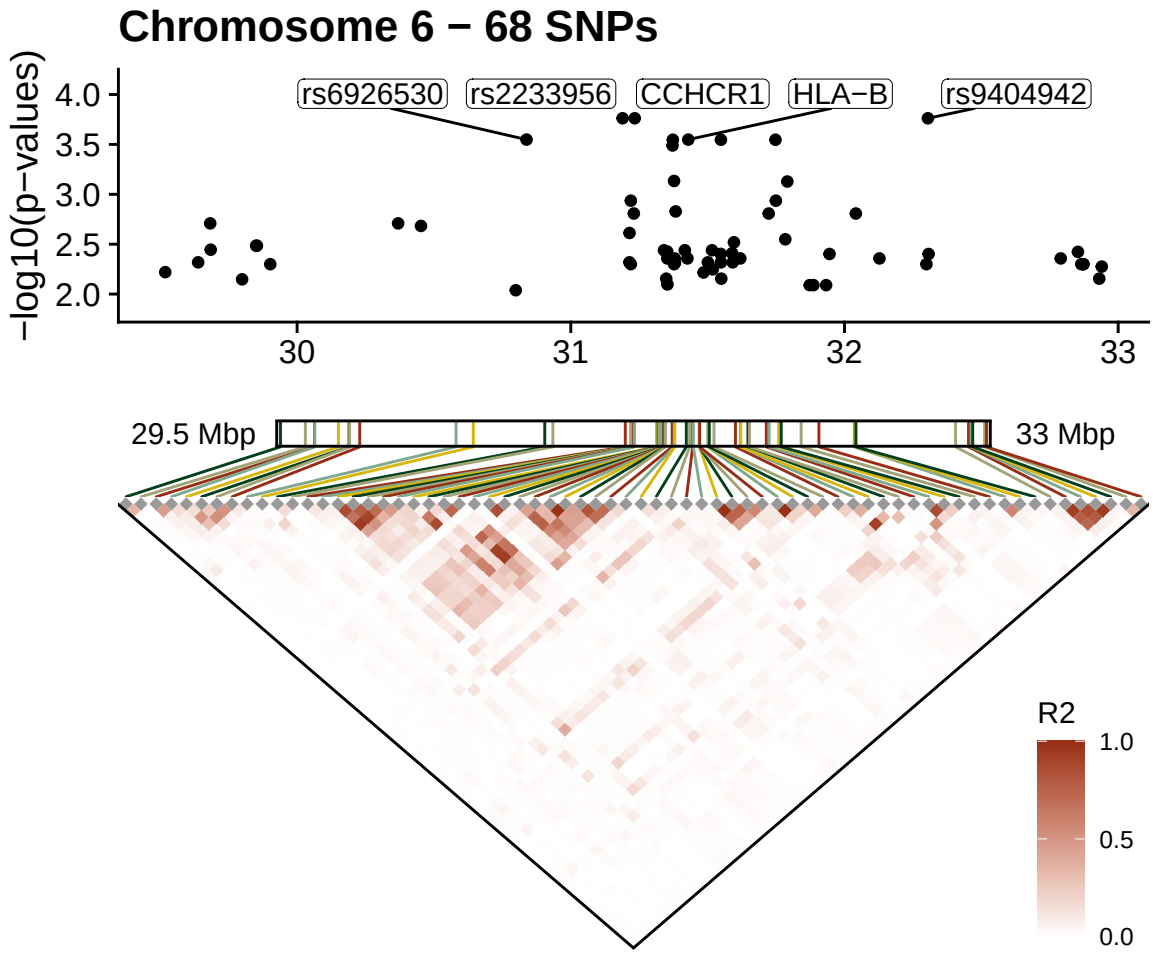

Figure 7: Linkage disequilibrium visualization of all the 68 SNPs significantly associated (FDR p-value < 0.05) shows the diffuse haplotype formed by SNPs from *CCHCR1* and *HLA-B* genes.

5. TagSNP feature selection

The `gtable_ld_gdata` function can be used to visualize the effect of feature selection by MAF filtering and TagSNP selection using the `maf` and `r2` parameters. Using the previous set of 20 SNPs from the 1p13.2 region, 5 SNPs are removed after applying TagSNP selection ($r^2 = 0.8$, 500 kbp window) (Figure 8).

```
R>    plt <- gtable_ld_gdata(qc$gdata, snp_idxs_1p13, r2 = 0.8)
R>    grid::grid.draw(plt)
```

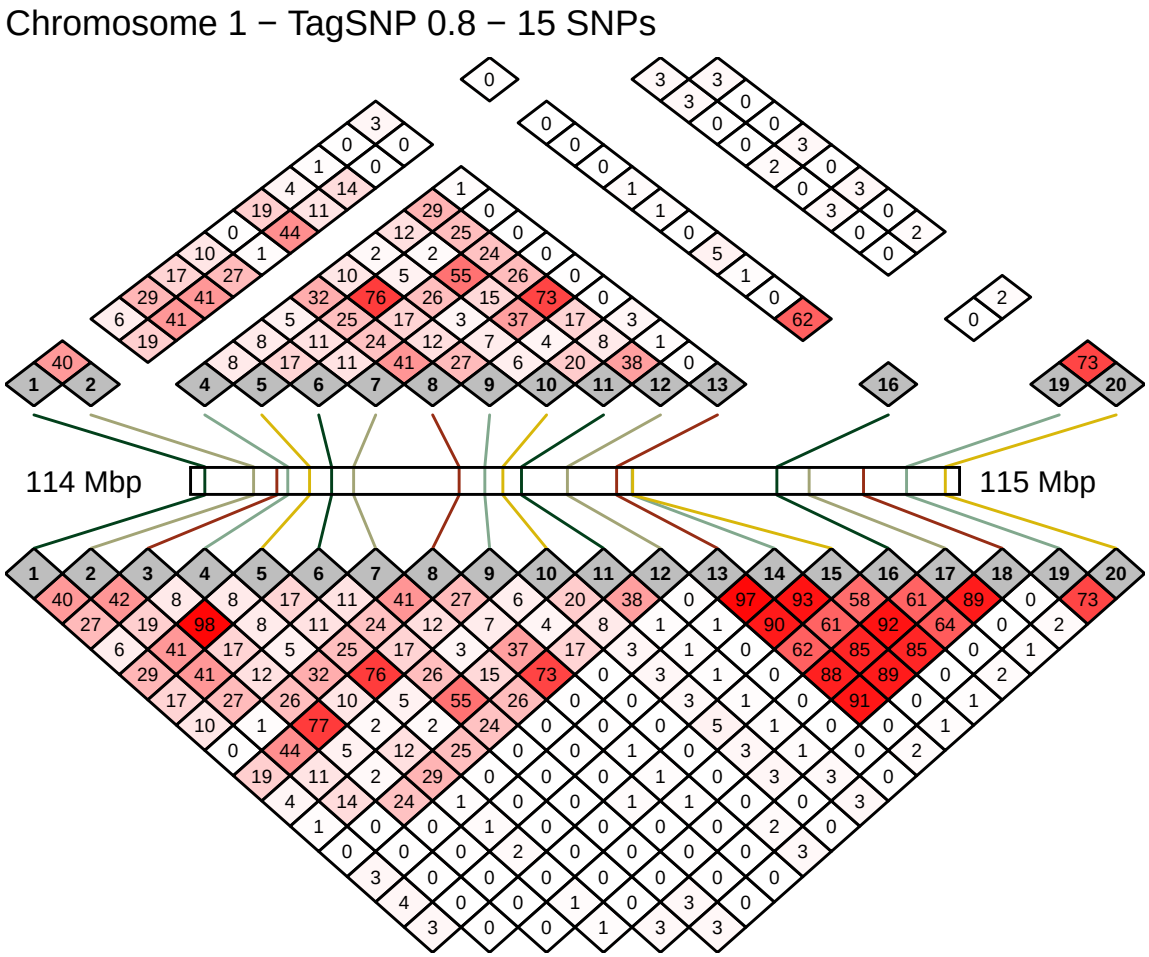

Figure 8: TagSNP selection of 20 SNPs in the 1p13.2 region removes 5 SNPs ($r^2 = 0.8$, 500 kbp window).

Feature selection visualization also uses points by default for more than 40 SNPs. Using the previous set of 100 SNPs from the 1p13.2 region, 27 SNPs are removed after TagSNP selection ($r^2 = 0.8$, 500 kbp window) (Figure 9).

```
R>   plt <- gtable_ld_gdata(qc$gdata, snp_idxs_1p13_large, r2 = 0.8)
R>   grid::grid.draw(plt)
```

6. Conclusions

The **snplinkage** provides linkage disequilibrium visualizations for small and large sets of SNPs that can be combined to association studies results or be used to investigate feature selection methods. The correlation patterns found in the 3 regions studied in this article illustrate the complexity and interdependence of genetic variations found genome-wide.

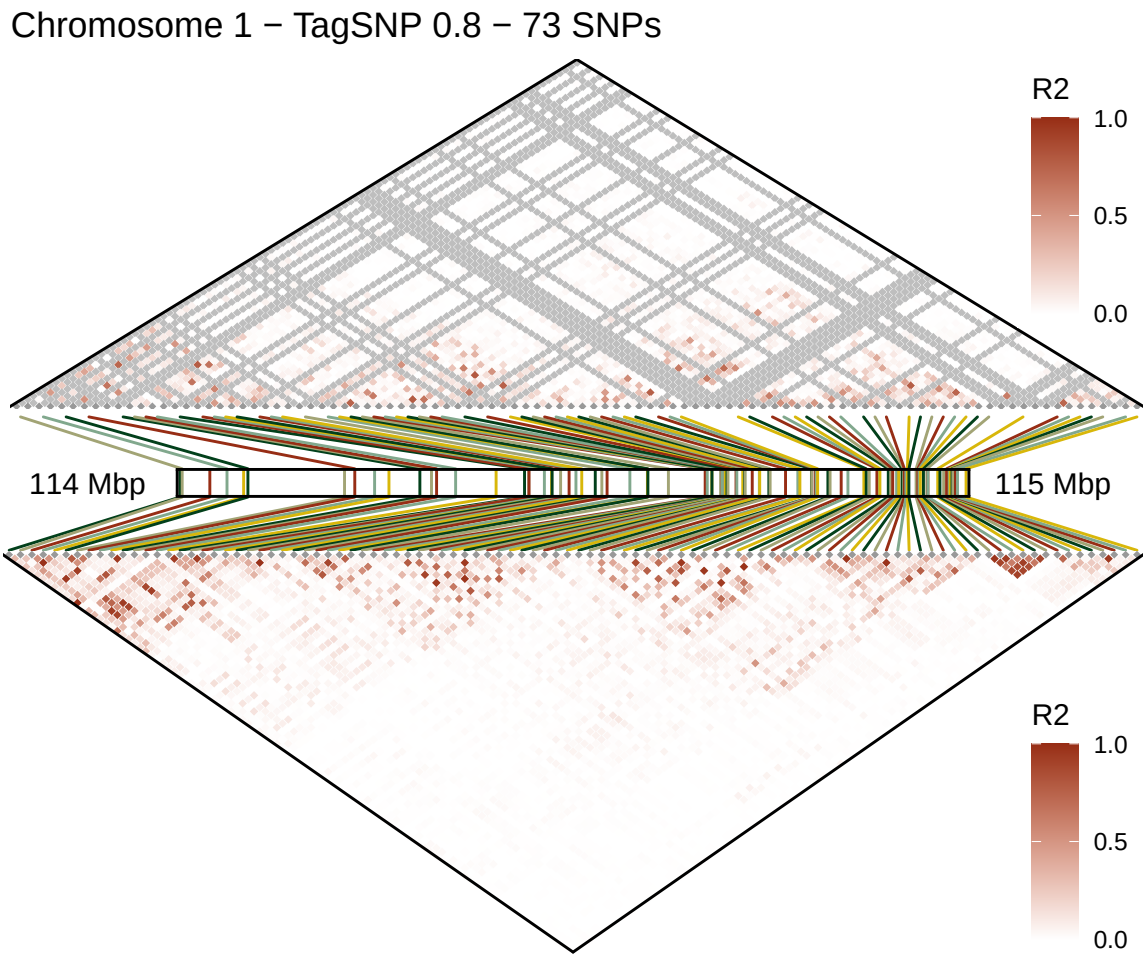

Figure 9: TagSNP selection of 100 SNPs in the 1p13.2 region removes 27 SNPs, indicated by grey lines in the correlation values ($r^2 = 0.8$, 500 kbp window).

7. Further reading

The code in this package was used to perform large-scale visualizations of up to 500 SNPs of three genomic regions (*MHC*, *1p13.2*, *8p23*) in a population of more than 1,000 systemic autoimmune diseases patients and healthy controls, sampled by the European research project PreciseSADs, in my Ph.D. thesis (Charlon 2019).

8. Acknowledgements

The package uses code first published in the **snpclust** package (https://github.com/ThomasChln/snpclust), which was co-authored with Karl Forner, Alessandro Di Cara and Jérôme Wojcik.

References

Awadalla MS, Thapa SS, Burdon KP, Hewitt AW, Craig JE (2011). "The association of hepatocyte growth factor (HGF) gene with primary angle closure glaucoma in the Nepalese population." *Molecular vision*, **17**, 2248.

Begovich AB, Carlton VE, Honigberg LA, Schrodi SJ, Chokkalingam AP, Alexander HC, Ardlie KG, Huang Q, Smith AM, Spoerke JM, *et al.* (2004). "A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis." *The American Journal of Human Genetics*, **75**(2), 330–337.

Benjamini Y, Hochberg Y (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300.

Charlon T (2019). *Genetic clustering for the discovery of a new classification of systemic autoimmune diseases*. Ph.D. thesis, University of Geneva.

Group ISMW, *et al.* (2001). "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms." *Nature*, **409**(6822), 928.

Namjou B, Ni Y, Harley IT, Chepelev I, Cobb B, Kottyan LC, Gaffney PM, Guthridge JM, Kaufman K, Harley JB (2014). "The Effect of Inversion at 8p23 on BLK Association with Lupus in Caucasian Population." *PloS one*, **9**(12), e115614.

Simmonds M, Gough S (2007). "The HLA region and autoimmune disease: associations and mechanisms of action." *Current genomics*, **8**(7), 453–465.

Affiliation:

Thomas Charlon, Ph.D.
E-mail: charlon@protonmail.com