BBS3011 Computer lab 1: Genetic Diversity of Populations and Disease Variants

Introduction

In year 2, in course BBS2002, you have already encountered the Ensembl and NCBI genome databases. Also, you have briefly explored the resources they offer for the study of genetic variations and differences in allele frequencies between populations in the world, including OMIM, Ensembl SNP pages, and NCBI dbSNP.

Today we will use these resources again, refreshing our memories and applying them to study the information they provide in more detail. For your reference, the 'Reference guide online resources' that has been provided to you last year, is added to the practical materials on Student Portal again. Chapter 3 deals with genetic variations and population diversity, chapter 1 and 2 introduce Ensembl and NCBI and how to find (basic information on) genes and transcripts.

In addition, we will use the GWAS Catalog, which is available from https://www.ebi.ac.uk/gwas/ and contains results of Genome Wide Association Studies.

Next week's computer lab will further explore the consequences of genetic variations at protein level and for drug functioning. Today we will mainly focus on alleles, their frequencies, global differences in occurrence, and the connections to human disease.

The topic of the practical continues on the disease you studied in the first journal club, Systemic Lupus Erythematosus (SLE). We will revisit one of the genes discussed in the paper by Goulielmos *et al.*[1], PTPN22. We hope that this lab shows you how you could find genetic variants that contribute to disease risk, and which studies have been performed, either when you are studying the molecular biology or pathophysiology of a disease, or when you are investigating a disease in your own research, for example during your internship.

Good luck and enjoy!

Practical coordinator: Lars Eijssen, l.eijssen@maastrichtuniversity.nl

[1] Goulielmos GN, Zervou MI, Vazgiourakis VM, Ghodke-Puranik Y, Garyfallos A, Niewold TB. The genetics and molecular pathogenesis of systemic lupus erythematosus (SLE) in populations of different ancestry. Gene. 2018 Aug 20;668:59-72. doi:10.1016/j.gene.2018.05.041. Review. PubMed PMID: 29775752.

Variations in the PTPN22 gene and SLE susceptibility

Assignment 1 – Information given in the Goulielmos review paper 🕘 15 min.

As described in the paper by Goulielmos *et al.*, the gene PTPN22 is associated to SLE risk as well. In the introductory paragraph of section 4, the authors state that "the autoimmune disease risk variant of the protein tyrosine phosphatase type 22 gene (PTPN22) is almost absent in African-derived populations". In section 5.1 of the paper, more information on the specific risk allele for this gene is given.

a) What function does the protein encoded by the PTPN22 gene have, according to the paper?
It is a protein tyrosine phosphatase which functions as a negative regulator of T-cell activation.

b) Which variant is described in the paper: what is the nucleotide change and what is the position?
It is a C -> T change at position 1858.

c) Which of the alleles is the risk allele for this variant?
The T allele is given to be associated with a higher risk of SLE.

d) Which other diseases besides SLE, is the variant associated with, based on the information in the paper?
Autoimmune thyroid disease, Wegener's granulomatosis, myasthenia gravis, systemic sclerosis, juvenile idiopathic arthritis, rheumatoid arthritis, generalized vitiligo, Addison's disease, and type I diabetes.

Later in the section, the paper discusses the SNP with identifier rs2476601.

e) Which amino acid change at which position does this variant cause?
An arginine (R) to tryptophan (W) change at position 620 in the protein.

f) Is this likely to be the same variant as the paper refers to at the start of section 5.1?
Yes it is, as a change in amino acid 620, should be at position 1858, 1859, or 1860 from the start codon.

g) In which population(s) have associations between variations in the PTPN22 gene and SLE been found according to the paper and which ones? In which populations not?
Europeans and Hispanics for a panel of 107 SNPs in the gene. For SNP rs2476601 mainly Europeans (European Americans), but also Hispanics – the SNP was found to be associated with skewing of cytokine levels towards higher IFNα activity, a factor known to correlate with SLE (and lower TNFα). SNP rs3765598 has been mainly found in Hispanics. No significant associations have been found in Africans (African Americans) and Asians.

Assignment 2 – OMIM: Online Mendelian Inheritance in Man 30 min.

Look up SLE in OMIM (you may have already done this during the Journal Club).

a) What is the OMIM identifier of SLE?
152700.

b) Which genes are associated with SLE and in which way and with which inheritance mode, according to OMIM?
PTPN22, FCGR2B, CTLA4, TREX1, DNASE1 are associated with susceptibility to SLE; FCGR2A is associated with susceptibility to Lupus Nephritis, a consequence.
All risk alleles follow an autosomal dominant (AD) inheritance patterns.

Scroll down to the molecular genetics section and find the information about PTPN22.

c) Which allele is given as the risk allele for the variant that is also mentioned by the Goulielmos paper? Does this correspond to what the paper indicates?
The T allele is indicated to increase to risk of SLE. Yes, it corresponds.

d) What are the odds ratios for heterozygotes and homozygotes for this variant and to which inheritance mode does this correspond?
The Odds Ratio (OR) is 1.37 for heterozygotes and 4.37 for homozygotes. This corresponds to a multiplicative inheritance mode: the OR for homozygotes is much larger than double that for heterozygotes. If we have to pick either autosomal recessive or dominant (from classical genetics perspective), it depends on whether you would consider 1.37 already a strong enough difference from 1. Clearly the OMIM page does, as it indicates an AD inheritance for this gene and SLE. The notions of AR and AD are of course more difficult for diseases that are not strictly monogenic, as the total risk is composed of many contributing factors.

e) Which other PTPN22 variant is reported to be associated with SLE?
The rs33996649 variant, G788A (R263Q).
Note: for the rs2476601 variant mentioned, OMIM only gives the rs number, but clicking the link to Ensembl (or referring back to the Goulielmos paper) shows you it is the previously mentioned one R620W.

f) Which other gene is reported to interact with PTPN22 with respect to SLE risk? Briefly describe how the interaction connects to SLE risk.
The c-Src tyrosine kinase CSK. By physical interaction of its encoded protein with the phosphatase LYP, encoded by PTPN22, it can modify the activation state of downstream Src kinases, such as LYN in lymphocytes. Increased CSK expression augments inhibitory phosphorylation of LYN. The LYP-CSK complex activates B-cells and leads to increased susceptibility to SLE.
Note: this is also mentioned in the Goulielmos paper in section 5.1: "disrupting the Lyp-Csk interaction".

Open the page about PTPN22 in OMIM, by clicking it from the SLE table or using the search box.

g) To which other diseases is this gene associated and in which way and with which inheritance mode?
Susceptibility to T1DM (autosomal recessive, AR) and susceptibility to rheumatoid arthritis (inheritance mode not given).

h) Which type of cells does this OMIM page report to play a central role in the connection of PTPN22 variants and risk of the diseases mentioned in the previous question?
T-cells.
Note: the Goulielmos paper also mentions the function of the PTPN22 protein as a negative regulator of T-cell activation.

i) What does this page tell us about the susceptibility effect of the G788A variant that we have already encountered?
It leads to a reduced susceptibility (by reducing phosphatase activity of the protein).

Assignment 3 – The R620W variant in Ensembl and dbSNP 🕘 40 min.

First look up the PTPN22 R620W variant (rs2476601) in Ensembl.

a) Do the alleles and the positions in the RNA and protein correspond to what was indicated by the Goulielmos paper and OMIM?
Yes, they do.
Note: the alleles are given as A/G, but A is the minor allele: the order does not necessarily give the major allele first.
Note: as you have already seen in BBS2002, multiple exact positions are given for most SNPs, depending on the splice variant considered; one of those corresponds.
Also recall that positions in the naming of SNPs are positions in the coding sequence (so counting from the start coding), not positions in the mRNA (including 5'UTR).

b) What are the given overall minor allele frequency, and the highest population minor allele frequency?
Overall: 0.03; highest: 0.15.

We will now focus on the information with respect to allele frequencies in populations. We will use the 1000Genomes results, which include the most individuals, to draw some conclusions and verify what the Goulielmos paper states.

Remember that in Ensembl you can find the results of population studies for a SNP on a dedicated 'Population genetics' page.

The 1000Genomes results have been aggregated to some extent, and individual populations studies have been combined in five main regions, abbreviated as follows: AFR, African – AMR, Ad Mixed American – EAS, East Asian – EUR, European – SAS, South Asian.
For your information: you can find which population are included for each main region at: http://www.internationalgenome.org/faq/which-populations-are-part-your-study.

Explore the 1000Genomes results for rs2476601 in Ensembl, given by study identifier ss1292502155.

c) Complete the table:

Population	Number of individuals	Fraction G	Fraction A
AFR	661	0.997	0.003
AMR	347	0.964	0.036
EAS	504	1.000	0.000
EUR	503	0.906	0.094
SAS	489	0.987	0.013
total	2504	0.973	0.027

Note that the number of alleles is commonly reported for population studies, so to obtain the number of individuals the value has to be divided by 2.

d) Briefly describe what the results in the table show.
The A allele is overall quite rare. However there are substantial differences in its occurrence between populations, ranging from absent (East Asian), almost absent (African), only ~1% (South-East Asian), to rare (Admixed American), to almost 10% (European).

e) Does the minor allele occur more frequently in northern or in southern Europe? Explain your answer.
In Northern Europe. When looking at the individual populations, frequencies are somewhat higher in Northern/Western European populations (FIN, CEU, GBR) than in Southern European populations (IBS, TSI).

f) What are the genotype frequencies for the Europeans? Complete the table.

Population	Fraction GG	Fraction GA	Fraction AA
EUR	0.815	0.181	0.004

g) Are the values for the Europeans in Hardy-Weinberg equilibrium? Write down your computation. What does that mean?

Expected:
GG 0.906^2 = 0.821
GA 0.906*0.094*2 = 0.170
AA 0.094^2 = 0.009

So the SNP seems to be in Hardy-Weinberg equilibrium.
Note: to really compute this, one has to apply a statistical test, but that is not needed now. The conclusion is based on the very small deviation of the observed frequencies from the expected frequencies. Note also that although the frequency of the AA genotype is not even half of that expected, this does not mean much with such small fractions (the absolute deviation is still small).

The genotype frequencies are as expected based on the allele frequencies. This means there is no evidence of sample artefacts/errors, and no evidence of strong differences in evolutionary fitness between the genotypes.

h) How many homozygous individuals for the alternative allele have been observed in total over all populations in this 1000Genomes study?
Only 3 (out of 2540 individuals).

i) Do the 1000Genomes results confirm the statement in the paper about the frequency of the alternative allele in African-derived populations?
The paper says the alternative allele is almost absent from African-derived populations. This is also what the study results show (only 4 A alleles, versus 1318 G alleles; all A alleles are present in heterozygotes).

Now also look up the PTPN22 R620W variant (rs2476601) in NCBI dbSNP.

j) Does dbSNP also confirm the alleles and the positions in the RNA and protein as given by the Goulielmos paper and OMIM?
Yes; where the same notes hold as given for Ensembl.

Also look up the 1000Genomes results as presented by dbSNP; the study identifier of this study is ss1292502155.

k) Do the given frequencies correspond to those given by Ensembl?
Yes, they do.
Note: click the study identifier to open an easy readable table.

Assignment 4 – SLE and the PTPN22 gene in the GWAS Catalog 🕘 15 min.

Open the GWAS Catalog at https://www.ebi.ac.uk/gwas/ and search for Systemic Lupus Erythematosus or SLE.

Note: when showing the results, GWAS Catalog only shows part of the tables to keep the page compact. Click the 'Show more results' button for the part of the results where you want to see more.

a) Name six genes that are reported to have one or more variants associated with SLE?
CFB, STAT4, TNXB, DQA1, DQB1, IRF5 (and many more).

b) Why are there so many more genes reported in GWAS Catalog than there were described in OMIM?
GWAS Catalog reports on all associations found in GWAS studies. These include many genes with small effects on polygenic trait, or small modifier effects on familiar hereditary diseases. OMIM reports curated information and focuses on Mendelian diseases and on stronger and/or well-studied effects.

c) How many studies have reported the PTPN22 gene in the list of associations for SLE retrieved from GWAS Catalog?
Eight studies have reported it (use the find function of the browser).
Note that for six of the studies the given phenotype is "Systemic lupus erythematosus", for one it is "Systemic lupus erythematosus or rheumatoid arthritis" and for one it is "Pediatric autoimmune diseases".

d) Which are the SNPs reported to be associated to SLE for the PTPN22 gene by GWAS Catalog?
The reported SNPs are rs2476601 (this is the R620W variant) and rs6679677.

e) What is the approximate reported odds ratio for the risk of PTPN22 for each of these SNPs?
For both SNPs it is around 1.4; this is reported in the table of results.

f) How can you explain that we have not encountered the SNP rs6679677 in either the review paper or the OMIM curated report?
SNP rs6679677 is an intergenic variant (as mentioned by GWAS Catalog, click the rs number in the list of results). Its odds ratio is roughly equal to that of the rs2476601 SNP. It could well be that this SNP is only an associated marker/tagging SNP (that happens to be included in the measurement panel of several studies) without a direct functional effect itself.

Now search GWAS Catalog for the PTPN22 gene.

g) Give six traits other than SLE to which the PTPN22 gene has been associated.
Hypothyroidism, Hashimoto's thyroiditis, rheumatoid arthritis, vitiligo, late-onset myasthenia gravis, Graves' disease (and many more).