Chapter 3

Attempting to identify the genetic determinants of thiopurine-induced liver injury

3.1 Introduction

Thiopurines are a type of immunosuppressive drug that have found their application as treatments for a variety of conditions, including immune-mediated disorders and acute lymphoblastic leukaemia. They are also used as maintenance therapy for patients who have received an organ transplant to suppress the immune reaction to the graft. More recently, thiopurines have been used in conjunction with anti-TNF (‘combination therapy’) in order to improve the treatment outcomes of the biologic therapy [41, 96] (see Sections 1.4.4 and 2.1). Several thiopurines are available on the market: azathioprine (AZA), mercaptopurine (6-mercaptopurine or 6MP), and thioguanine (6-thioguanine or 6TG).

Despite the arrival of the biologic therapies, thiopurines are still used for remission maintenance in both Crohn’s disease and ulcerative colitis [103]. In addition, a mounting body of evidence suggests a significant advantage of prescribing thiopurines alongside anti-TNF in order to reduce the immunogenicity risk and, ultimately, improve treatment outcomes [41, 96]. There is also a growing interest in exploring the use of thiopurines alongside vedolizumab, although the evidence of clinical benefit is sparse [79].

Similar to other immunosuppressive treatment options for IBD, thiopurines have a burden on the patients' quality of life and may increase the risk of opportunistic infections (especially viral), lymphomas, myeloid disorders, and skin cancers [13]. In addition, patients treated with thiopurines have a high rate of adverse drug reactions of varying severity. In clinical trials, 0–15% of patients discontinued the treatment due to adverse drug effects [144, 69]. Some of the adverse treatment effects are rather trivial and can be managed – rashes, fevers, nausea. Others, like severe thiopurine-induced myelosuppression (TIM), can be potentially lethal and require utmost caution. Thiopurine-induced myelosuppression has a strong genetic component. Several risk-increasing variants in *TPMT* and *NUDT15* have been identified and are now being used for targeted clinical genotyping prior to treatment (see Section 1.4.5).

Another adverse effect causing major concern when prescribing thiopurine therapy is liver injury. The reported rates of thiopurine-induced liver injury (TILI) vary greatly across studies, with retrospective studies reporting a mean of 3% [69], while the only prospective study has reported a rate of 10% [16]. Several sub-types of TILI exist. Hepatocellular injury is the most common, most asymptomatic type of TILI that is associated with transaminase enzyme elevation. The condition occurs within 12 weeks of treatment start or after dose escalation [24, 185]. Hepatocellular TILI is usually resolved after thiopurine dosage reduction or treatment withdrawal [65]. Cholestatic liver injury is observed in 1 in 1,000 thiopurine-treated patients and is associated with jaundice, itching, and fatigue [80, 186]. Cholestatic TILI occurs between 2 and 12 months after the start of the treatment. The condition can often be mitigated by stopping the therapy, though it can continue after its cessation [80, 186]. Finally, TILI can present itself with both the symptoms and biomarkers of hepatocellular and cholestatic injury (and so will be referred to as 'mixed TILI').

Several risk factors have been associated with TILI: use of corticosteroids was associated with hepatocellular TILI; while concomitant anti-TNF appears to reduce the risk [16], nonalcoholic fatty liver disease, a condition more frequent in IBD patients, has been associated with a higher risk of hepatocellular TILI [166, 179].

As of September 2019, no genetic associations for thiopurine-induced liver damage have been reported. However, several associations, all within the HLA region, for non-thiopurine liver injury have been reported: HLA-B*57:01 for flucloxacillin (antibiotic) [45], HLA-A*02:01 and HLA-DRB1*15:01 for amoxicillin-clavulanate (AC, antibiotic) [114], HLA-B*35:02 for minocycline (antibiotic) [181], and HLA-A*33:01 for terbinafine (antifungal) [133]. The associations are thought to be drug-specific.

In a recent study, Cirulli et al. [39] performed a genome-wide association study of idiosyncratic drug-induced liver injury (DILI). They assembled a cohort of 2,048 individuals with DILI and 12,429 unmatched controls. The cohort was primarily of European ancestry (1,806 cases and 10,397 controls), but included a small number of African American and Hispanic cases with matched population controls. The authors also assembled a replication cohort, consisting of 113 individuals and 239,304 controls of Icelandic ancestry. The cases included subjects of cholestatic (26% amongst the primary European cohort), hepatocellular (41%), mixed (26%), and unknown (7%) DILI. The study included DILI cases caused by a variety of drugs, and herbal and dietary preparations.

The major finding of the study was the association in *PTPN22* – rs2476601 ($N_{cases} = 444$; OR=1.44; 95% CI, 1.28 to 1.62; P=1.2x10^{-9}), which was replicated in the Icelandic cohort (OR=1.48; 95% CI, 1.09 to 1.99; P=0.01). The association is primarily driven by amoxicillin-clavulanic acid combination therapy, which is used as an antibiotic (OR=1.62; 95% CI, 1.32 to 1.98; P=4x10^{-6} amongst the patients of European Ancestry). The authors argue that the association is not driven by any particular category of drugs, demonstrating that association has a consistent direction of effect (OR>1) across 39 drugs. However, only seven of these reach the nominal significance level of 0.05. Some therapies, like the antibiotic flucloxacillin, do not demonstrate any evidence of association ($N_{cases} = 195$; OR=1.24, P=0.18). The replication cohort was more homogeneous, consisting of individuals with DILI due to amoxicillin-clavulanic acid and other antimicrobial drugs. As such, it is reasonable to assume that the *PTPN22* association is not a universal risk variant for drug-induced liver injury, but it only causes an adverse reaction in a subset of treatments. The effect of the association on thiopurine-induced liver injury remains uncertain, as the cohort only had 10 cases on mercaptopurine (OR=1.72; 95% CI, 0.5 to 5.97; P=0.39). This will be revisited in the chapter below.

Thiopurines continue to be a widely used therapy for treating patients who suffer from inflammatory bowel disease and a variety of other conditions. Severe drug reactions, like the thiopurine-induced myelosuppression and liver injury, pose a serious challenge and require a better understanding of the associated risk factors in order to enable better therapeutic drug monitoring and personalised prescription. In this chapter, I describe the results from a genome-wide association study of TILI subjects that were recruited as a part of the Predicting Serious Drug Side Effects in Gastroenterology (PRED4) study.

3.2 Methods

3.2.1 Genotyping, cohort assembly, and quality control

Assembling the dataset

Patients were recruited as a part of the Predicting Serious Drug Side Effects in Gastroenterology (PRED4) study (REC number 11/SW/0222).

The final phenotype revision of the PRED4 TILI cohort included 859 individuals – 278 TILI cases (32%) and 581 controls (68%). Cases included 126 subjects with hepatocellular TILI (45% of the cases), 41 with cholestatic (15%), 106 with mixed (38 %), and 5 with an unknown type (2%).

Unfortunately, the genotyping of the entire cohort was not performed in one batch. 1,221 genotyped samples belonging to 786 phenotyped patients were identified (up to 4 genotyped samples per patient). The samples were spread across five genotyping cohorts:

- Two cohorts (153 [`broad1`] and 485 [`broad2`] samples) genotyped at the Broad Institute as a byproduct of the G4L WES pipeline (based on the Illumina Infinium Genome-Wide Association Study array, contain ~245,000 markers).
- One produced at the Sanger institute using the Illumina HumanCoreExome-12 – 245 samples previously included in the de Lange et al. study [49] (imputed to HRC, ~11 million markers [gwas3]).
- Two cohorts produced at the Sanger Institute using the HumanCoreExome-24 array – 150 (imputed to HRC, underwent QC described in de Lange et al. [49], [newwave]) and newly genotyped 188 samples (522,049 markers [newgeno]).

The 1,221 samples were combined together, taking an overlapping set of variants across all five cohorts (1,221 samples, 230,597 variants).

In order to verify to the correctness of the phenotype-to-genotype ID matching and to remove related samples, the genetic kinship across samples was calculated. The assumption was that samples from different genotyping cohorts, tentatively assigned to the same ID, should have a high PI_HAT score.

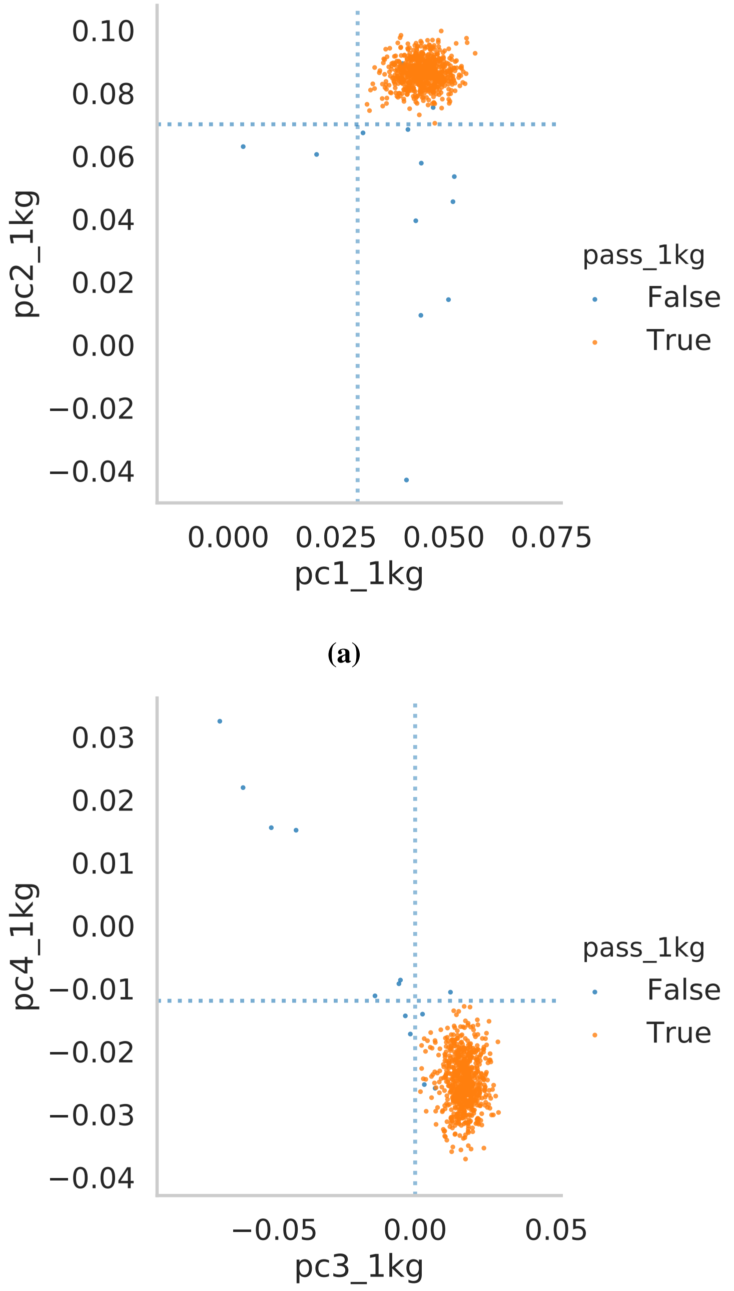
For calculating the relatedness, the full genotyping cohort was filtered to include only accurately genotyped, common SNPs (MAF > 5%, call rate > 0.95, $P_{hwe} > 10^{-6}$, 216,278 variants post-filtering). Variants in linkage disequilibrium with each other were removed via the LD-prune procedure ($r^2 = 0.2$, window = 500,000 bp, 77,053 variants post-filtering). In addition, five poorly genotyped samples were excluded (sample call rate < 0.8), as they resulted in spurious low-level relatedness with tens of otherwise unrelated samples.

The filtered cohort was used to run kinship estimation (identity by descent, similar to the method described in [151]). Pairwise relatedness was estimated for all individuals in the cohort. Pairs with PI_HAT > 0.18 were retained (473 pairs; PI_HAT = 0.1875 is halfway between the expected pi-hat for third- and second-degree relatives [7]; adjusted to be marginally lower in order to account for genotyping heterogeneity). Amongst the pairs of genotyped samples thought to belong to the same patient ID, the minimal PI_HAT was 0.96 (PI_HAT range is between 0 [unrelated] to 1 [duplicate samples or monozygotic twins]), suggesting that the ID matching was done correctly. One sample from each related pair was removed, prioritising samples with denser genotyping, to ensure that no two sample pair had a PI_HAT > 0.18. Post filtering, the cohort retained 778 samples.

Sample quality control

Next, I have removed poor-quality samples. Minimal variant-level QC was applied (MAF > 1%, missingness < 5%, P_{hwe} in controls $> 10^{-6}$) to avoid the sample-level metrics being affected by low-quality variants. Samples with a minimal call rate of 99% and within three standard deviations from the median of the heterozygosity ratio and the call rate were retained. The filter was applied on both individual cohorts and on the overall dataset. 34 samples were removed, retaining 744 samples.

The absolute majority of the individuals in the cohort were of self-reported European ancestry. Weights from the 1000 Genomes project principal component analysis (PCA) were obtained from [9], and projected on the TILI samples. Thirteen outliers (Figure 3.1), based on the first four principal components were removed. The remaining 731 samples were well-mixed, forming one cluster.



(a)

(b)

Figure 3.1 Projection of weights derived from 1000 Genome Project principal component analysis onto 778 samples from the PRED4 TILI cohort. The absolute majority of the samples are clustered together (731, orange). Thirteen outliers (blue) were removed after manually inspecting the first four principal components. Exclusion thresholds are shown with dashed blue lines.

Within-cohort Hardy-Weinberg-normalised PCA was conducted to identify outlier samples (Figure 3.2). 10 principal components were calculated using the LD-pruned variant subset (described above). Regions of the genome with long-range LD were excluded, as described in [149]. The eigenvalues were small, suggesting that there is no substantial variance across the genotyping batches (eigenval$_{PC1}$ = 1.42, eigenval$_{PC10}$ = 1.24). Thirty-three outliers, based on the first four principal components were removed. The first five principal components were used as covariates during the association tests.

The final sample set contained 698 samples – 207 TILI cases and 491 matched non-TILI controls.

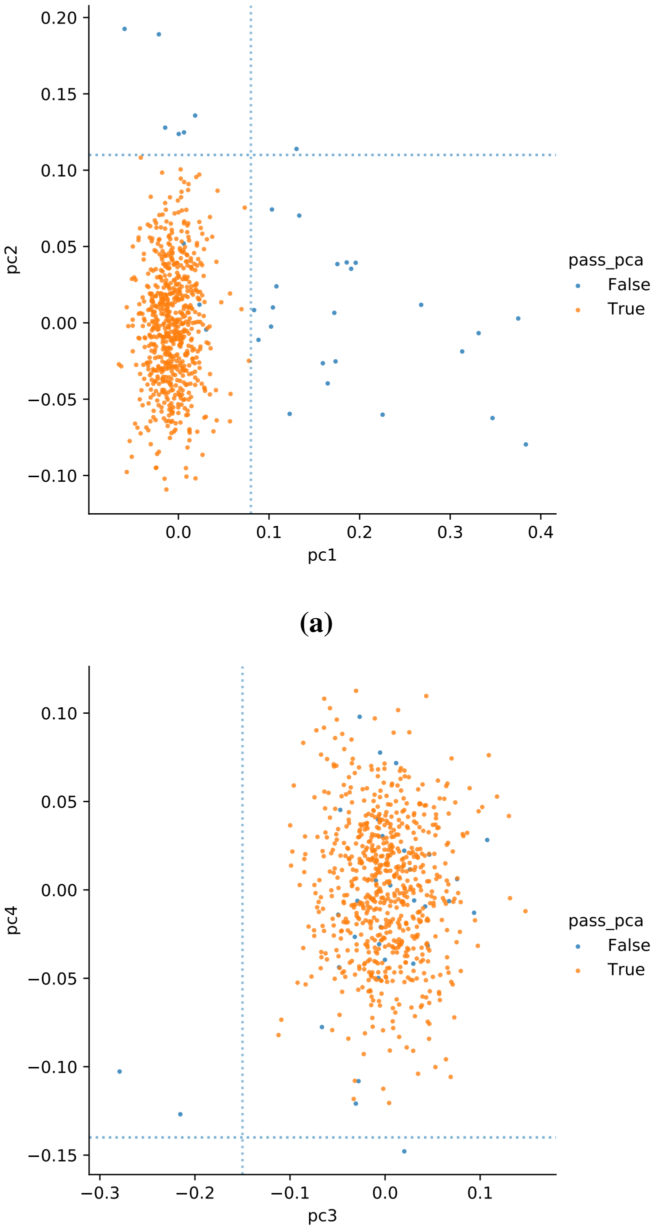
Variant QC

Rare and low-quality variants were removed prior to the association testing. The following filters were applied to the dataset: MAF > 1%, missingness < 5%, P$_{hwe}$ in controls > 10^{-6}. The final dataset contained 226,337 SNPs. At this stage, I did not impute the dataset to any of the reference panels. Primarily, this was due to the low density genotyping and the strong exonic bias of the Broad G4L chip, and partially due to time constraints. A decision was made to impute the dataset only if any significant associations were uncovered (the entire GWAS signal ‘peak’ is unlikely to contain only imputed SNPs).

3.2.2 Statistical testing

For the case-control association tests, logistic regression was carried out using the Wald test. Sex, disease type (CD versus UC), and the first five principal components were used as the covariates.

After performing the draft case-control association analysis, I identified several spurious associations (single variant with no variants in LD with similar p-values). Upon further inspection, these associations were driven by samples from one of the cohorts. To correct for these batch effects, I included a series of case-case and control-control tests (e.g., cases from `GWAS3` versus cases from other cohorts). The variant was only considered significant if the p-value in neither of the batch-effect tests exceeded α=1×10^{-3}.



(a)

(b)

Figure 3.2 Principal component analysis of the PRED4 TILI cohort. Thirty-three outliers (blue) were removed after manually inspecting the first four principal components. Exclusion thresholds are shown with dashed blue lines.

Sample and variant QC, PCA, and association testing was performed using the Hail 0.2 framework [77].

3.2.3 Power calculation

Using the methodology described by Johnson and Abecasis [88], power to detect single-variant associations for TILI was calculated. The following parameters were used: significance threshold α=5×10^{-8}, trait prevalence 10% [16], additive model, 207 cases and 491 controls. Considering that the controls were screened for TILI, the genotype relative risk ratios can be used as the odds ratio estimates. Scenarios where statistical power exceeded 0.8 were considered as 'well powered'.

The power calculation (Figure 3.3) suggests that at the current sample size, I was poorly powered to detect associations with relative risks below 2.5 for variants of all minor allele frequencies. For rare variants (frequency of 1%), the risk would have to exceed 10. For common variants with MAF = 10% and above, I only had power to detect variants with an odds ratio of 3 and above. It should be noted that power calculations tend to overestimate the power, as they do not take into account the genotyping errors, cryptic population structure, and batch effects that have a negative impact on the ability to detect a true association.

3.3 Results

3.3.1 Case-control analysis

The genome-wide case-control analysis did not identify any robust associations for thiopurine-induced liver injury. The QQ-plot did not indicate a major inflation of the p-values. The genetic inflation factor was low (λ = 1.01). A single variant that passed the genome-wide significance threshold – rs10935807 – appears to be significantly associated with TILI (OR = 2.32; 95% CI 2.07 to 2.57; 6.61×10^{-11}). In the GTEx dataset, the variant is significantly associated with the expression of the *EIF2A* gene in 18 tissues (not including liver). No formal colacolisation analysis was performed. Eukaryotic translation initiation factor 2-alpha kinase 3 (*EIF2AK3*), often referred to as protein kinase R (PKR)-like endoplasmic reticulum kinase (*PERK*), is known to be involved in phosphorylation of *EIF2A* [160]. Hopper et al.

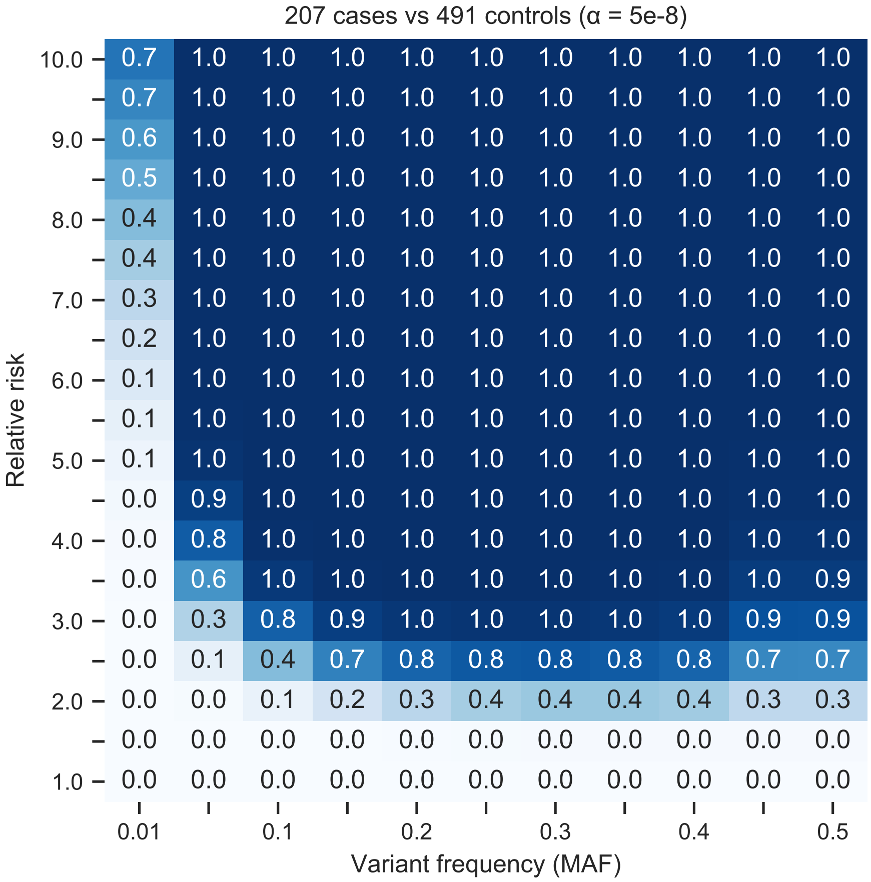


Figure 3.3 Power to detect single-variant associations for PRED4 TILI cohort. X axis – relative risk of the variant. Y axis – variant frequency. Individual segments on the heatmap – power to detect the association at the genome-wide significance level. Methodology as described in [88]. Original code translated from JavaScript to Python in order to run power calculations for larger sets of parameters simultaneously. Parameters used: significance threshold α=5×10^{-8}, trait prevalence 10% [16], additive model, 207 cases and 491 controls.

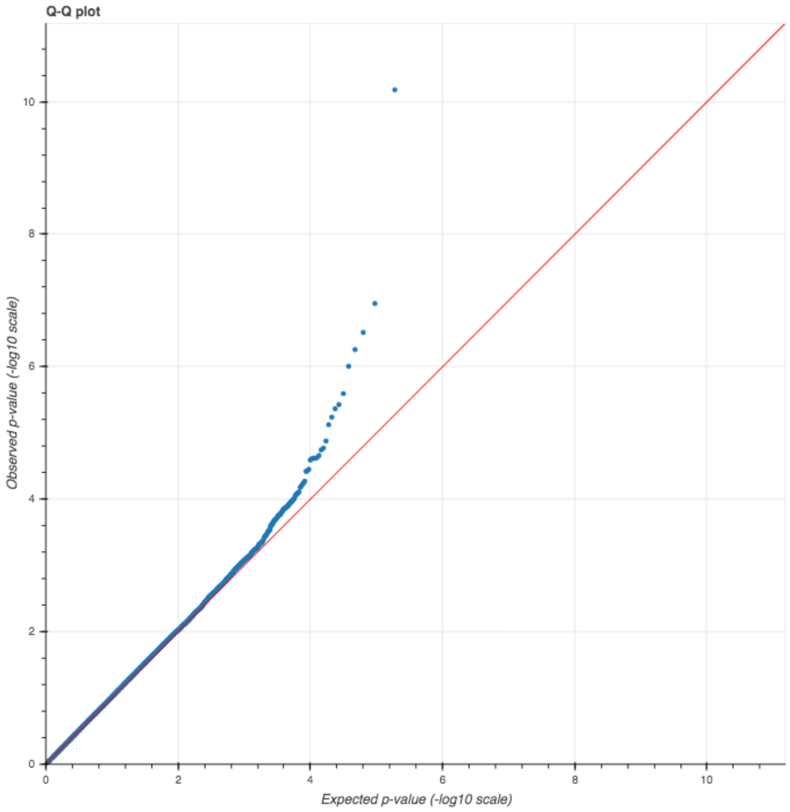


Figure 3.4 Quantile-quantile plot for the case-control analysis of the PRED4 TILI cohort. The inflation factor λ is 1.01, suggesting a good fit to the uniform distribution.

[81] have demonstrated that azathioprine induces autophagy, partially via stimulation of the unfolded protein response (UPR) sensor PERK. However, despite the tentative biologic explanation, I do not believe that the association is truly robust.

While I could not identify any obvious issues with the genotyping quality, there are several properties that make me cautious about declaring this association to be true. Firstly, as visible on the Manhattan plot, there are no other SNPs with a p-value close to rs10935807. The SNP with the highest R^2 with rs10935807 (out of those currently genotyped) – rs9883613 (R^2 = 0.6352, D' = 0.9728) demonstrates no evidence of association (P = 0.31). In addition, amongst the two biggest case-containing batches, the variant has a substantially different minor allele frequency (`broad1` – 0.45, `newgeno` – 0.60). The frequency in the biggest control batches closely matches that reported in gnomAD for individuals of the European ancestry (∼0.35). As it currently stands, I am hesitant to claim that the variant is associated with TILI. The variant could be potentially re-imputed. Unfortunately, the G4L chip does not include any of the five variants that are in high LD (> 0.95) with the rs10935807, making

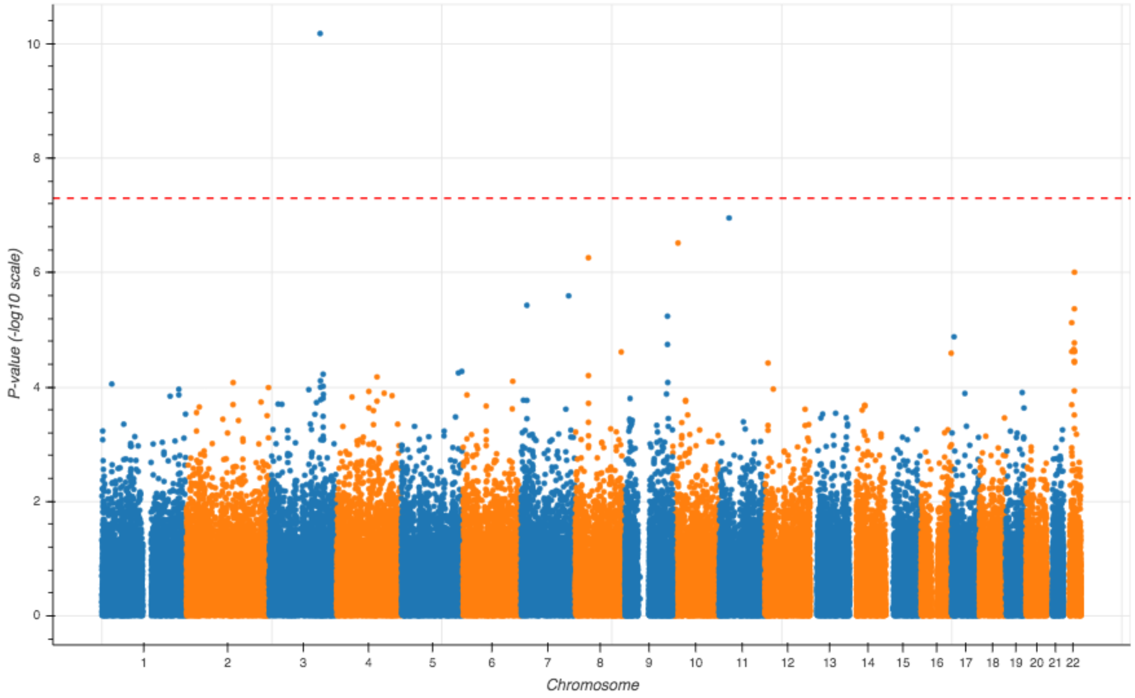


Figure 3.5 Manhattan plot for case-control analysis of the PRED4 TILI cohort. Only one variant (rs10935807) reached the genome-wide significance level (OR = 2.32; 95% CI 2.07 to 2.57; 6.61×10^{-11}). The supporting text describes the arguments why it requires further investigation before conclusive statements can be made with regard to its association with TILI.

high-quality imputation unlikely. Alternatively, the cases and controls genotyped exclusively with the G4L chip can be re-genotyped on the CoreExome chip in order to match the rest of the cohort. Potentially, one could try validating this tentative association via replication.

3.3.2 rs2476601 in *PTPN22* does not appear to be associated with TILI

I have attempted to replicate the result reported by Cirulli et al. [39] (see the detailed description in the introduction). They describe a finding that a missense variant in *PTPN22* is associated with idiosyncratic drug-induced liver injury. The association is largely driven by liver injury caused by amoxicillin-clavulanic acid combination therapy, though a variety of other therapies have an effect size in the same direction. The cohort included 10 TILI cases caused by mercaptopurine which, when analysed alone, had a similar effect size to the overall association but was not significant (OR=1.72; 95% CI, 0.5 to 5.97; P=0.39). Convincingly replicating the DILI association in a TILI cohort would be of great interest and could be considered the first robust genetic association for TILI.

The rs2476601 SNP was only genotyped in the CoreExome subset of the PRED4 TILI cohort – 152 cases and 364 controls. The particular variant was well genotyped: 100% call rate, P_{hwe}=0.17, MAF=0.91 comparable to MAF=0.90 reported in gnomAD for Northwestern Europeans. The case-control analysis was repeated on this subset containing rs2476601.

No evidence of replication was uncovered (OR=0.88; 95% CI, 0.41 to 1.35; P=0.60). It should be noted, that I did not have the complete statistical power to replicate the association: assuming $\alpha = 0.05$ (replication-level significance) and OR=1.72 (the reported odds ratio for mercaptopurine), the power was 0.76; assuming $\alpha = 0.05$ and OR=1.44 (the pooled OR) it was 0.44.

The work of Cirulli et al. is an important step towards understanding the genetic determinants of drug-induced liver injury. It demonstrates consistent effect in DILI caused by several drugs (namely, antibiotics and antifungals). However, at this stage there is no evidence that rs2476601 is associated with liver injury caused by thiopurines.

3.4 Discussion

In this chapter, I have described the first genome-wide association study of thiopurine-induced liver injury. Unfortunately, at this stage, it did not result in any robust associations. I believe there are several potential reasons for this.

The success of GWAS as a technique for studying complex disease genetics comes down to the consistent application of rigorous statistical approaches to ever-growing sample sizes. The hypothesis-free nature of GWAS requires applying a stringent genome-wide significance threshold for p-values, correcting for inevitable multiple testing (typically, α=5×10^{-8}, see Section 1.3.2 for a more extensive discussion). Therefore, GWAS study cohorts need to be sufficiently big to detect truly associated variants. In this chapter, I have analysed a cohort of 207 cases and 491 controls – well below the sample size that is expected for modern complex trait GWASs, which often exceed hundreds of thousands of cases and controls.

The small sample size is not unusual for pharmacogenetic studies. Cholestatic TILI occurs in 1 in 1,000 thiopurine patients who are treated for IBD, a disease that occurs in approximately 0.5–0.8% individuals in the UK. The rarity of the condition makes collecting even a few hundred cases a challenging task. Past pharmacogenetic GWASs have yielded results at modest sample sizes, due to the high odds ratios of the uncovered variants (e.g., the coding *NUDT15* variant increasing the risk of thiopurine-induced myelosuppression [OR=27.3] [186]).

An assumption often made when designing pharmacogenetic studies is that, in contrast to the majority of complex diseases, adverse drug reactions might be associated with common genetic variants (say, MAF>5%) and have a high effect size, as they have not gone through the purifying selection due to the recency of the therapy's arrival. This assumption, however, is only partially correct: therapies with severe side effects that are strongly associated with common genetic variants are not expected to pass stages II and III of clinical trials due to safety concerns. I have performed a case-control power calculation (Figure 3.3), suggesting that the study had 80% power to detect variants with relative risk of 2.5 and above for all minor allele frequencies. It is entirely possible that thiopurine-induced liver damage is a polygenic trait that is not associated with such high-effect size SNP variants.

Another limitation of the study is the heterogeneity of the genotyping arrays, resulting in only around 226,000 markers being tested for associations (discussed in Methods). The number of tested markers can be potentially increased via imputation. However, as discussed in the Methods section, the sparsity and the exonic bias of the G4L array is likely to be detrimental to the overall imputation quality. We are considering including the TILI samples into the ongoing IBD WES study in order to whole-exome sequence them at 60x depth. Exome sequencing will allow us to search for low-frequency variation that may be associated with TILI, but is poorly captured by the current genotyping arrays. Considering the small cohort size, this would be sufficient only for finding associations of a high effect size and it is likely that such variants are within the exonic regions of the genome. However, it is worth noting that, based on the power calculation, a 1% MAF variant would need to have a relative risk of 9 and above in order to be associated at least at an exome-wide significance level (α=1×10^{-6}).

Finally, I believe that the TILI study can be extended further. The NIHR IBD BioResource project [143] is finalising the recruitment of its first 25,000 IBD patients, who will undergo genotyping. The participants will be given questionnaires that include timings (treatment start and discontinuation), therapy response, and information on adverse drug reactions. Assuming that 75% of those 25,000 patients underwent thiopurine therapy and a 10% incidence of TILI, one would expect 1,875 TILI cases amongst the BioResource patients. The difficulty is that TILI cases might not be encoded as such.

One way to get around the scarcity of the TILI cohorts is to perform a discovery GWAS for all-cause therapy thiopurine failure. The proxy-phenotype can be further improved by including the information on when the patients have ceased the therapy, expecting the hepatocellular TILI cases to stop the therapy within the first 12 weeks of treatment. This approach is unlikely to work for patients with cholestatic TILI since the condition occurs within the first 12 months of the treatment, which is hard to distinguish from discontinuation due to treatment ineffectiveness. Besides, considering their 1 in 1,000 frequency, the number of cholestatic patients in the entire BioResource will be very low. If the analysis results in significant associations, these can be validated in the PRED4 cohort by confirming that a variant is at least nominally significant in TILI and has a similar OR.

Ultimately, this chapter demonstrates that the search for pharmacogenetic associations remains a challenging task, largely limited by the difficulty of assembling the cohorts. The arrival of large biobanks and bioresource services could enable studies that leverage imperfect proxy phenotypes in order to maximise the statistical power at the discovery stage. However, the presence of the clinically validated datasets, like the PRED4 TILI cohort, will remain important for verifying such associations.