Basic Guide to HLA

Written by: Seth A Sharp
Email: S.sharp@exeter.ac.uk

You may have seen HLA alleles represented in any of these forms: DR3, DR3-DQ2, DQ2.5, DR2-DQ2.5 which actually often have the same intended meaning. In the past HLA alleles were discovered by reacting one patient's serum with white blood cells from another patient, as different reactions were observed different alleles were documented such as DR3, DR4, DR15 etc. Later molecular study identified genes within these serotypes such as HLA-A, DRB1, DQB1 and subtypes of these alleles were named based on both serological reaction and molecular product for example DRB1*03:01 where the 03 represents a DR3 serotype coding protein subtype 01. This can be expanded as far as 8 digits e.g. DRB1*03:01:01:02 to describe variation in genetic sequence in both coding and non-coding regions.

What HLA alleles mean in terms of genetic variation

We will use made up sequence as an example, in reality the sequence for each allele is far longer.

HLA-DRB1*[02]

ATATATTTTTTAGTCCCCGTGAGTAAATAGGGCTATTTTAGTCCCC
GTGAGTCTGTTAGGGGTCGGTAGTCGTGAGTAAATATTAGTCCC
CCCGTAAATATTAGTCGTAAATATTAGTCTAGTCCCCGTGATAGTT
TGGCTCTCTACTCTACTACTAGTCGTGAGTAAATCATAGTGTGTA
GTCGTGAGTAAATATCCCCCCTACTGTCGTGAGTAAATATGTGTA

2-digit resolution: imagine this sequence is found on one two copies of chromosome 6. The yellow area represents the base pairs that correspond to the DRB1 gene with serotype DR2. There is significant genetic variation but as long as the serotype stays the same then the allele is the same.

HLA-DRB1*02:[01]

ATATATTTTTTAGTCCACGTGAGTAAATAGGGCTATTTTAGTCCCC
GTGAGTCTGTTAGGAGTCGGTAGTCGTGAGTAAATATTAGTCCCC
CCGTAAATATTAGTCGTAAATATTAGTCTAGTCCCCGTGATAGTT
TGGCTCTCTACTCTACTACTAGTCGTGAGTAAATCATAGTGTGTA
GTCGTGAGTAAATATCCCCCCTACTGTCGTGAGTAAATATGTGTA

4-digits resolution: Looking at the DRB1*02 sequence we can identify variation that is specific to encoding a protein subtype. The sequence still has variation, but we can be sure that we will be coding the same 02:01 protein and this will be different from 02:02 etc. Past this level the biological implications are much less, thus for genetic study we normally only care about 4-digit allele types.

HLA-DRB1*02:01:[01:02]

ATATATTTTTTAGTCCACGTGAGTAAATAGGGCTATTTTAGTCCCC
GTGAGTCTGTTAGGAGTCGGTAGTCGTGAGTAAATATTAGTCCCC
CCGTAAATATTAGTCGTAAATATTAGTCTAGTCCCCGTGATAGTT
TGGCTCTCTACTCTACTACTAGTCGTGAGTAAATCATAGTGTGTA
GTCGTGAGTAAATATCCCCCCTACTGTCGTGAGTAAATATGTGTA

6/8-digit resolution: Let's say the first part of our sequencing is in a non-coding region and the second part is in a coding region. At 6 digits we're talking about the coding part being a fixed sequence, at 8 digits we're talking about the non-coding part as a fixed sequence. Thus at 8 digits you are describing a single fixed sequence. You may also see a letter as a 9th digit but this describes changes in expression rather than genomic sequence.

How the nomenclature is used

Now you hopefully understand the relevance of serotype and what that means in terms of genetic variation you will understand the different nomenclature better. HLA alleles are subdivided by class of protein produced, for example class 1 proteins are encoded by HLA-A,B,C genes and class 2 by HLA-DR, DQ, DP, DO, DM. Class 3 encodes a less important system called the complement system. Biologically a particular section of HLA-DR and DQ has demonstrated the most importance in disease, this is the genes DRB1 – DQA – DQB1.

So when people talk about DR3 in Type 1 diabetes technically they could mean DRB1*03:XX:XX:XX so any protein, not a fixed sequence just the same serotype. Similarly, DR3-DQ2 tells us a bit more but still isn't perfect, now we know the DQB part of the chain is serotype 2. Then expanding to DR3-DQ2.5 tells us the DQA part of the chain is serotype 5. Now we know we're dealing with DRB1*03:XX – DQA1*05:XX – DQB1*02:XX which is a haplotype i.e. a "block" of genetic variation inherited together.

Table 1: An example of the kind of language you might hear talked about in Type 1 diabetes and what is intended versus what is likely meant.

What you hear	What it actually means	What they likely meant
DR3	DRB1*03:XX – DQA1*XX:XX – DQB1*XX:XX	DRB1*03:01 – DQA1*05:01 – DQB1*02:01
DR3-DQ2	DRB1*03:XX – DQA1*XX:XX – DQB1*02:XX	DRB1*03:01 – DQA1*05:01 – DQB1*02:01
DR3-DQ2.5	DRB1*03:XX – DQA1*05:XX – DQB1*02:XX	DRB1*03:01 – DQA1*05:01 – DQB1*02:01
DQ2	DRB1*XX:XX – DQA1*XX:XX – DQB1*02:XX	DRB1*03:01 – DQA1*05:01 – DQB1*02:01
DQ2.5	DRB1*XX:XX – DQA1*05:XX – DQB1*02:XX	DRB1*03:01 – DQA1*05:01 – DQB1*02:01

What's the deal with the DQX.X nomenclature? It is a kind of lazy shorthand to describe what's going on with DQA1 and DQB1, traditional immunogeneticists do not like it as it is not consistent with their carefully developed naming conventions. You will still see it frequently in genetics publications as bioinformaticians like it (because it's lazy).

In Type 1 diabetes, fortunately we know from modern genetic study that what all of these different nomenclatures are trying to describe is DRB1*03:01 – DQA1*05:01 – DQB1*02:01. However, people who have been in the field a while will still talk in serotypes such as just saying DR3 as they were taught the serology only and don't know the genetic implications. You will find this in many diseases e.g. the use of DR15 as shorthand to describe DRB1*15:01 – DQA1*01:02 – DQB1*06:02 as the highest risk allele for multiple sclerosis.

How this ties to SNP array data and imputation

So two important points from above to keep in mind (1) most of the biological insight can be determined by 4 digit HLA alleles e.g. DRB1*02:01 and (2) at 4 digits a lot of the sequence is fixed but not all of it and specific parts of sequence correspond to specific 4 digit alleles e.g 02:01 vs 02:02 etc (c) a haplotype describes a block of multiple genes often inherited together e.g. DRB1-DQA1-DQB1

Given a set of SNPs how do we determine 4-digit HLA allele type or even better full haplotypes? We can combine information from many SNPs to infer the HLA allele type, this is HLA imputation such as SNP2HLA, HIBAG, HLA*IMP etc. Accuracy can be good, but the downside is we then need lots of SNPs just to identify one gene type, that makes it practically really difficult. We also need to determine which chromosome each allele type is on to determine the full haplotypes which we call **phasing** and can be really difficult.

What we can do instead is pull out single SNPs that correlate highly with HLA haplotypes or specific alelles. We call these SNPs proxy variants. Let's take rs2187668, it tags the haplotype DRB1*03:01 – DQA1*05:01 – DQB1*02:01 with correlation r^2=0.98 in white European populations, that's exceptionally accurate and we only had to genotype a single SNP in a huge block of sequence!

Special care should be taken in non-White ethnic populations as linkage disequilibrium (LD) structures can vary massively, so we cannot assume proxy SNP and haplotypes discovered in white European ancestry populations correlate in the same manner.