Genes and variants in hematopoiesis-related pathways are associated with gemcitabine/carboplatin-induced thrombocytopenia

Niclas Björn, Benjamín Sigurgeirsson, Anna Svedberg, Sailendra Pradhananga, Eva Brandén, Hirsh Koyi, Rolf Lewensohn, Luigi de Petris, Maria Apellániz-Ruiz, Cristina Rodríguez-Antona, Joakim Lundeberg and Henrik Gréen

The self-archived postprint version of this journal article is available at Linköping University Institutional Repository (DiVA):

<http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-162137>

N.B.: When citing this work, cite the original publication.

The original publication is available at www.springerlink.com:

Björn, N., Sigurgeirsson, B., Svedberg, A., Pradhananga, S., Brandén, E., Koyi, H., Lewensohn, R., de Petris, L., Apellániz-Ruiz, M., Rodríguez-Antona, C., Lundeberg, J., Gréen, H., (2019), Genes and variants in hematopoiesis-related pathways are associated with gemcitabine/carboplatin-induced thrombocytopenia, *The Pharmacogenomics Journal*. <https://doi.org/10.1038/s41397-019-0099-8>

Original publication available at:

<https://doi.org/10.1038/s41397-019-0099-8>

Copyright: Springer Nature [academic journals on nature.com] (Hybrid Journals)

<https://www.nature.com/>

Genes and variants in hematopoiesis-related pathways are associated with gemcitabine/carboplatin-induced thrombocytopenia

Niclas Björn¹, Benjamín Sigurgeirsson^{2,3}, Anna Svedberg¹, Sailendra Pradhananga², Eva Brandén^{4,5}, Hirsh Koyi^{4,5}, Rolf Lewensohn⁶, Luigi De Petris⁶, Maria Apellániz-Ruiz⁷, Cristina Rodríguez-Antona⁷, Joakim Lundeberg^{2,8} and Henrik Gréen^{1,2,8,&,*}

¹Clinical Pharmacology, Division of Drug Research, Department of Medical and Health Sciences, Linköping University, Linköping, Sweden. ²Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, Division of Gene Technology, KTH Royal Institute of Technology, Solna, Sweden. ³School of Engineering and Natural Sciences, University of Iceland, Reykjavík, Iceland ⁴Department of Respiratory Medicine, Gävle Hospital, Gävle, Sweden. ⁵Centre for Research and Development, Uppsala University/Region Gävleborg, Gävle, Sweden. ⁶Thoracic Oncology Unit, Tema Cancer, Karolinska University Hospital, and Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden. ⁷Hereditary Endocrine Cancer Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain. ⁸Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Linköping, Sweden.

& JL and HG share the last authorship.

* Corresponding author: Henrik Gréen, Clinical Pharmacology, Division of Drug Research, Department of Medical and Health Sciences, Linköping University, SE-58185 Linköping, Sweden.

E-mail: henrik.green@liu.se, Telephone: +46101031544, Fax: +4613104195

Running title: Genetics and chemotherapy-induced thrombocytopenia

Abstract

Chemotherapy-induced myelosuppression, including thrombocytopenia, is a recurrent problem during cancer treatments that may require dose alterations or cessations that could affect the anti-tumor effect of the treatment. To identify genetic markers associated with treatment-induced thrombocytopenia, we whole-exome sequenced 215 non-small cell lung cancer patients homogeneously treated with gemcitabine/carboplatin. The decrease in platelets (defined as nadir/baseline) was used to assess treatment-induced thrombocytopenia. Association between germline genetic variants and thrombocytopenia was analyzed at single-nucleotide variant (SNV) (based on the optimal false discovery rate, the severity of predicted consequence, and effect), gene, and pathway levels. These analyses identified 130 SNVs and 25 genes associated with thrombocytopenia (P -value <0.002). 23 SNVs were validated in an independent genome-wide association study (GWAS). The top associations include rs34491125 in *JMJD1C* (P -value= 9.07×10^{-5}), the validated variants rs10491684 in *DOCK8* (P -value= 1.95×10^{-4}), rs6118 in *SERPINA5* (P -value= 5.83×10^{-4}) and rs5877 in *SERPINC1* (P -value= 1.07×10^{-3}), and the genes *CAPZA2* (P -value= 4.03×10^{-4}) and *SERPINC1* (P -value= 1.55×10^{-3}). The SNVs in the top-scoring pathway “*Factors involved in megakaryocyte development and platelet production*” (P -value= 3.34×10^{-4}) were used to construct weighted genetic risk scores (wGRS) and logistic regression models that predict thrombocytopenia. The wGRS predict which patients are at high or low toxicity risk levels, for CTCAE (OR=22.35, P -value= 1.55×10^{-8}), and decrease (OR=66.82, P -value= 5.92×10^{-9}). The logistic regression models predict CTCAE grades 3-4 (receiver operator characteristics (ROC) area under the curve (AUC) = 0.79), and large decrease (ROC AUC=0.86). We identified and validated genetic variations within hematopoiesis-related pathways that provide a solid foundation for future studies using genetic markers for predicting chemotherapy-induced thrombocytopenia and personalizing treatments.

Introduction

Lung cancer is both common and lethal. It represents 13% of all newfound cancer cases, causes 19% of all cancer-related deaths, and has a 5-year survival of only 18%.^{1, 2} Classical chemotherapeutic treatment of non-small cell lung cancer (NSCLC) commonly consists of gemcitabine and carboplatin, depending on the result of the primary treatment using either PD-1 inhibitors or targeted therapies. The regime is known to cause severe toxicity that may require postponement or reduction of the treatment dose and, in some cases, even discontinuation of the therapy. Myelosuppression is the most commonly induced adverse drug reaction (ADR), manifesting itself as anemia, leukopenia, neutropenia, and thrombocytopenia.^{3, 4} Severe thrombocytopenia of Common Terminology Criteria for Adverse Events (CTCAE) grade 3-4 is commonly observed in patients in clinical studies,⁵⁻¹⁰ in most cases, reported in 50-70% of the patients. There is a vast inter-individual variability in thrombocytopenia, and with equal doses, some patients exhibit no or mild symptoms, while others experience severe thrombocytopenia even after the first course of treatment.

The underlying mechanisms for induced ADRs in patients treated with chemotherapy are, to date, not fully understood for myelosuppression or thrombocytopenia. Germline genetic variability is thought to affect toxicity, and efforts have been made to find ways to predict myelosuppression in chemotherapy treatments by using candidate gene approaches and genome-wide association studies (GWAS).¹⁰⁻¹⁶ These previous studies, mostly of Asian origin, have investigated either a single treatment regimen in patients with various tumor types or multiple treatment regimens in patients with a single tumor type.

Being able to predict patient risk of thrombocytopenia, and adjust doses and treatments accordingly, would most likely be beneficial for both patient quality of life and response to treatment, and it would reduce hospitalization times and costs related to treatment-induced toxicity.

The purpose of this study was, therefore, to identify the genes and variants underlying the inter-individual variability in gemcitabine/carboplatin-induced thrombocytopenia, with the goal of identifying genetic variations that can be used for the prediction of toxicity. To do this, we conducted whole-exome sequencing of NSCLC patients homogeneously treated with gemcitabine/carboplatin chemotherapy using high-throughput next-generation sequencing. We, subsequently, performed association analyses of thrombocytopenia at the single nucleotide variant (SNV), gene, and pathway levels and validated the results in an independent patient cohort.

Materials and Methods

Study population

Between 2006 and 2008, 215 NSCLC patients at Karolinska University Hospital, Stockholm, Sweden, were included in the study. The regional ethics committee in Stockholm, Sweden, approved the study (DNR-03-413 with amendment 2016/258-32/1), and all patients provided written and oral informed consent as per the Declaration of Helsinki. A subsample of these patients had previously been included in another toxicity study of 32 patients.¹⁰ The 215 participants received at least one cycle of carboplatin (area under the curve (AUC) = 5, at day 1) and gemcitabine (1250 mgm⁻² at day 1 and day 8), which was the standard of care for NSCLC patients at the time and place of the study. Platelet counts were registered at baseline and monitored at day 8, 15, and 21 during the first cycle. The platelet nadir value and the decrease in platelets (defined in Equation 1) were used as toxicity parameters for thrombocytopenia.

$$\text{Equation 1:} \quad \text{decrease} = \text{nadir/baseline}$$

DNA extraction, exome enrichment, and sequencing

DNA was extracted from whole blood using QIAamp DNA mini-kits (VWR International, Stockholm, Sweden) according to the manufacturer's protocol. Target enrichment and library preparation were performed using the Nextera Rapid Capture Exome kit (FC-140-1003, Illumina, San Diego, California, USA) according to the manufacturer's instructions and utilizing Agilent Bravo for automation. All samples were whole-exome sequenced on Illumina HiSeq 2500 v4 (generating 125 base-pair long paired-end reads) at the Science for Life Laboratory, Stockholm, Sweden.

Quality control, alignment, and variant calling

Sequencing reads were quality and adapter trimmed using TrimGalore! (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and cutadapt.¹⁷ Trimmed reads with a Phred quality score < 25 and read-pairs with one read length < 25 were removed. Alignment to the human genome (GRCh37.72, <http://www.ensembl.org/>) was done using Bowtie2.¹⁸ Mapped reads were filtered, using SAMtools,¹⁹ to only utilize primary aligned reads that mapped in proper pairs. Duplicate reads were discarded using Picard Tools (<http://www.picard.sourceforge.net/>). Variant calling was conducted using the Genome Analysis Toolkit (GATK) version 3.1.1. supplied with the Nextera Rapid Capture Exome Targeted Regions Manifest version 1.2 and following their best practice pipeline.²⁰

Post-variant calling quality control and outlier removal

VCFtools²¹ was used to discard variants not labeled as PASS, with genotyping rate < 0.95, with mean coverage < 10 across all samples, and that failed the Hardy-Weinberg test with a P-value of 0.0001. Detection of outliers was conducted using the metrics identity by descent (IBD) and identity by missingness (IBM) in PLINK.²² Three samples were identified as outliers, see Supplementary Figure S1, and deemed to be unreliable and, consequently, were removed from all subsequent analyses.

Association analysis

Phenotype values and transformations

The two toxicity parameters of thrombocytopenia, nadir values and the decrease in the nadir values, were normalized using the natural logarithmic transformation in R version 3.4.1²³ and a rank-based transformation using van der Waerden normal scores implemented in the tRank function in the R-package multic. Supplementary Figure S2 shows the distribution before and after the transformation of nadir values and the decrease. The transformation gave four normal-distributed phenotype values (PVs) for thrombocytopenia that were used for the association analyses: the natural logarithm transformation of nadir values (PV1) and the decrease (PV2), and the rank-based transformation of nadir values (PV3) and the decrease (PV4).

SNV association analysis

Common variants (minor allele frequency (MAF) ≥ 0.01) were included in the SNV association analyses that were carried out individually for the four PVs of thrombocytopenia using an additive genetic model in PLINK²² with the covariates age and gender. To evaluate the different PVs and to estimate a suitable P-value cut-off, 1000 permutations with randomly shuffled PVs were performed. Supplementary Figure S3 contains the results of the permutation tests. It is evident from the results that the decrease parameters PV2 and PV4 had an optimal P-value cut-off of ≤ 0.002 , which was determined to yield the largest difference from the random permutations with the lowest false discovery rate (FDR). The results also show that the nadir values in PV1 and PV3 reflect the random effect and, therefore, these values were not investigated further. All SNVs from the PV2 and PV4 analyses were scored using Combined Annotation-Dependent Depletion (CADD),²⁴ a method that integrates various annotations into a single score for each variant.

Gene-based association analysis

The combined effect of common ($MAF \geq 0.01$) and rare ($MAF < 0.01$) genetic variants within a gene region (exon region in RefSeq GRCh37/hg19 ± 6 base pairs) on thrombocytopenia was evaluated with the optimal sequence kernel association test in the R-package SKAT²⁵ using default settings, assigning equal weight to all variants, and with age and gender as covariates. Genes harboring only one genetic variant were excluded.

Potentially pertinent variants and genes

We used a stringent requirement that SNVs and genes had to surpass the P-value cut-off < 0.002 for both PV2 and PV4 as a measure to reduce the likelihood of false positives. SNVs and genes that met this requirement were considered to infer association and were denoted as potentially pertinent variants (PPVs) and potentially pertinent genes (PPGs).

Validation

Validation cohort

As a validation, SNVs with $MAF \geq 0.01$ and a P-value < 0.002 for the association to thrombocytopenia ($n = 1595$) from an independent GWAS generated in²⁶ (briefly described in Supplementary Material S1 and accompanied with information on the successfully validated variants in Supplementary Table S1) were used.

Validation method

Linkage disequilibrium (LD) between SNVs from the validation cohort and our PPVs was examined using the Ensembl REST API (version 6.3). Pair-wise LD was evaluated for variant pairs closer than 500 kilobases using the European population panels CEU (Utah residents with Northern and Western European ancestry), FIN (Finnish in Finland), GBR (British in England and Scotland), IBS (Iberian populations in Spain) and TSI (Toscani in Italia) from the 1000 Genomes Phase 3 data. A $D' > 0.33$ was considered to indicate LD²⁷ and validate the possible importance of the genetic variants for treatment-induced thrombocytopenia.

Pathway analysis

Over-representation analysis of pathway-based sets was performed using the online tool ConsensusPathDB-human (Release 33, <http://cpdb.molgen.mpg.de/>)²⁸ to find genes enriched in predefined pathways in the ConsensusPathDB-human meta-database containing data from various heterogeneous resources. The tool was supplied with all unique genes represented by PPGs, validated PPVs, or PPVs with a CADD > 20 . The over-representation analysis used the default settings, which consisted of a minimum two-gene overlap and a P-value < 0.01 .

Weighted genetic risk score (wGRS) for thrombocytopenia

We constructed weighted genetic risk scores (wGRS) using the SNVs represented in the top associated pathway to predict the risk of thrombocytopenia in our patients. Beta values for the minor alleles acquired in the SNV association analyses were used as weights for the included SNVs. The wGRS were created by multiplying the beta values by the number of minor alleles (0, 1, or 2) summarized across all model SNVs for each patient. The patients were split into four toxicity risk levels (low, medium-low, medium-high, and high) based on the quantiles of the wGRS. The toxicity risk levels were then plotted against thrombocytopenia as CTCAE grades 0-2 and 3-4 or as the magnitude of the decrease in platelets (small decrease: $\geq 25\%$, medium decrease: 25-10%, or large decrease: $\leq 10\%$). Differences in the distribution of thrombocytopenia between the toxicity risk levels were assessed using two-sided Fisher's exact test in R version 3.4.1.²³

Logistic regression model for thrombocytopenia

Using the SNVs represented in the top associated pathway two separate binomial logistic regression models were constructed, one for CTCAE grades 0-2 versus 3-4, and one for a small decrease versus a large decrease in platelets (as defined in the previous section), using the function glm in R version 3.4.1.²³ The predictive capacity of the models were evaluated with receiver operating characteristics (ROC) and AUC using the R-package ROCR version 1.0-7.²⁹

Results

Patient characteristics

The patient characteristics and thrombocytopenia graded according to the National Cancer Institute's CTCAE (version 4.03) are presented in Table 1. The gender of the patients from the clinical report forms was confirmed genetically using check sex in PLINK.²²

Exome target enrichment, sequencing, and variant calling

The average number of paired reads from the sequencing was 38.7 million. The average mapping yield was 99.2%, and the average exome coverage was 74X. There were about 28000 non-reference variants per sample, and a total of 148148 variants were identified, of which 71374 were common (MAF $\geq$ 0.01).

Association analysis

SNV association analysis

The SNV association analysis identified a total of 130 PPVs (SNVs with P-values < 0.002 for both PV2 and PV4) associated with thrombocytopenia. Detailed information on all the PPVs is listed in Supplementary Table S2. Of the PPVs, 97 had negative beta values, meaning that the minor allele is associated with higher toxicity. Interestingly, 50 of the associated variants resided on chromosome 6. In total, the PPVs were represented by 103 unique genes. Of all PPVs, 11 had CADD scores above > 20 , indicating that they can be predicted to be among the top 1 % most deleterious variants in the human genome. Of these 11 variants, rs6907580 (minimal P-value = 6.41×10^{-4} , CADD = 35.00), a stop-gain variant in *GPRC6A*, and rs34491125 (minimal P-value = 9.07×10^{-5} , CADD = 22.10), a missense variant in *JMJD1C*, represent the variants with the highest CADD score and the lowest P-value, respectively.

Gene-based association analysis

The gene-based analysis associated 25 PPGs (genes with P-values < 0.002 for both PV2 and PV4) with thrombocytopenia, listed in Supplementary Table S3. Of the PPGs, 18 were also represented by PPVs in the same genes, of which *CHM*, *OR2B2*, and *PPP1R18* included the top CADD scoring variants.

Validation

By using the 1595 SNVs in the validation data, we found 112 pairs of PPVs and SNVs that were within 500 kilobases of each other. LD data from the European populations in 1000 Genomes Phase 3 data was available for 48 of these pairs (note that European LD data is missing for the remaining 64 pairs). The 48 pairs represented 23 unique PPVs that could be validated, meaning that the validation SNVs and PPVs indicated an LD with $D' > 0.33$. All validated PPVs and their corresponding validation SNV (or SNVs, as some PPVs were validated by multiple variants) are listed in Table 2. The validated PPVs with the two highest CADD scores were rs6118 (missense, minimal P-value = 5.83×10^{-4} , CADD = 22.30), validated by rs3790036 (Hazard ratio (HR) = 1.91), and rs10491684 (synonymous, minimal P-value = 1.95×10^{-4} , CADD = 11.71), validated by rs7025610 (HR = 2.35), located in *SERPINA5* and *DOCK8*, respectively. Further, the three PPGs, *CYP2C8*, *PPP1R18*, and *SERPINC1* also included validated variants

Pathway analysis

The pathway analysis used 45 genes as input (all unique genes represented by PPGs, validated PPVs or PPVs with a CADD > 20) and identified 14 enriched pathways, listed in Table 3. Of the pathways, two were considered very relevant for the investigated phenotype, gemcitabine/carboplatin-induced thrombocytopenia. These two were also the pathways with the lowest associated P-values. They are briefly presented here. The “Hemostasis” pathway (P-value = 1.27×10^{-3}) was identified to be enriched via the association to thrombocytopenia of seven genes; *CAPZA2*, *DGKD*, *DOCK8*, *JMJD1C*, *KIF6*, *SERPINA5*, and *SERPINC1*. Even more interesting, a subpathway to “Hemostasis” called “Factors involved in megakaryocyte development and platelet production” (P-value = 3.34×10^{-4}) was enriched via the genes *CAPZA2*, *DOCK8*, *KIF6*, and *JMJD1C*. These genes included genetic variants with high CADD scores (among the variants in Supplementary Table S2), validated variants (Table 2), and genes from the gene-based association analysis (Supplementary Table S3).

Risk prediction of thrombocytopenia

wGRS

Next, we constructed wGRSs for each patient using the 17 SNVs covered by the genes *CAPZA2*, *DOCK8*, *KIF6*, and *JMJD1C* represented in the pathway “Factors involved in megakaryocyte development and platelet production”. Supplementary Table S4 lists the 17 SNVs together with the beta values and the minor alleles used for calculating the wGRS. The patients were split into the toxicity risk levels low, medium-low, medium-high, and high based on the quantiles of the wGRS. Figure 1A shows the distribution of CTCAE grades 0-2 and 3-4 in the toxicity risk levels, and Figure 1B shows the distribution of the magnitude of decrease in the toxicity risk levels. Fisher’s exact test showed that the wGRS can classify patients accurately to low and high toxicity risk levels (Table 4). Patients in the high toxicity risk level are more likely than patients in the low toxicity risk level to experience CTCAE grades 3-4 (odds ratio (OR) = 22.35, P-value = 1.55×10^{-8}) or a large decrease to $\leq 10\%$ of the baseline value (OR = 66.82, P-value = 5.92×10^{-9}). This indicates that these 17 SNVs can be used in wGRS models to predict risk of severe thrombocytopenia, although further validation is needed.

Logistic regression

Using the same 17 SNVs, as for the wGRSs above, logistic regression models for predicting CTCAE grades 3-4 and a large decrease of platelets were implemented. Figure 2A shows the thrombocytopenia CTCAE model predictions of CTCAE grades 3-4 and Figure 2B shows the associated ROC curve (AUC = 0.79). The decrease model predictions of large decrease are shown in Figure 2C and the associated ROC curve (AUC = 0.86) is shown in Figure 2D. The coefficients that were used in the two logistic regression models, determined by the function glm, are listed in Supplementary Table S5.

Overview

An overview of the statistical approach and the most relevant associated pathways for gemcitabine/carboplatin-induced thrombocytopenia are shown in Figure 3. It can be seen how the results come together in pathways that include genes and SNVs consistent over many analyses that were partly validated and of high CADD scores. These results indicate that SNVs can be used for risk prediction of thrombocytopenia.

Discussion

In this study, exome variants from 212 (after quality control) NSCLC patients undergoing gemcitabine/carboplatin treatment were analyzed for genetic association with chemotherapy-induced thrombocytopenia. The main strengths of the study are the uniform NSCLC patient population, the use

of a single treatment regimen of gemcitabine/carboplatin, and the sample size, which, to the best of our knowledge, is the most extensive study using whole-exome sequencing to investigate drug-induced thrombocytopenia as its primary outcome.

The study found 130 SNVs, of which 23 were validated (43% of those that could be validated), 25 genes, and 14 enriched pathways associated with thrombocytopenia. By using the analysis strategy presented, we could concentrate on the associations consistent over many analyses that were validated, and that were in pathways relevant to the investigated phenotype thrombocytopenia (for an overview see Figure 3). These results could be used to construct wGRS models and logistic regression models that can predict patients risk and probability of CTCAE grades 3-4 and a large decrease in platelets. However, other individual SNV, genes, and pathways could be pertinent to treatment-induced thrombocytopenia, therefore, all PPVs, PPGs, and pathways are reported in the associated supplements.

Many associations on chromosome 6 (50 SNVs, 14 genes and 13 validated variants) were found. These included the gene *PPRP18* with the validated high CADD scoring (32.00) variant rs9262143, and the gene *KIF6*, which is among the genes in the relevant pathways. This finding suggests the importance of genetic regions on chromosome 6 for thrombocytopenia.

The primary findings of the study are the enriched pathways that clearly anchor back to thrombocytopenia. First, the broad “*Hemostasis*” pathway in which we found associations to thrombocytopenia for the following genes *CAPZA2*, *DGKD*, *DOCK8*, *JMJD1C*, *KIF6*, *SERPINA*, and *SERPINC1*, which are all potentially interesting for further investigation. The variant rs6118 in *SERPINA5* had a high CADD score (22.30) and was validated in the GWAS dataset, which added to its relevance. *SERPINC1* was associated through five variants in the gene-based analysis, of which rs5877 is a synonymous variant that was also among the validated PPVs found. *SERPINC1* deficiency is known to cause thrombophilia,^{30,31} a condition in which blood has an increased tendency to form clots, which indicates that the gene has major effects in the body. This indicates that variation in or near *SERPINC1* could also affect thrombocytopenia. The four genes *JMJD1C*, *CAPZA2*, *DOCK8*, and *KIF6* in the “*Hemostasis*” pathway are of particular interest as they are all also involved in the top enriched pathway “*Factors involved in megakaryocyte development and platelet production*”. This pathway seems to be highly involved in the thrombocytopenia investigated in this study. It included the PPV rs34491125 in *JMJD1C* a missense variant with a high CADD score (22.10), indicating a deleterious effect. The gene *JMJD1C* is known to have a function in the formation of platelets, both in model organisms and in humans. In zebrafish, silencing *JMJD1C* leads to failed erythropoiesis and thrombopoiesis, and in humans, the variants rs10761731 and rs7075195 in *JMJD1C* have been correlated to platelet count and mean platelet volume, respectively.³² The mean platelet volume has previously been correlated to the *JMJD1C* variants rs10761741³³ and rs4379723.³⁴ Interestingly, after assessing the pairwise LD (using the same approach as in the validation process) of these four SNVs and the PPV rs34491125 in *JMJD1C*, we found that rs34491125 is in LD with rs10761731 ($r^2 = 0.05$, $D' = 1.00$, CEU), rs7075195 ($r^2 = 0.05$, $D' = 1.00$, CEU), and rs10761741 ($r^2 = 0.05$, $D' = 1.00$, CEU), however, no European LD data was available for rs4379723. The region around and including *JMJD1C* has also been reported in relation to mean platelet volume.³⁵ In addition, the gene has been shown to affect the proliferation of megakaryocytes in mice.³⁶ These findings strongly indicate the possible relevance for *JMJD1C* in relation to thrombocytopenia in our study. The variant rs10813766 in *DOCK8* has previously been correlated to mean platelet volume in humans.³² The PPV reported in our material, rs10491684, is in LD with rs10813766 ($r^2 = 0.05$, $D' = 1.00$, GBR). Further, rs10491684 was also validated with the independent GWAS data. The last two genes in the pathway, *CAPZA2* and *KIF6*, correlated to thrombocytopenia via the gene-based association analysis and the SNV association analysis, respectively. None of their associated variants were validated or had high CADD scores, however, because of the effects of the other genes within the same pathway, we still

consider *CAPZA2* and *KIF6* to be potentially important. By combining the 17 SNVs represented by the four genes in the pathway “*Factors involved in megakaryocyte development and platelet production*”, we could show that a wGRS model can predict which patients are at high or low toxicity risk levels, at least in this cohort. The 17 SNVs predictive validity was strengthened by showing how logistic regression models can predict CTCAE grade 3-4 and a large decrease in platelets with an AUC of 79% and 86%, respectively. This gives sound credibility to the findings and adds to their potential clinical relevance. The results of “*Hemostasis*” and primarily “*Factors involved in megakaryocyte development and platelet production*” suggest that genetic variation within these pathways and genes, especially *JMJD1C* and *DOCK8*, might not only affect platelet status but could also have an important underlying effect on gemcitabine/carboplatin-induced thrombocytopenia.

The findings of this study have no overlaps with the previous extreme toxicity study that included 32 of these patients.¹⁰ There are probably multiple reasons for this; in the present study, we used the decrease and not the nadir, and we included more patients along with patients with intermediate toxicity (not only the extremes).

Initial association analyses showed that the investigated effects would be inadequate to demonstrate exome-wide statistical significance after correction for multiple testing using the Benjamini-Hochberg adjustment. This phenomenon is not limited to our study alone, it is the case for many studies of similar character.³⁷ Therefore, we evaluated the nadir and decrease parameters using randomly shuffled permutations, which guided us to remove the nadir parameter completely as its effects seemed random. An analysis of the permutations showed that a P-value cut-off of 0.002 for the decrease toxicity parameters would yield the highest specificity in the analysis. One interesting side note is that the lowest P-value is not necessarily the one with the greatest deviation from the random permutations; some false positives seem to have very low P-values. With that in mind, it is understandable why top-hits in previous genome-wide studies have proven difficult to validate. We understand that the approach used in the present study will likely increase the number of false positives, however, the approach should also keep the true positives from the multi-factorial phenotypes within the reported results of the study. As a measure to reduce the number of false positives, we implemented the use of potentially pertinent variants and genes (PPVs and PPGs) where an SNV or gene association had to have a P-value < 0.002 for two PVs (PV2 and PV4) to be considered to infer association. These results were then layered with validation and pathway enrichment to strengthen the credibility of the findings and circumvent the problem of false positives. This was done by anchoring the results to independently reported SNVs in LD with our PPVs and to relevant pathways for gemcitabine/carboplatin-induced thrombocytopenia. Another thing that should be kept in mind is that $D' > 0.33$ was used to indicate LD and, thus, add to the validity of the findings. This might not be optimal. However, r^2 does not give an accurate representation of LD as we had low-frequency variants. D' , on the other hand, is more robust and preferable when investigating low-frequency variants.^{38, 39} However, both r^2 and D' are reported in the validation results, Table 2.

The decrease toxicity parameter represents the magnitude of a change from baseline. From a dose-response and a pharmacological perspective, this is what you would expect to have a correlation to genotype. Interestingly, the decrease parameter had a more pronounced optimum in the permutation tests than the nadir. From a clinical perspective, the nadir value might be of more interest, however, the decrease can easily be established by accounting for the baseline value.

Concluding remarks

We reported on variants and genes associated with the gemcitabine/carboplatin-induced ADR thrombocytopenia. When the results from the SNV and gene associations were layered with pathways

and previously published findings, the results formed evidence for variants and genes possibly important for treatment-induced thrombocytopenia. The main pathways included “*Hemostasis*” and “*Factors involved in megakaryocyte development and platelet production*”, which contain genes harboring the top SNVs found in the association analyses. We were able to validate some of these SNVs along with genes previously linked to platelet/thrombocyte function and formation. The SNVs in the latter pathway could be used to predict patients’ risk/probability of severe thrombocytopenia (CTCAE 3-4) using both wGRSs and logistic regression models, something we hope can be validated in forthcoming studies. These results strongly support further investigation into using genetic markers as predictors for chemotherapy-induced thrombocytopenia; something that should not be limited to non-small cell lung cancer or gemcitabine/carboplatin chemotherapy but preferably also extended to other therapies where thrombocytopenia is an ADR.

Acknowledgements

This work was financially supported by grants from the Swedish Cancer Society (HG), the Swedish Research Council (HG), Linköping University (HG), ALF grants Region Östergötland (HG), the Funds of Radiumhemmet (RL and LDP), Marcus Borgströms stiftelse (HG), and the Spanish Ministry of Economy and Competitiveness [SAF2015-64850-R] (CR-A). The funders had no role in study design, data collection, data analysis, decision to publish, or preparation of the manuscript. We gratefully acknowledge the Science for Life Laboratory (SciLifeLab, Stockholm, Sweden), National Genomics Infrastructure (NGI, Sweden), NBIS (National Bioinformatics Infrastructure Sweden), and UPPMAX (Uppsala Multidisciplinary Center for Advanced Computational Science, Uppsala, Sweden) for providing massive parallel sequencing, computational infrastructure, and support.

Conflict of interest

The authors declare no conflicts of interest.

Data availability statement

The raw sequencing datasets generated and analyzed in this study are not publicly available because this is not permitted according to the ethical approval of the study. However, the datasets are available from the corresponding author upon reasonable request with the appropriate ethical approval.

Supplementary Information

Supplementary information is available at *The Pharmacogenomics Journal’s* website.

References

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA Cancer Journal for Clinicians* 2015; **65**(2): 87-108.
2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. *CA Cancer Journal for Clinicians* 2015; **65**(1): 5-29.
3. Barton-Burke M. Gemcitabine: A pharmacologic and clinical overview. *Cancer Nursing* 1999; **22**(2): 176-183.
4. Calvert AH, Harland SJ, Newell DR, Siddik ZH, Jones AC, McElwain TJ, *et al.* Early clinical studies with cis-diammine-1,1-cyclobutane dicarboxylate platinum II. *Cancer Chemotherapy and Pharmacology* 1982; **9**(3): 140-147.
5. Gronberg BH, Bremnes RM, Flotten O, Amundsen T, Brunsvig PF, Hjelde HH, *et al.* Phase III study by the Norwegian lung cancer study group: pemetrexed plus carboplatin compared with gemcitabine plus carboplatin as first-line chemotherapy in advanced non-small-cell lung cancer. *J Clin Oncol* 2009; **27**(19): 3217-3224.
6. Imamura F, Nishio M, Noro R, Tsuboi M, Ikeda N, Inoue A, *et al.* Randomized phase II study of two schedules of carboplatin and gemcitabine for stage IIIB and IV advanced non-small cell lung cancer (JACCRO LC-01 study). *Chemotherapy* 2011; **57**(4): 357-362.
7. Rudd RM, Gower NH, Spiro SG, Eisen TG, Harper PG, Littler JAH, *et al.* Gemcitabine plus carboplatin versus mitomycin, ifosfamide, and cisplatin in patients with stage IIIB or IV non-small-cell lung cancer: A phase III randomized study of the London Lung Cancer Group. *Journal of Clinical Oncology* 2005; **23**(1): 142-153.
8. Sederholm C, Hillerdal G, Lamberg K, Kolbeck K, Dufmats M, Westberg R, *et al.* Phase III trial of gemcitabine plus carboplatin versus single-agent gemcitabine in the treatment of locally advanced or metastatic non-small-cell lung cancer: the Swedish Lung Cancer Study Group. *J Clin Oncol* 2005; **23**(33): 8380-8388.
9. Zatloukal P, Petruželka L, Zemanová M, Kolek V, Skříčková J, Pešek M, *et al.* Gemcitabine plus cisplatin vs. gemcitabine plus carboplatin in stage IIIB and IV non-small cell lung cancer: A phase III randomized trial. *Lung Cancer* 2003; **41**(3): 321-331.
10. Green H, Hasmats J, Kupersmidt I, Edsgard D, de Petris L, Lewensohn R, *et al.* Using Whole-Exome Sequencing to Identify Genetic Markers for Carboplatin and Gemcitabine-Induced Toxicities. *Clin Cancer Res* 2015.
11. Han B, Gao G, Wu W, Gao Z, Zhao X, Li L, *et al.* Association of ABCC2 polymorphisms with platinum-based chemotherapy response and severe toxicity in non-small cell lung cancer patients. *Lung Cancer* 2011; **72**(2): 238-243.
12. Kiyotani K, Uno S, Mushiroda T, Takahashi A, Kubo M, Mitsuhata N, *et al.* A genome-wide association study identifies four genetic markers for hematological toxicities in cancer patients receiving gemcitabine therapy. *Pharmacogenet Genomics* 2012; **22**(4): 229-235.
13. Qian J, Qu HQ, Yang L, Yin M, Wang Q, Gu S, *et al.* Association between CASP8 and CASP10 polymorphisms and toxicity outcomes with platinum-based chemotherapy in Chinese patients with non-small cell lung cancer. *Oncologist* 2012; **17**(12): 1551-1561.
14. Low SK, Chung S, Takahashi A, Zembutsu H, Mushiroda T, Kubo M, *et al.* Genome-wide association study of chemotherapeutic agent-induced severe neutropenia/leucopenia for patients in Biobank Japan. *Cancer Sci* 2013; **104**(8): 1074-1082.
15. Lamba JK, Fridley BL, Ghosh TM, Yu Q, Mehta G, Gupta P. Genetic variation in platinating agent and taxane pathway genes as predictors of outcome and toxicity in advanced non-small-cell lung cancer. *Pharmacogenomics* 2014; **15**(12): 1565-1574.

16. Cao S, Wang S, Ma H, Tang S, Sun C, Dai J, *et al.* Genome-wide association study of myelosuppression in non-small-cell lung cancer patients with platinum-based chemotherapy. *The pharmacogenomics journal* 2016; **16**(1): 41-46.
17. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* 2011; **17**(1): 10-12.
18. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 2012; **9**(4): 357-359.
19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**(16): 2078-2079.
20. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 2011; **43**(5): 491-501.
21. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, *et al.* The variant call format and VCFtools. *Bioinformatics* 2011; **27**(15): 2156-2158.
22. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 2007; **81**(3): 559-575.
23. RCoreTeam (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
24. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* 2014; **46**(3): 310-315.
25. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. *American Journal of Human Genetics* 2013; **92**(6): 841-853.
26. Leandro-García LJ, Inglada-Pérez L, Pita G, Hjerpe E, Leskelä S, Jara C, *et al.* Genome-wide association study identifies ephrin type a receptors implicated in paclitaxel induced peripheral sensory neuropathy. *Journal of Medical Genetics* 2013; **50**(9): 599-605.
27. Strachan T, Read AP. Human molecular genetics. 2011; **4 ed.**: 482.
28. Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herwig R. ConsensusPathDB: Toward a more complete picture of cell biology. *Nucleic Acids Research* 2011; **39**(SUPPL. 1): D712-D717.
29. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCr: visualizing classifier performance in R. *Bioinformatics* 2005; **21**(20): 3940-3941.
30. Ding Q, Wang M, Xu G, Ye X, Xi X, Yu T, *et al.* Molecular basis and thrombotic manifestations of antithrombin deficiency in 15 unrelated Chinese patients. *Thromb Res* 2013; **132**(3): 367-373.
31. van Boven HH, Vandenbroucke JP, Briet E, Rosendaal FR. Gene-gene and gene-environment interactions determine risk of thrombosis in families with inherited antithrombin deficiency. *Blood* 1999; **94**(8): 2590-2594.
32. Gieger C, Radhakrishnan A, Cvejic A, Tang W, Porcu E, Pistis G, *et al.* New gene functions in megakaryopoiesis and platelet formation. *Nature* 2011; **480**(7376): 201-208.
33. Eicher JD, Xue L, Ben-Shlomo Y, Beswick AD, Johnson AD. Replication and hematological characterization of human platelet reactivity genetic associations in men from the Caerphilly Prospective Study (CaPS). *J Thromb Thrombolysis* 2016; **41**(2): 343-350.
34. Shameer K, Denny JC, Ding K, Jouni H, Crosslin DR, de Andrade M, *et al.* A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. *Hum Genet* 2014; **133**(1): 95-109.

35. Soranzo N, Spector TD, Mangino M, Kuhnel B, Rendon A, Teumer A, *et al.* A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat Genet* 2009; **41**(11): 1182-1190.
36. Kitajima K, Kojima M, Kondo S, Takeuchi T. A role of jumonji gene in proliferation but not differentiation of megakaryocyte lineage cells. *Exp Hematol* 2001; **29**(4): 507-514.
37. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* 2014; **15**(5): 335-346.
38. Lin CY, Xing G, Ku HC, Elston RC, Xing C. Enhancing the power to detect low-frequency variants in genome-wide screens. *Genetics* 2014; **196**(4): 1293-1302.
39. Slatkin M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* 2008; **9**(6): 477-485.

Figures and figure legends

Figure 1. Weighted genetic risk score (wGRS) for thrombocytopenia.

A) Shows the distributions of Common Terminology Criteria for Adverse Events (CTCAE) grades grouped as grades 0-2 and 3-4 at the different toxicity risk levels (low, medium-low, medium-high, and high) that patients were predicted to using wGRS. B) Shows the distributions of the platelet decrease (small, medium, or large) at the different toxicity risk levels (low, medium-low, medium-high, and high) that patients were predicted to using wGRS.

Figure 2. Logistic regression models for thrombocytopenia.

A) Predicted probability of thrombocytopenia CTCAE grades 3-4 for the patients separated on registered CTCAE grades. B) Receiver operator characteristics (ROC) curve of the thrombocytopenia CTCAE model predictions. C) Predicted probability of a large decrease of platelets ($\leq 10\%$ of the baseline value) for the patients separated on if they experienced a small ($> 25\%$) or a large decrease. D) ROC curve of the platelet decrease model predictions.

Figure 3. Overview of the statistical methods and their output.

The associations found 130 SNVs (PPVs) and 25 genes (PPGs). The variants were scored using CADD (Combined Annotation-Dependent Depletion), and 23 variants were validated using independent data. Using all unique genes (n = 45) represented by PPGs, validated PPVs, or PPVs with a CADD score > 20, we found 14 enriched pathways. The two top pathways are of primary interest and showed high relevance to the investigated phenotype, thrombocytopenia. The genes in bold include SNVs with high CADD scores or validated variants. Lastly, the SNVs included in the most relevant pathway were used to construct weighted genetic risk scores (wGRS) and logistic regression models for predicting the risk of thrombocytopenia.

Tables and table legends

Table 1. Patient Characteristics.

Baseline characteristics of patients and thrombocytopenia during the first cycle graded according to the National Cancer Institute's Common Terminology Criteria for Adverse Events (CTCAE) version 4.03.

Gender, N (%)			
Male	100	(47.2%)	
Female	112	(52.8%)	
Age, in years, median (range)			
All	64	(45-82)	
Male	66.5	(45-82)	
Female	63.5	(47-80)	
Clinical stage, N (%)			
Stage I	40	(18.9%)	
Stage II	28	(13.2%)	
Stage III	63	(29.7%)	
Stage IV	79	(37.3%)	
not specified	2	(0.9%)	
Histological classifications, N (%)			
Adenocarcinoma (AC)	132	(62.3%)	
Squamous cell carcinomas (SCC)	40	(18.9%)	
Non-small cell lung cancer (NSCLC)	29	(13.7%)	
Large cell carcinoma (LLC)	10	(4.7%)	
Other	1	(0.5%)	
Smoking history, N (%)			
Current	92	(43.4%)	
Former	99	(46.7%)	
Never	21	(9.9%)	
Thrombocytopenia, CTCAE grade, N (%)			
0	43	(20.3%)	
1	50	(23.6%)	
2	44	(20.8%)	
3	44	(20.8%)	
4	31	(14.6%)	
5	0	(0.0%)	

1 **Table 2. Validated PPVs for thrombocytopenia.**
2 Validated PPVs for thrombocytopenia and their respective validation SNV(s) along with r^2 and D' values. A $D' > 0.33$ was considered to indicate linkage
3 disequilibrium.

PPV							Validation data					CEU		FIN		IBS		TSI		GBR	
rsID	Minor allele	Gene name	Consequence	CADD	Min P-value	MAF	rsID	Minor allele	Hazard ratio	P-value	MAF	r^2	d'	r^2	d'	r^2	d'	r^2	d'	r^2	d'
rs9262132	A	C6orf136	upstream	9.61	6.07E-04	0.11	rs915664	G	0.48	7.68E-04	0.33	0.06	1.00	NA	NA	NA	NA	NA	NA	NA	NA
rs2233980	A	C6orf15	synonymous	0.27	7.58E-04	0.11	rs915664	G	0.48	7.68E-04	0.33	0.06	1.00	NA	NA	NA	NA	NA	NA	NA	NA
							rs2596531	G	0.54	1.84E-03	0.39	0.08	0.85	0.10	1.00	0.14	0.87	0.11	0.77	0.10	0.85
							rs2844511	A	0.54	1.84E-03	0.39	0.08	0.85	0.10	1.00	0.14	0.87	0.11	0.77	0.10	0.85
							rs2516448	A	0.54	1.84E-03	0.39	0.08	0.85	0.10	1.00	0.14	0.87	0.11	0.77	0.10	0.85
rs3130985	T	CDSN	intron	6.04	3.13E-04	0.12	rs915664	G	0.48	7.68E-04	0.33	0.06	1.00	NA	NA	NA	NA	NA	NA	NA	NA
							rs2596531	G	0.54	1.84E-03	0.39	0.08	0.85	0.10	1.00	0.16	0.88	0.11	0.77	0.10	0.85
							rs2844511	A	0.54	1.84E-03	0.39	0.08	0.85	0.10	1.00	0.16	0.88	0.11	0.77	0.10	0.85
							rs2516448	A	0.54	1.84E-03	0.39	0.08	0.85	0.10	1.00	0.16	0.88	0.11	0.77	0.10	0.85
rs1058932	A	CYP2C8	3 prime UTR	1.28	4.14E-04	0.19	rs11812285	C	2.03	8.86E-05	0.31	0.09	1.00	0.07	0.66	0.08	0.76	0.10	1.00	0.07	0.82
							rs1934975	G	2.01	1.68E-04	0.30	0.09	1.00	0.07	0.68	0.13	1.00	0.10	1.00	0.07	0.83
rs11572078	TA	CYP2C8	splice, intron	0.54	1.09E-04	0.21	rs11812285	C	2.03	8.86E-05	0.31	0.10	1.00	0.09	0.73	0.13	0.82	0.16	1.00	0.07	0.75
							rs1934975	G	2.01	1.68E-04	0.30	0.10	1.00	0.09	0.74	0.17	1.00	0.16	1.00	0.09	0.87
rs10491684	A	DOCK8	synonymous	11.71	1.95E-04	0.08	rs7025610	G	2.35	5.19E-04	0.07	NA	NA	0.16	0.40	NA	NA	NA	NA	NA	NA
rs3094086	A	DPCR1	synonymous	8.28	3.65E-04	0.11	rs915664	G	0.48	7.68E-04	0.33	0.06	1.00	NA	NA	NA	NA	0.06	1.00	0.05	1.00
							rs2596531	G	0.54	1.84E-03	0.39	0.07	0.73	NA	NA	0.19	1.00	0.17	0.83	0.12	0.87
							rs2844511	A	0.54	1.84E-03	0.39	0.07	0.73	NA	NA	0.19	1.00	0.17	0.83	0.12	0.87
							rs2516448	A	0.54	1.84E-03	0.39	0.07	0.73	NA	NA	0.19	1.00	0.17	0.83	0.12	0.87
rs73450548	A	ETS2 *	regulatory	8.83	1.27E-03	0.10	rs2142113	A	0.59	1.42E-03	0.50	0.05	0.64	NA	NA	NA	NA	NA	NA	NA	NA
rs34007703	TG	FUT8 *	regulatory	10.28	2.47E-04	0.04	rs2300871	C	2.57	1.86E-03	0.08	0.48	1.00	0.92	1.00	0.19	0.68	0.39	1.00	0.39	1.00
rs3130907	G	HCP5	non-coding exon	5.27	3.18E-04	0.13	rs2596531	G	0.54	1.84E-03	0.39	0.11	1.00	0.11	1.00	0.18	1.00	0.19	1.00	0.14	1.00
							rs2844511	A	0.54	1.84E-03	0.39	0.11	1.00	0.11	1.00	0.18	1.00	0.19	1.00	0.14	1.00
							rs2516448	A	0.54	1.84E-03	0.39	0.11	1.00	0.11	1.00	0.18	1.00	0.19	1.00	0.14	1.00
rs1049709	C	HLA-C	3 prime UTR	5.41	5.04E-04	0.13	rs915664	G	0.48	7.68E-04	0.33	0.07	1.00	0.05	1.00	NA	NA	NA	NA	NA	NA
							rs2596531	G	0.54	1.84E-03	0.39	0.13	1.00	NA	NA	NA	NA	NA	NA	0.14	1.00
							rs2844511	A	0.54	1.84E-03	0.39	0.13	1.00	NA	NA	NA	NA	NA	NA	0.14	1.00
							rs2516448	A	0.54	1.84E-03	0.39	0.13	1.00	NA	NA	NA	NA	NA	NA	0.14	1.00

rs886424	T	LINC00243	non-coding exon	4.70	2.74E-04	0.10	rs915664	G	0.48	7.68E-04	0.33	0.07	1.00	NA	NA	NA	NA	NA	NA	NA	NA
rs886423	C	LINC00243	non-coding exon	2.52	7.87E-04	0.14	rs915664	G	0.48	7.68E-04	0.33	0.08	1.00	0.05	1.00	0.11	1.00	0.05	1.00	0.05	1.00
rs3115672	T	MSH5	synonymous	11.39	3.43E-04	0.13	rs2596531	G	0.54	1.84E-03	0.39	0.08	1.00	0.11	1.00	0.08	0.70	0.09	0.75	0.12	1.00
							rs2844511	A	0.54	1.84E-03	0.39	0.08	1.00	0.11	1.00	0.08	0.70	0.09	0.75	0.12	1.00
							rs2516448	A	0.54	1.84E-03	0.39	0.08	1.00	0.11	1.00	0.08	0.70	0.09	0.75	0.12	1.00
rs9262143	T	PPP1R18	missense	32.00	1.34E-04	0.11	rs915664	G	0.48	7.68E-04	0.33	0.06	1.00	NA	NA	NA	NA	NA	NA	NA	NA
rs10885	T	PRRC2A	missense	24.20	8.61E-04	0.20	rs2596531	G	0.54	1.84E-03	0.39	0.12	1.00	0.22	0.92	NA	NA	0.06	0.35	0.15	0.81
							rs2844511	A	0.54	1.84E-03	0.39	0.12	1.00	0.22	0.92	NA	NA	0.06	0.35	0.15	0.81
							rs2516448	A	0.54	1.84E-03	0.39	0.12	1.00	0.22	0.92	NA	NA	0.06	0.35	0.15	0.81
rs11229	G	PRRC2A	synonymous	5.59	8.61E-04	0.20	rs2596531	G	0.54	1.84E-03	0.39	0.12	1.00	0.22	0.92	NA	NA	0.06	0.35	0.15	0.81
							rs2844511	A	0.54	1.84E-03	0.39	0.12	1.00	0.22	0.92	NA	NA	0.06	0.35	0.15	0.81
							rs2516448	A	0.54	1.84E-03	0.39	0.12	1.00	0.22	0.92	NA	NA	0.06	0.35	0.15	0.81
rs6118	T	SERPINA5	missense	22.30	5.83E-04	0.11	rs3790036	G	1.91	1.47E-03	0.16	0.07	0.33	NA	NA	NA	NA	NA	NA	NA	NA
rs6113	C	SERPINA5	synonymous	0.01	1.63E-03	0.11	rs3790036	G	1.91	1.47E-03	0.16	0.07	0.33	NA	NA	NA	NA	NA	NA	NA	NA
rs6119	G	SERPINA5	missense	0.00	6.13E-04	0.10	rs3790036	G	1.91	1.47E-03	0.16	0.07	0.33	NA	NA	NA	NA	NA	NA	NA	NA
rs5877	C	SERPINC1	synonymous	0.10	1.07E-03	0.33	rs12079820	G	2.07	9.42E-04	0.40	0.41	0.64	NA	NA	0.28	0.54	0.34	0.59	0.49	0.84
							rs10912773	A	2.07	9.42E-04	0.40	0.41	0.64	NA	NA	0.28	0.54	0.34	0.59	0.49	0.84
rs1800629	A	TNF *	regulatory	1.54	2.92E-04	0.19	rs2596531	G	0.54	1.84E-03	0.39	0.13	0.82	0.10	0.71	0.09	0.47	0.14	0.74	0.10	0.75
							rs2844511	A	0.54	1.84E-03	0.39	0.13	0.82	0.10	0.71	0.09	0.47	0.14	0.74	0.10	0.75
							rs2516448	A	0.54	1.84E-03	0.39	0.13	0.82	0.10	0.71	0.09	0.47	0.14	0.74	0.10	0.75
rs6504649	G	XYLT2	missense	6.18	1.52E-03	0.37	rs2103273	A	2.38	1.51E-04	0.18	NA	NA	0.06	0.55	NA	NA	NA	NA	NA	NA

Note: PPV rsIDs in the first column are sorted according to Gene Name, CADD score and validation SNP P-value. Variants validated by multiple SNVs are within the same horizontal boxes marked with lines. Variants within the same genes are marked with thick border lines. * gene closest to the variant.

Abbreviations: NA, not available; PPV, potentially pertinent variant; CADD, Combined Annotation-Dependent Depletion; SNV, single nucleotide variant; PV, phenotype value; Min P-value, the minimal P-value from the SNV association analysis; MAF, minor allele frequency; CEU, Utah residents with Northern and Western European ancestry; FIN, Finnish in Finland; GBR, British in England and Scotland; IBS, Iberian populations in Spain; TSI, Toscani in Italia; UTR, untranslated region.

10 **Table 3. Enriched Pathways.**

11 All enriched pathways with a P-value < 0.01 and a minimum two-gene overlap using all unique genes (n = 45) represented by PPGs, validated
 12 PPVs, or PPVs with a CADD > 20.

Pathway name	Genes matched to the pathway	Pathway size	P-value	Pathway source	Pathway identification in source
Factors involved in megakaryocyte development and platelet production	KIF6; CAPZA2; DOCK8; JMJD1C	131	3.34E-04	Reactome	R-HSA-983231
Hemostasis	CAPZA2; DOCK8; SERPINA5; SERPINC1; KIF6; JMJD1C; DGKD	668	1.27E-03	Reactome	R-HSA-109582
Intrinsic Pathway of Fibrin Clot Formation	SERPINA5; SERPINC1	22	1.43E-03	Reactome	R-HSA-140837
Common Pathway of Fibrin Clot Formation	SERPINA5; SERPINC1	22	1.43E-03	Reactome	R-HSA-140875
Membrane Trafficking	CAPZA2; KIF6; CHM; CLTCL1; SPTBN5; BIN1	582	3.25E-03	Reactome	R-HSA-199991
Endocytosis - Homo sapiens (human)	HLA-C; CLTCL1; BIN1; CAPZA2	244	3.35E-03	KEGG	path:hsa04144
Allograft rejection - Homo sapiens (human)	HLA-C; TNF	38	4.01E-03	KEGG	path:hsa05330
Vesicle-mediated transport	CAPZA2; KIF6; CHM; CLTCL1; SPTBN5; BIN1	620	4.43E-03	Reactome	R-HSA-5653656
Formation of Fibrin Clot (Clotting Cascade)	SERPINA5; SERPINC1	39	4.45E-03	Reactome	R-HSA-140877
Graft-versus-host disease - Homo sapiens (human)	HLA-C; TNF	41	4.91E-03	KEGG	path:hsa05332
Type I diabetes mellitus - Homo sapiens (human)	HLA-C; TNF	43	5.39E-03	KEGG	path:hsa04940
keratinocyte differentiation	ETS2; TNF	53	8.09E-03	BioCarta	keratinocytepathway
Transport to the Golgi and subsequent modification	CAPZA2; SPTBN5; FUT8	168	8.94E-03	Reactome	R-HSA-948021
Complement and Coagulation Cascades	SERPINA5; SERPINC1	59	9.95E-03	Wikipathways	WP558

13 Abbreviations: PPV, potentially pertinent variant; PPG, potentially pertinent gene; CADD, Combined Annotation-Dependent Depletion.

14 **Table 4. Statistics of the wGRS.**

15 Comparison of the distribution of CTCAE and decrease in the wGRS toxicity risk levels using Fisher's exact test in R.

wGRS		CTCAE					Decrease			CTCAE 0–2 vs. 3–4			Small vs. Large Decrease		
Toxicity risk level	Risk score	0	1	2	3	4	Small	Medium	Large	OR	95% CI	P-value	OR	95% CI	P-value
Low	> -0.2225	16	7	10	2	1	25	10	1	REF	-	-	REF	-	-
Medium low	-0.5267 – -0.2225	15	26	15	12	7	37	30	8	3.69	0.98 – 20.96	4.25E-02	5.31	0.64 – 248.98	1.41E-01
Medium high	-0.8667 – -0.5267	8	10	13	10	7	17	22	9	5.91	1.50 – 34.54	4.33E-03	12.66	1.52 – 600.73	1.09E-02
High	< -0.8667	4	7	6	20	16	10	17	26	22.35	5.82 – 129.50	1.55E-08	66.82	8.83 – 3015.75	5.92E-09
Total:		43	50	44	44	31	89	79	44						

16 Abbreviations: wGRS, weighted genetic risk scores; CTCAE, common terminology criteria for adverse events; OR, odds ratio; CI, confidence
 17 interval; REF, reference.

Supplementary Figure S1

Supplementary Figure S1. Identification of sample outliers by identity by descent (IBD) and identity by missingness (IBM).

As mentioned in Materials and Methods, three samples out of the original 215 were removed from the analysis. After all filtering steps the program PLINK was used to identify outliers; first through identity by descent (IBD) and then through identity by missingness (IBM). IBD identifies how many of the variants are shared between the samples (A). There it is apparent that samples S0580 and S0664 share variants with other samples at a much higher rate than the rest of the cohort. While the exact reasons for this remain unknown, one possible explanation is that these samples were contaminated and/or mixed in some way prior to sequencing. In any case, to not include them in the association analysis, was regarded as the safe approach. IBM clusters the samples based on missing genotypes (B). There it is shown that samples S0328 and S0664 deviate from the other samples. Sample S0328 was, therefore, also regarded as unreliable and removed from the analysis. Thus, 212 of the original 215 samples were used in the subsequent association analysis.

Supplementary Figure S2

Supplementary Figure S2. Visualization of normalization of nadir and decrease using natural logarithmic and rank-based (tRank) transformation.

The top row shows the right-skewed distribution of the unnormalized distribution of nadir (A) and decrease (B). The middle row shows the nadir (C) and the decrease (D) after normalization using the natural logarithmic transformation in R. The bottom row shows the nadir (E) and the decrease (F) after normalization using the rank-based transformation tRank in the R-package multic. Both transformations give the distributions a more Gaussian curve compared to the distribution of the untransformed data. The logarithmic distributions shown in the middle row is what is referred to in the main text as phenotype value 1 (PV1) and phenotype value 2 (PV2) for nadir and decrease, respectively. The rank-based distribution in the bottom row is what is referred to in the main text as phenotype value 3 (PV3) and phenotype value 4 (PV4) for nadir and decrease, respectively.

Supplementary Figure S3

Supplementary Figure S3. Visualization of permutations for PV1-PV4 and choice of P-value cut-off.

Results from the permutations ($n=1000$) using randomly shuffled values for PV1-PV4 comparing the false discovery rate (FDR) between the permutations and the test with the true phenotype values (all done using the SNV association analysis setup adjusted for age and gender in PLINK) at different P-value cut-offs ranging from 0.01 – 0.00001. For the nadir values in PV1 and PV3, we see that the FDR is between 90-100 %, even for low P-values. Therefore, PV1 and PV3 were discarded from the remaining statistical analyses as the results from them would be deemed highly unreliable. For the decrease parameter in PV2 and PV4, however, we see the FDR going down with a lower P-value cut-off until a point where the FDRs goes up sharply again. From this, the P-value cut-off 0.002 (represented by the dotted line) was chosen to be used in the study. This cut-off shows the lowest FDR (roughly 77 %) for PV2 and PV4 combined. The cut-off will include many false positives and at the same time still, include the true positives. Lowering the cut-off more will increase the number of false positives and decrease the number of true positives. An interesting note here is that the lowest P-values tend to be among the worst, meaning that they are likely to reflect a random effect in the data. This is somewhat alarming considering that studies are, naturally, focused on the variants with the lowest P-values in their dataset.

Genes and variants in hematopoiesis-related pathways are associated with gemcitabine/carboplatin-induced thrombocytopenia

Niclas Björn¹, Benjamín Sigurgeirsson^{2,3}, Anna Svedberg¹, Sailendra Pradhananga², Eva Brandén^{4,5}, Hirsh Koyi^{4,5}, Rolf Lewensohn⁶, Luigi De Petris⁶, Maria Apellániz-Ruiz⁷, Cristina Rodríguez-Antona⁷, Joakim Lundeberg^{2,&} and Henrik Gréen^{1,2,8,&,*}

¹ Clinical Pharmacology, Division of Drug Research, Department of Medical and Health Sciences, Linköping University, Linköping, Sweden

² Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, Division of Gene Technology, KTH Royal Institute of Technology, Solna, Sweden

³ School of Engineering and Natural Sciences, University of Iceland, Reykjavík, Iceland

⁴ Department of Respiratory Medicine, Gävle Hospital, Gävle, Sweden

⁵ Centre for Research and Development, Uppsala University/Region Gävleborg, Gävle, Sweden

⁶ Thoracic Oncology Unit, Tema Cancer, Karolinska University Hospital, and Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden

⁷ Hereditary Endocrine Cancer Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain

⁸ Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Linköping, Sweden

& JL and HG share the last authorship.

* Corresponding author E-mail: henrik.green@liu.se

Supplementary Material S1

Independent GWAS data used for validation

Germline DNA samples from 144 cancer patients (among which 89% had ovary or lung malignancies and 94% received paclitaxel/carboplatin chemotherapy as first-line treatment) were genotyped by Leandro-Garcia *et al.*¹. In brief, the Infinium BeadChip Human 660WQuad assay (Illumina, San Diego, California, USA) was used and the GenomeStudio software was applied to extract genotype data. One sample with a call rate <0.95 was excluded and after filtering 559348 single nucleotide variants (SNVs) were annotated and used in the subsequent association analysis.

Patients' blood status was monitored throughout the treatment and the time to first toxic event (CTCAE grade ≥ 1) was used as the toxicity parameter for thrombocytopenia. Associations of SNVs to thrombocytopenia was assessed with Cox regressions using an additive genetic model adjusted for age in PLINK (version 1.07). A hazard ratio > 1 indicates that the minor allele is associated with toxicity.

From this study, we used SNVs with P-values < 0.002 (the same cut-off as we used for our variants) and MAF ≥ 0.01 for the association to thrombocytopenia (n = 1595). The SNVs that we could successfully validate in the presented manuscript are listed in S1 Table.

References

1. Leandro-García LJ, Inglada-Pérez L, Pita G, Hjerpe E, Leskelä S, Jara C, *et al.* Genome-wide association study identifies ephrin type a receptors implicated in paclitaxel induced peripheral sensory neuropathy. *Journal of Medical Genetics* 2013; **50**(9): 599-605.

Supplementary Table S1. Validated SNVs from the independent GWAS dataset.

Chromosome	Position	rsID	Minor allele	MAF	Hazard ratio	P-value
1	174209909	rs12079820	G	0,40	2,07	9,42E-04
1	174343705	rs10912773	A	0,40	2,07	9,42E-04
6	30794617	rs915664	G	0,33	0,48	7,68E-04
6	31387557	rs2596531	G	0,39	0,54	1,84E-03
6	31389784	rs2844511	A	0,39	0,54	1,84E-03
6	31390410	rs2516448	A	0,39	0,54	1,84E-03
9	753093	rs7025610	G	0,07	2,35	5,19E-04
10	96760812	rs11812285	C	0,31	2,03	8,86E-05
10	96769769	rs1934975	G	0,30	2,01	1,68E-04
14	66147694	rs2300871	C	0,08	2,57	1,86E-03
14	94773121	rs3790036	G	0,16	1,91	1,47E-03
17	48290686	rs2103273	A	0,18	2,38	1,51E-04
21	40506615	rs2142113	A	0,50	0,59	1,42E-03

Abbreviations: MAF, minor allele frequency.

Supplementary Table S2. All PPVs (SNVs with P-values < 0.002 for both PV2 and PV4).

Chromosome	Position	rsID	Ref	Alt	Type	Annotation	Consequence	Gene name	CADD	P-value PV2	beta PV2	P-value PV4	beta PV4	MAF	Minor allele	Major allele	Number of homozygous minor allele	Number of heterozygous	Number of homozygous major allele
1	40555274	rs6689837	C	G	SNV	Transcript	intron	PPT1	2,94	1,62E-03	-0,514	1,31E-03	-0,577	0,071	G	C	2	26	184
1	173878862	rs5877	T	C	SNV	Coding Transcript	synonymous	SERPINC1	0,10	1,07E-03	0,298	1,23E-03	0,324	0,328	C	T	24	91	97
1	247886949	rs36075193	T	C	SNV	Transcript	splice, intron,nc	RP11-63487.5	1,06	1,12E-03	0,413	1,40E-03	0,446	0,142	C	T	3	54	155
2	37468837	rs17409514	G	A	SNV	Coding Transcript	synonymous	NDUFAF7	10,63	1,67E-05	-1,575	6,03E-05	-1,620	0,014	A	G	0	6	206
2	170063471	rs35114151	A	G	SNV	Coding Transcript	synonymous	LRP2	1,78	5,50E-04	0,872	1,62E-04	1,045	0,026	G	A	1	9	202
2	170097707	rs17848149	T	G	SNV	Coding Transcript	missense	LRP2	0,76	5,50E-04	0,872	1,62E-04	1,045	0,026	G	T	1	9	202
2	170129416	rs10515931	C	T	SNV	Transcript	intron	LRP2	2,49	9,58E-04	0,866	2,69E-04	1,048	0,024	T	C	1	8	203
2	227853157	rs112368937	G	A	SNV	Transcript	intron	RHBDD1	6,10	5,53E-05	-1,495	1,24E-03	-1,327	0,014	A	G	0	6	206
2	234294634	rs202178942	CGTGTGCGTGT	C	DEL	Regulatory Feature	regulatory	DGKD *	6,74	6,83E-04	-0,761	1,50E-03	-0,784	0,040	C	CGTGTGCGTGT	0	17	195
3	43074430	rs2948619	A	G	SNV	Coding Transcript	synonymous	FAM198A	1,83	3,32E-04	-0,307	3,97E-04	-0,333	0,441	A	G	43	101	68
3	173997153	rs7646919	G	A	SNV	Coding Transcript	synonymous	NLGN1	9,90	1,67E-03	-0,345	1,49E-03	-0,384	0,201	A	G	8	69	135
4	40045025	rs73229741	T	C	SNV	Non Coding Transcript	non coding exon	RP11-333E13.4	1,22	1,09E-03	0,550	7,08E-04	0,627	0,071	C	T	1	28	183
4	123192240	rs45608936	A	C	SNV	Coding Transcript	missense	KIAA1109	19,99	1,50E-03	-0,812	1,93E-03	-0,874	0,031	C	A	0	13	199
5	14741984	rs2288474	T	C	SNV	Coding Transcript	synonymous	ANKH	3,19	1,11E-03	0,775	5,06E-04	0,908	0,035	C	T	0	15	197
5	161522556	rs11135176	C	T	SNV	Coding Transcript	synonymous	GABRG2	13,74	1,17E-03	0,464	1,75E-03	0,492	0,111	T	C	1	45	166
6	25862466	rs1165165	C	T	SNV	Coding Transcript	missense	SLC17A3	8,30	4,76E-04	-0,353	1,79E-03	-0,348	0,231	T	C	12	74	126
6	27327000	rs72839477	C	T	SNV	Non Coding Transcript	non coding exon	ZNF204P	16,39	2,56E-04	-0,561	2,93E-04	-0,612	0,083	T	C	2	31	179
6	27775674	rs200484	A	G	SNV	Coding Transcript	missense	HIST1H2BL	2,68	1,23E-03	-0,450	1,47E-03	-0,488	0,104	G	A	3	38	171
6	27835218	rs17763089	G	A	SNV	Coding Transcript	synonymous	HIST1H1B	6,33	6,76E-04	-0,528	8,21E-04	-0,572	0,087	A	G	2	31	169
6	27839746	rs200956	T	C	SNV	Coding Transcript	synonymous	HIST1H3I	0,01	9,46E-04	-0,418	1,32E-03	-0,447	0,125	C	T	5	43	164
6	27879200	rs34788973	C	A	SNV	Coding Transcript	missense	OR2B2	22,30	2,56E-04	-0,561	2,93E-04	-0,612	0,083	A	C	2	31	179
6	27879982	rs61742093	A	G	SNV	Coding Transcript	missense	OR2B2	13,66	2,56E-04	-0,561	2,93E-04	-0,612	0,083	G	A	2	31	179
6	28228342	rs1679709	A	G	SNV	Coding Transcript	missense	NKAPL	12,58	1,09E-04	-0,476	1,30E-04	-0,518	0,118	A	G	7	36	169
6	28239915	rs11965538	G	A	SNV	Non Coding Transcript	non coding exon	RP5-874C20.3	1,74	9,35E-05	-0,499	1,10E-04	-0,544	0,120	A	G	6	37	162
6	28268824	rs33932084	A	G	SNV	Coding Transcript	missense	PGBD1	3,91	2,62E-04	-0,539	2,59E-04	-0,594	0,092	G	A	2	35	175
6	28297313	rs853678	T	A	SNV	Coding Transcript	missense	ZSCAN31	12,47	9,72E-04	-0,416	1,20E-03	-0,449	0,116	A	T	6	37	169
6	28366151	rs2232423	A	G	SNV	Coding Transcript	missense	ZSCAN12	0,00	5,91E-04	-0,518	6,30E-04	-0,567	0,087	G	A	2	33	177
6	28891176	rs2230683	T	C	SNV	Coding Transcript	synonymous	TRIM27	4,34	5,33E-05	-0,599	6,56E-05	-0,651	0,086	C	T	3	30	176
6	28963248	rs6456880	T	G	SNV	Coding Transcript	missense	ZNF311	0,39	5,24E-04	0,282	8,70E-04	0,299	0,453	G	T	50	92	70
6	29342775	rs3749971	G	A	SNV	Coding Transcript	missense	OR12D3	1,61	5,93E-05	-0,586	1,76E-04	-0,604	0,094	A	G	2	36	174
6	29523957	rs404240	A	G	SNV	Coding Transcript	synonymous	UBD	3,11	2,76E-05	-0,616	1,13E-04	-0,627	0,092	G	A	2	35	175
6	30078275	rs2523989	C	T	SNV	Coding Transcript	missense	TRIM31	0,04	5,67E-04	-0,367	1,18E-03	-0,381	0,184	T	C	10	58	144
6	30078330	rs2239529	C	T	SNV	Coding Transcript	synonymous	TRIM31	12,41	1,01E-03	-0,352	1,90E-03	-0,367	0,179	T	C	10	56	146
6	30080274	rs2517598	G	A	SNV	Coding Transcript	synonymous	TRIM31	0,09	8,79E-04	-0,354	1,75E-03	-0,368	0,186	A	G	10	54	143
6	30080496	rs2844793	G	A	SNV	Coding Transcript	synonymous	TRIM31	9,47	1,01E-03	-0,352	1,90E-03	-0,367	0,179	A	G	10	56	146
6	30611350	rs9262132	C	A	SNV	Intergenic	upstream	C6orf136	9,61	6,07E-04	-0,445	1,80E-03	-0,447	0,111	A	C	5	37	170
6	30652781	rs9262143	C	T	SNV	Coding Transcript	missense	PPP1R18	32,00	1,34E-04	-0,508	3,64E-04	-0,523	0,109	T	C	4	38	170
6	30782002	rs886424	C	T	SNV	Non Coding Transcript	non coding exon	LINC00243	4,70	2,74E-04	-0,496	1,13E-03	-0,490	0,102	T	C	4	35	172
6	30782205	rs886423	G	C	SNV	Non Coding Transcript	non coding exon	LINC00243	2,52	7,87E-04	-0,409	1,98E-03	-0,415	0,142	C	G	5	50	157
6	30919391	rs3094086	G	A	SNV	Coding Transcript	synonymous	DPCCR1	8,28	3,65E-04	-0,482	1,16E-03	-0,484	0,114	A	G	3	42	166
6	31079644	rs2233980	G	A	SNV	Coding Transcript	synonymous	C6orf15	0,27	7,58E-04	-0,439	1,72E-03	-0,450	0,113	A	G	5	37	166
6	31085356	rs3130985	C	T	SNV	Transcript	intron	CDSN	6,04	3,13E-04	-0,457	1,07E-03	-0,458	0,120	T	C	5	41	166
6	31236854	rs1049709	T	C	SNV	Transcript	3 prime UTR	HLA-C	5,41	5,04E-04	-0,428	8,34E-04	-0,453	0,134	C	T	5	47	160
6	31431813	rs3130907	A	G	SNV	Non Coding Transcript	non coding exon	HCP5	5,27	3,18E-04	-0,443	6,43E-04	-0,463	0,125	G	A	7	37	160
6	31543031	rs1800629	G	A	SNV	Regulatory Feature	regulatory	TNF *	1,54	2,92E-04	-0,408	6,63E-04	-0,423	0,186	A	G	6	67	139
6	31603770	rs111229	A	G	SNV	Coding Transcript	synonymous	PRRC2A	5,59	8,61E-04	-0,358	1,92E-03	-0,368	0,198	G	A	9	66	137
6	31604591	rs10885	C	T	SNV	Coding Transcript	missense	PRRC2A	24,20	8,61E-04	-0,358	1,92E-03	-0,368	0,198	T	C	9	66	137
6	31727897	rs3115672	C	T	SNV	Coding Transcript	synonymous	MSH5	11,39	3,43E-04	-0,434	8,16E-04	-0,448	0,126	T	C	7	39	164
6	32029205	rs1150757	G	A	SNV	Coding Transcript	synonymous	TNXB	6,86	1,58E-04	-0,539	7,56E-04	-0,531	0,095	A	G	4	31	171
6	32064726	rs1150752	T	C	SNV	Coding Transcript	missense	TNXB	0,00	5,37E-04	-0,462	1,95E-03	-0,457	0,112	C	T	4	39	167
6	32223632	rs3132966	G	A	SNV	Non Coding Transcript	non coding exon	XXbac-BPG154L12.4	11,37	1,52E-03	-0,311	1,93E-03	-0,334	0,257	A	G	15	79	118
6	32230608	rs9268055	T	C	SNV	Non Coding Transcript	non coding exon	XXbac-BPG154L12.4	0,74	8,28E-04	-0,329	9,11E-04	-0,359	0,264	C	T	15	81	114
6	32261507	rs3749966	T	C	SNV	Coding Transcript	missense	C6orf10	0,00	4,26E-04	-0,344	5,34E-04	-0,372	0,267	C	T	15	83	114
6	32270363	rs7775332	A	T	SNV	Transcript	intron	C6orf10	0,11	3,02E-04	-0,353	3,72E-04	-0,383	0,269	T	A	15	84	113
6	49479750	rs13193063	T	C	SNV	Coding Transcript	missense	GLYATL3	18,12	1,19E-03	-0,300	5,01E-04	-0,354	0,311	C	T	21	90	101
6	56426931	rs9367689	C	T	SNV	Transcript	splice, intron	DST	11,85	1,11E-03	0,292	1,73E-03	0,309	0,340	T	C	26	92	94
6	56482801	rs6459166	C	G	SNV	Transcript	intron	DST	15,46	9,40E-04	0,298	1,81E-03	0,310	0,335	G	C	25	92	95
6	96034895	rs200988008	CTG	C	DEL	Transcript	intron, feature truncation	MANEA	4,46	1,63E-03	0,445	5,92E-04	0,534	0,092	C	CTG	4	31	177
6	117150008	rs6907580	G	A	SNV	Coding Transcript	stop gained	GPRC6A	35,00	1,80E-03	0,513	6,41E-04	0,616	0,075	A	G	1	30	181
6	150005250	rs56246065	A	G	SNV	Coding Transcript	synonymous	LATS1	1,10	4,60E-04	1,004	7,76E-05	1,242	0,024	G	A	0	10	202
6	153365100	rs2295230	A	C	SNV	Coding Transcript	synonymous	RGS17	0,02	5,08E-04	-0,334	1,53E-03	-0,335	0,293	C	A	17	90	105
6	159653330	rs117860656	G	A	SNV	Coding Transcript	missense	FNDC1	7,99	1,65E-03	1,278	1,90E-04	1,660	0,012	A	G	0	5	206
6	170845354	rs9348278	A	G	SNV	Transcript	intron	PSMB1	4,67	8,91E-04	0,284	7,44E-04	0,318	0,460	G	A	46	103	63
6	170862300	rs12717	G	C	SNV	Coding Transcript	missense	PSMB1	0,06	1,27E-03	0,278	6,42E-04	0,324	0,408	C	G	37	99	76
6	170887693	rs3173219	G	C	SNV	Transcript	intron	PDCD2	0,21	8,14E-04	0,286	4,33E-04	0,330	0,408	C	G	38	97	77
7	5540769	rs6953642	A	C	SNV	Coding Transcript	synonymous	FBXL18	0,00	1,43E-03	-0,453	6,81E-04	-0,530	0,097	A	C	3	35	173
7	44148572	rs376965157	CTGAG	C	DEL	Intergenic	upstream	MIR4649	32,00	7,60E-04	-1,369	8,79E-04	-1,489	0,012	C	CTGAG	0	5	207
7	44180652	rs4440539	G	T	SNV	Coding Transcript	synonymous	MYL7	10,35	7,97E-04	0,628	2,68E-04	0,749	0,054	T	G	1	21	190
7	149523195	rs78492471	C	G	SNV	Non Coding Transcript	non coding exon	SSPO	5,29	1,64E-03	-0,470	1,44E-03	-0,524	0,099	G	C	1	40	171
8	144642281	rs146277165	C	T	SNV	Transcript	intron	GSDMD	3,86										

9	452149	rs16909246	G	C	SNV	Intergenic	downstream	RP11-165F24.3	0,60	1,74E-03	-0,465	1,59E-03	-0,516	0,101	C	G	1	41	170
9	84200529	rs2228172	T	C	SNV	Coding Transcript	synonymous	TLE1	8,67	1,94E-04	-0,944	1,58E-03	-0,885	0,031	C	T	0	13	199
9	135763816	rs2073869	C	T	SNV	Coding Transcript	synonymous	C9orf9	6,62	1,22E-03	0,358	8,71E-04	0,406	0,201	T	C	7	71	134
10	6008172	rs2296139	C	T	SNV	Coding Transcript	synonymous	IL15RA	10,86	8,17E-04	-0,474	1,59E-03	-0,493	0,116	T	C	1	47	164
10	15649698	rs9333269	T	G	SNV	Coding Transcript	missense	ITGA8	1,13	1,46E-03	0,478	5,92E-04	0,566	0,097	G	T	1	39	172
10	33214802	rs2298141	A	G	SNV	Coding Transcript	synonymous	ITGB1	5,00	7,44E-04	-0,402	4,48E-04	-0,460	0,156	G	A	5	56	151
10	64937501	rs34491125	G	C	SNV	Coding Transcript	missense	JMJD1C	22,10	9,07E-05	-1,175	8,35E-04	-1,109	0,021	C	G	0	9	203
10	96117799	rs140800381	C	T	SNV	Transcript	intron	NOC3L	0,14	2,90E-05	-1,672	5,90E-04	-1,522	0,012	T	C	0	5	207
10	96796861	rs1058932	G	A	SNV	Transcript	3 prime UTR	CYP2C8	1,28	5,87E-04	0,363	4,14E-04	0,410	0,186	A	G	11	57	144
10	96827119	rs11572078	T	TA	INS	Transcript	splice, intron, feature elongation	CYP2C8	0,54	1,67E-04	0,383	1,09E-04	0,433	0,208	TA	T	13	137	
10	100195130	rs11539873	G	A	SNV	Coding Transcript	synonymous	HP51	4,39	1,45E-03	-0,565	1,29E-03	-0,628	0,068	A	G	0	29	183
10	127504998	rs17153575	C	T	SNV	Transcript	splice, intron	UROS	14,66	2,10E-05	-1,693	9,26E-04	-1,463	0,012	T	C	0	5	207
11	6578792	rs112656142	T	C	SNV	Coding Transcript	missense	DNHD1	0,00	1,51E-03	1,166	2,45E-04	1,478	0,014	C	T	0	6	206
11	6584806	rs73407197	G	A	SNV	Regulatory Feature	regulatory	DNHD1 *	10,63	1,51E-03	1,166	2,45E-04	1,478	0,014	A	G	0	6	206
11	6588256	rs12418976	G	A	SNV	Coding Transcript	synonymous	DNHD1	0,37	1,51E-03	1,166	2,45E-04	1,478	0,014	A	G	0	6	206
11	6591380	rs7942310	G	C	SNV	Transcript	intron	DNHD1	0,62	1,51E-03	1,166	2,45E-04	1,478	0,014	C	G	0	6	206
11	61257355	rs12787061	G	C	SNV	Coding Transcript	missense	PPP1R32	0,00	1,86E-04	-0,710	1,71E-04	-0,786	0,052	C	G	1	20	191
11	71850130	rs61734430	C	T	SNV	Coding Transcript	synonymous	FOLR3	11,70	9,91E-04	-0,776	6,97E-04	-0,879	0,031	T	C	1	11	199
11	108384207	rs2846412	C	T	SNV	Coding Transcript	missense	EXPH5	0,00	7,65E-04	-0,293	6,66E-04	-0,326	0,484	T	C	49	107	56
11	117899620	11:117899620	CAGG	C	DEL	Transcript	intron, feature truncation	TMPPRSS4-AS1	5,40	1,05E-04	-1,557	1,74E-03	-1,392	0,012	C	CAGG	0	5	207
11	129722553	rs558813	C	T	SNV	Coding Transcript	missense, splice	TMEM45B	0,00	1,94E-03	-0,329	1,56E-03	-0,369	0,196	T	C	10	63	139
12	95228286	rs2289030	G	C	SNV	Non Coding Transcript	non coding exon	MIR492	1,35	1,15E-03	-0,531	6,68E-04	-0,611	0,078	C	G	1	31	180
13	78216915	rs2813739	C	T	SNV	Coding Transcript	synonymous	SCEL	12,69	1,82E-03	-0,352	1,04E-03	-0,407	0,165	T	C	8	54	150
14	65878847	rs34007703	T	TG	INS	Regulatory Feature	regulatory	FUT8 *	10,28	1,41E-03	0,768	2,47E-04	0,967	0,035	TG	T	0	15	197
14	75279395	rs2302834	T	A	SNV	Coding Transcript	synonymous	YLPM1	9,74	6,74E-04	-0,314	1,03E-03	-0,334	0,361	T	A	25	103	84
14	77931034	rs61755659	A	G	SNV	Transcript	splice, intron	AHSA1	5,32	4,68E-05	-1,487	6,42E-04	-1,380	0,014	G	A	0	6	206
14	95053849	rs6113	T	C	SNV	Coding Transcript	synonymous	SERPINA5	0,01	1,63E-03	-0,458	1,67E-03	-0,503	0,106	C	T	1	43	168
14	95053863	rs6118	C	T	SNV	Coding Transcript	missense	SERPINA5	22,30	5,83E-04	-0,483	7,02E-04	-0,524	0,110	T	C	2	42	166
14	95054012	rs6119	A	G	SNV	Coding Transcript	missense	SERPINA5	0,00	6,13E-04	-0,506	6,67E-04	-0,553	0,102	G	A	1	41	169
15	42154738	rs16972193	G	A	SNV	Coding Transcript	missense, splice	SPTBN5	21,90	1,46E-04	-0,461	6,75E-04	-0,455	0,137	A	G	7	42	155
16	5122072	rs112683515	G	GGTCT	INS	Regulatory Feature	regulatory	ALG1 *	7,46	1,06E-03	-0,290	1,73E-03	-0,306	0,453	G	GGTCT	42	108	62
16	56686470	rs12051311	C	T	SNV	Intergenic	downstream	MT1CP	0,75	9,24E-04	-0,287	7,59E-04	-0,320	0,493	C	T	52	105	55
17	6115	rs61739713	G	C	SNV	Coding Transcript	synonymous	DOC2B	9,04	4,20E-04	-0,344	1,09E-03	-0,352	0,224	G	C	15	65	132
17	6157	rs2294076	A	G	SNV	Coding Transcript	synonymous	DOC2B	0,02	4,79E-04	-0,337	1,18E-03	-0,345	0,226	A	G	16	64	132
17	7224433	rs79385421	C	T	SNV	Coding Transcript	missense, splice	NEURL4	15,42	2,05E-04	-1,272	3,91E-04	-1,340	0,017	T	C	0	7	204
17	8064779	rs75664430	C	G	SNV	Intergenic	upstream	PER1	0,32	1,06E-03	-0,352	1,92E-03	-0,367	0,252	G	C	8	91	113
17	48437456	rs6504649	C	G	SNV	Coding Transcript	missense	XYLT2	6,18	1,52E-03	-0,311	1,73E-03	-0,338	0,370	G	C	21	115	76
17	56395757	rs2072147	C	T	SNV	Coding Transcript	missense	BZRAP1	17,08	3,38E-04	-0,984	8,12E-04	-1,014	0,026	T	C	0	11	201
17	56396609	rs2072145	T	C	SNV	Coding Transcript	missense	BZRAP1	9,23	3,38E-04	-0,984	8,12E-04	-1,014	0,026	C	T	0	11	201
18	29432624	rs3737374	T	C	SNV	Coding Transcript	missense, splice	TRAPP8	11,20	5,84E-04	0,329	3,08E-04	0,380	0,366	C	T	22	111	79
18	34647322	rs72890602	G	T	SNV	Coding Transcript	missense	KIAA1328	0,46	5,16E-04	-0,732	1,27E-03	-0,749	0,040	T	G	1	15	196
19	7831628	rs2277998	G	A	SNV	Coding Transcript	missense	CLEC4M	0,01	1,85E-03	-0,274	1,44E-03	-0,308	0,347	A	G	29	89	94
19	10444603	rs2230399	C	G	SNV	Coding Transcript	missense	ICAM3	4,28	1,05E-03	-0,565	6,02E-04	-0,650	0,073	G	C	0	31	181
19	10543272	rs139275433	CCT	C	DEL	Regulatory Feature	regulatory	PDE4A *	10,02	8,55E-04	-0,505	6,99E-04	-0,565	0,094	C	CCT	1	38	173
19	35633601	rs149407483	C	T	SNV	Regulatory Feature	regulatory	FXYP1 *	15,73	2,07E-05	-1,700	1,07E-03	-1,450	0,012	T	C	0	5	207
19	46269313	rs2014377	G	C	SNV	Coding Transcript	missense	SIX5	0,09	8,87E-04	0,362	3,92E-04	0,425	0,195	G	C	9	63	136
19	58945962	rs9941466	C	T	SNV	Coding Transcript	synonymous	ZNF132	10,29	1,10E-03	0,533	8,65E-04	0,599	0,078	C	T	1	31	180
21	40177325	rs73450548	G	A	SNV	Regulatory Feature	regulatory	ETS2 *	8,83	1,27E-03	-0,479	1,92E-03	-0,508	0,099	A	G	1	40	171
22	19165216	rs2854642	G	A	SNV	Regulatory Feature	regulatory	SLC25A1 *	7,23	1,26E-03	-0,625	8,31E-04	-0,712	0,050	A	G	1	19	191
22	19183787	rs1633399	A	G	SNV	Coding Transcript	missense	CLTCL1	15,45	1,65E-03	-0,636	1,27E-03	-0,716	0,045	G	A	1	17	194
22	19195680	rs807547	T	C	SNV	Coding Transcript	missense	CLTCL1	6,37	1,18E-03	-0,641	8,77E-04	-0,723	0,047	C	T	1	18	193
22	19197949	rs712952	G	A	SNV	Coding Transcript	missense	CLTCL1	32,00	1,18E-03	-0,641	8,77E-04	-0,723	0,047	A	G	1	18	193
22	19223352	rs807459	T	C	SNV	Coding Transcript	missense	CLTCL1	21,50	1,18E-03	-0,641	8,77E-04	-0,723	0,047	C	T	1	18	193
22	19750773	rs41298814	T	C	SNV	Coding Transcript	synonymous	TBX1	5,51	1,31E-03	0,354	1,65E-03	0,382	0,224	C	T	7	81	124
X	85233820	rs145707160	T	A	SNV	Coding Transcript	missense	CHM	23,30	5,69E-04	-1,168	1,10E-03	-1,230	0,031	A	T	0	7	105
X	114541027	rs2323736	A	G	SNV	Coding Transcript	synonymous	LUZP4	0,07	1,53E-03	-0,399	1,78E-03	-0,433	0,125	G	A	1	26	85

Note: The table is sorted on Chromosome and Position. Beta reflects the direction of toxicity - minor allele associated with high toxicity + minor allele associated with low toxicity. * the gene closest to the variant.

Abbreviations: NA, Not Available; Ref, Reference; Alt, Alternative; MAF, minor allele frequency; PV, phenotype value; PPV, potentially pertinent variant; CADD, Combined Annotation-Dependent Depletion; SNV, single nucleotide variant; DEL, deletion; INS, insertion; UTR, untranslated region.

Supplementary Table S3. All PPGs (genes with P-values < 0.002 for both PV2 and PV4). Gene-based association test results for thrombocytopenia using SKATO, adjusted for age and gender.

Chromosome	Gene name	PV2 P-value	PV4 P-value	Number of variants	Variants in gene
1	SERPINC1	1,55E-03	1,75E-03	5	rs5878; rs5877; rs2227627; rs121909551; rs2227624
2	DGKD	6,63E-04	1,44E-03	8	2:234344534; rs2305541; rs148025402; rs10208953; rs2228937; rs61739478; rs147134923; rs2228939
2	NDUFAF7	7,58E-04	1,83E-03	4	rs145277736; rs72875762; rs201231332; rs17409514
2	BIN1	9,19E-04	1,45E-03	13	rs375004668; rs138047593; rs61748155; 2:127811494; 2:127811592; rs2276579; rs371258305; rs2228955; rs114833236; rs61748157; rs41270608; 6:28887823; rs143463783; rs2230683
6	TRIM27	6,77E-05	7,86E-05	4	rs76463649; rs16893892; rs11965538; rs11965542; rs187327081; 6:28244225
6	ZSCAN26	1,28E-04	1,39E-04	6	rs2295230; rs41292882
6	RGS17	2,12E-04	7,13E-04	2	rs73392698; rs34788973; rs34957169; rs9368537; rs61742093
6	OR2B2	2,73E-04	2,92E-04	5	rs942379; rs56027330; rs1165165
6	SLC17A3	4,02E-04	1,41E-03	3	rs140243582; 6:30652729; rs9262143; 6:30652787; rs2269706
6	PPP1R18	4,29E-04	8,90E-04	5	rs11758339; rs144012429; rs201429821; 6:56334944; rs201115435; 6:56347530; rs187652380; 6:56362724; rs41271862; rs41271864; 6:56399939; rs4715630; rs4715631; 6:56417837; rs62621210; rs11756977; rs9367689; rs9464393; 6:56437548; rs375624981; rs201249286; rs4712138; rs150191284; rs112473525; rs115542549; rs115983672; rs6459166; rs141573097; rs12209266; 6:56483348; rs45487998; rs2230862; rs35497571; rs41271870; rs112640831; 6:56504813; rs2144407; rs151271595; rs113432929; rs1024196; rs41267673; rs16888128; rs182517599; rs11969595; rs150003298
6	DST	4,68E-04	1,20E-03	45	
6	ATAT1	5,04E-04	1,36E-03	6	6:30594726; rs146271858; rs6904236; 6:30610668; rs150614928; rs9262132
6	OR5V1	5,98E-04	6,94E-04	9	rs9257769; rs141436521; rs41271546; rs3117438; rs61742983; rs7761746; rs28359952; rs9257770; rs6930033
6	TRIM31	6,46E-04	1,22E-03	24	rs1116221; 6:30071354; rs17194258; rs146885411; 6:30071391; rs3132679; rs2023472; rs28400887; rs62624473; rs115893523; rs35775852; rs2523989; rs2239529; rs34801993; rs151255204; rs3734838; 6:30080265; rs2517598; rs35521684; rs34704616; rs143020056; 6:30080425; rs2844793; rs36063651
6	PSMB1	1,07E-03	6,78E-04	3	rs150833808; rs12717; rs10946279
6	KIF6	1,65E-03	7,41E-04	12	6:39311549; rs139582255; 6:39314286; rs20455; rs61748649; rs2273063; 6:39512404; rs7738892; 6:39580978; rs371477997; rs147204228; rs34016507
6	ZNF311	9,84E-04	1,54E-03	4	rs6456880; rs150648543; rs61751894; rs138000372
6	NKAPL	1,38E-03	1,56E-03	9	rs144553865; rs12000; rs1635; rs9461446; 6:28227824; rs3734564; rs147161729; rs201104720; rs1679709
7	MYL7	9,38E-04	3,39E-04	4	rs141852636; rs117678382; 7:44180361; rs4440539
7	CAPZA2	4,03E-04	4,21E-04	3	rs374052696; 7:116502628; rs4808
9	SPACA9	1,40E-03	1,02E-03	3	9:135762918; 9:135763786; rs2073869
10	CYP2C8	6,47E-04	4,18E-04	11	rs1058932; rs10509681; rs77147096; rs369479818; rs11188150; rs1058930; rs72558195; rs41286886; rs11572080; rs11572079; rs11572078
11	TMEM45B	4,52E-04	4,07E-04	3	rs145128105; rs558813; rs78733006
17	DOC2B	5,20E-04	1,24E-03	3	17:6052; rs61739713; rs2294076
X	CHM	1,86E-05	9,31E-05	2	X:85119740; rs145707160

Abbreviations: PV, phenotype value; PPG, potentially pertinent gene.

Supplementary Table S4. Variants included in the wGRS.

rsID	Minor allele	Beta for minor allele
rs34491125	C	-1,175
rs10491684	A	-0,627
rs374052696	T	-1,376
7:116502628	T	-0,691
rs4808	T	-0,331
6:39311549	A	0,186
rs139582255	T	-0,537
6:39314286	T	0,777
rs20455	G	-0,223
rs61748649	T	0,053
rs2273063	T	-0,412
6:39512404	G	-0,907
rs7738892	T	-0,304
6:39580978	T	-1,264
rs371477997	A	-1,407
rs147204228	T	-0,445
rs34016507	C	-0,743

Abbreviations: wGRS, weighted genetic risk scores.

Supplementary Table S5. Coefficients for the logistic regression models for the prediction of thrombocytopenia CTCAE and the decrease of platelets.

Model parameter	CTCAE model	Decrease model
Intercept	-65.5410027	-88.2483325
rs34491125	-0.4571852	-0.1265447
rs10491684	-17.5645283	0.0
rs374052696	-18.8444645	-36.8763664
7:116502628	0.3548092	0.3386908
rs4808	-0.6392537	-1.5352321
6:39311549	0.6752039	1.3721170
rs139582255	18.0025217	19.3322908
6:39314286	0.7531436	0.9941634
rs20455	18.5022042	19.8485981
rs61748649	18.9005386	20.5040707
rs2273063	17.0164587	0.0
6:39512404	1.8267082	16.1535803
rs7738892	18.2075950	19.7504007
6:39580978	1.3866380	2.4464639
rs371477997	1.1194446	1.1507203
rs147204228	0.9345956	1.3543144
rs34016507	3.3224846	20.9256816

Abbreviations: CTCAE, common terminology criteria for adverse events.