Leveraging GWAS Data Derived From a Large Cooperative Group Trial to Identify a SNP Cluster Associated With the Risk of Taxane-induced Peripheral Neuropathy

Maryam Lustberg (✉ Maryam.Lustberg@osumc.edu)

Ohio State University School of Medicine

Xuan Wu

Harvard School of Dental Medicine

Juan Luis Fernández-Martínez

primary endpoint solution, walthman

Enrique J. de Andrés-Galiana

University of Oviedo: Universidad de Oviedo

Santosh Philips

Indiana University School of Medicine

Jeffrey Leibowitz

Brigham and Women's Hospital

Bryan Schneider

Indiana University School of Medicine

Stephen Sonis

Dana-Farber Cancer Institute

Research article

Keywords: Chemotherapy-induced peripheral neuropathy (CIPN), neuropathy, clinical trial, Taxanes, Single nucleotide polymorphisms (SNPs)

Posted Date: February 23rd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-219649/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Leveraging GWAS data derived from a large cooperative group trial to identify a SNP cluster associated with the risk of taxane-induced peripheral neuropathy

Maryam Lustberg, MD, MPH^{1*}, Xuan Wu ², Juan Luis Fernández-Martínez, PhD^{4,5}, Enrique J. de Andrés-Galiana, PhD^{4,5}, Santosh Philips, PhD³, Jeffrey Leibowitz, PhD^{4,6}, Bryan Schneider, MD³, and Stephen Sonis, DMD, DMSc^{2,4,6,7}

¹Ohio State University School of Medicine, Columbus, Ohio, USA

²Harvard School of Dental Medicine, Boston, MA, USA

³Indiana University School of Medicine, Indianapolis, Indiana, USA

⁴Primary Endpoint Solutions, Waltham, MA, USA

⁵University of Oviedo, Oviedo, Spain

⁶Brigham and Women's Hospital, Boston, MA, USA

⁷Dana-Farber Cancer Institute, Boston, MA, USA

*Corresponding author

Maryam.Lustberg@osumc.edu

Abstract

Background

Chemotherapy-induced peripheral neuropathy (CIPN) is a common toxicity of taxanes for which there is no effective intervention. Genomic CIPN risk determination has yielded promising, but inconsistent results. The present study assessed the utility of a collective SNP cluster identified using novel analytic to describe taxane-associated CIPN risk.

Methods

We analyzed GWAS data derived from ECOG-5103, first identifying SNPs that were most strongly associated with CIPN using Fisher's ratio. We then ranked ordered those SNPs which discriminated CIPN-positive from CIPN-negative phenotypes based on their discriminatory power and developed the cluster of SNPs which provided the highest predictive accuracy using leave-one-out cross validation (LOOCV).

Results

Using GWAS aggregate data, we identified a 267 SNP cluster which was associated with a CIPN+ phenotype with an accuracy of 96.1%.

Conclusions

Identification of a 267 SNP cluster could accurately predict CIPN risk. Validation using an independent patient cohort should be performed.

Keywords: Chemotherapy-induced peripheral neuropathy (CIPN), neuropathy, clinical trial, Taxanes, Single nucleotide polymorphisms (SNPs)

Background

CIPN is a frequent dose-limiting toxicity [1] of taxanes, vinca alkaloids and platinum compounds which may persist for years following treatment. Consequences of CIPN include pain, numbness, falls, and difficulty in walking [2, 3]. Attempts at effective interventions for CIPN [4, 5] have not been successful and there are currently no confirmed preventative strategies.

Not all patients are at equal risk for CIPN. A relatively large number of studies have assessed the genomic basis for risk and reported a relationship between risk and specific single nucleotide polymorphisms (SNPs). However, the results of these investigations have been largely inconsistent [6-14]. Possible explanations for this observation include limitations of single SNP analysis methodology, candidate gene bias, differences in array platforms, and patient categorization. The reproducibility of standard genome-wide association studies (GWAS) has been hampered by range of other factors [15].

We demonstrated that cancer regimen- related toxicities can be reliably predicted using a learned SNP analysis strategy that is based on the hypothesis that risk identification is enhanced using a multiple-SNP signature [16]. Here, we applied a novel analytic to the comprehensive GWAS outputs from ECOG 5103 (Doxorubicin hydrochloride, cyclophosphamide, and paclitaxel with or without bevacizumab in treating patients with lymph node-positive or high-risk, lymph node-negative breast cancer; NCT00433511), and reported by Schneider et al [17] to identify a comprehensive, hierarchical single nucleotide polymorphism (SNP) cluster that could add to the robustness of CIPN risk prediction.

Methods

Our analyses were based on genotype data obtained from the ECOG-5103 clinical trial [17] in which two different arrays were used, HumanOmniExpress (727,227 SNPs, designated HOE) and HumanOmni1-Quad1 (1,131,857 SNPs and designated as HOQ1). The methods for

DNA isolation have been previously described [17]. HOE consisted of 322 samples; 13 samples from patients who did not develop CIPN and the remainder (n=309) from those who manifested moderate to severe CIPN. The HOQ1 cohort contained 581 samples. Forty-two samples (n=42) were from non-CIPN developers and the balance from patients who had significant symptoms. The two experiments were analyzed independently. Patients were divided into 2 classes: class 1, no CIPN; class 2, CIPN with grades of 3-4.

Alleles were designated as either A or B and parameterization of nucleotides was defined.

The parameterization used for the nucleotides was as follows:

A	1
C	2
T	3
G	4
I	5
D	6

With this parameterization, a mean of 3.77 for a particular SNP in one of the classes means that the corresponding nucleotide is mostly G (between T and G).

Analysis of the discriminatory power of the SNPs

We used Fisher's ratio as a measure of the individual discriminatory power of each SNP and then found the smallest-scale aggregate SNP cluster via leave-one-out cross validation (LOOCV). We then ordered SNPs by decreasing Fisher's ratio based on each SNP's discriminatory power.

We defined the Fisher's ratio for a SNP j in classes 1 and 2 (CIPN and no CIPN), c_1 , c_2 as:

$$FR_j = \frac{(\mu_{j1} - \mu_{j2})^2}{\sigma_{j1}^2 + \sigma_{j2}^2},$$

where μ_{j1}, μ_{j2} are measures of the center of the distribution (means) of the SNP j in classes 1 and 2 and $\sigma_{j1}^2, \sigma_{j2}^2$ are measures of the dispersion or variance within these classes. The use of Fisher's ratio as a measure of discriminatory power promotes SNPs that have centers of distribution in both classes separated with small dispersions within each class, allowing for a better discrimination among classes (CIPN and no CIPN).

Once we identified those SNPs which differentiated CIPN-positive from CIPN-negative, we ranked them in decreasing order based on their discriminatory power. We hypothesized that, rather than individual SNPs being most able to differentiate highest risk, the discriminatory power could be optimized by considering a cluster of SNPs. Consequently, we identified the smallest collection of SNPs with the highest prognostic accuracy using an algorithm based on recursive elimination of lower discriminatory SNPs. In our analysis, we differentiated those SNPs which were highly discriminatory from those which did not significantly contribute to informing differential risk. The predictive accuracy estimation identified the smallest-scale SNP signature via LOOCV and then ordering them by decreasing Fisher's ratio. This algorithm [18, 19] iteratively adds additional discriminatory SNPs to identify the minimum-scale signature with the maximum LOOCV predictive accuracy.

Next, we performed uncertainty sampling via the holdout sampler [19] to find those SNPs which appeared most often in all holdouts. Given the classifier

$$\mathbf{L}^*(\mathbf{g}): \mathbf{g} \in \mathbb{R}^S \rightarrow C = \{\text{CIPN}, \text{NoCIPN}\},$$

and having found the minimum scale signature with the highest predictive accuracy

$$Acc(\mathbf{g}) = 100 - \min O(\mathbf{g}),$$

with

$$O(\mathbf{g}) = \|\mathbf{L}^*(\mathbf{g}) - \mathbf{c}^{\text{obs}}\|,$$

representing the prediction error, that is, the number of uncorrected samples predicted by the classifier $\mathbf{L}^*(\mathbf{g})$, according to the observed class array of the training samples, $\mathbf{c}^{\text{obs}}$, the holdout

sampler serves to analyze the uncertainty space of the classifier, which is composed by the sets of high predictive genetic networks with similar predictive accuracy:

$$M_{tol} = \{\mathbf{g}: O(\mathbf{g}) < E_{tol} \rightarrow Acc(\mathbf{g}) > 100 - E_{tol}\}.$$

We included the uncertainty analysis because the number of monitored SNPS was much larger than the number of samples, resulting in an associated uncertainty space of the corresponding phenotype prediction problem. We dealt with this by limiting a minimum discriminatory power to the SNPs.

Results

Patient characteristics

Patients were part of the parent protocol (ECOG 5103) described by Schneider et al [17] and had consented for DNA analysis. ECOG 5103 was a phase III adjuvant breast cancer trial that randomized patients with node positive or high-risk node negative breast cancer to intravenous doxorubicin and cyclophosphamide every 2 or 3 weeks (at discretion of treating physician) for four cycles followed by 12 weeks of weekly paclitaxel alone (Arm A) or to the same chemotherapy with either concurrent bevacizumab or concurrent plus sequential bevacizumab. We focused our analysis on patients with available germline DNA who never developed CIPN (score of 0) vs. those who developed moderate to severe CIPN (scores of 3 of 4). We reasoned that focusing on the two most extreme phenotypes and excluding those patients who had mild CIPN, increased the likelihood of our more clearly differentiating genotype/phenotype differences.

Dataset analysis

Using the hold out experiment (HOE) dataset, 7.15% of the total number of alleles analyzed (104,063) had a discriminatory power > 0.5 as defined by their Fisher's ratio. The

median cumulative distribution function (cdf) for the set of the most discriminatory SNPs (n=110) was 1.97 with a low interquartile range (0.12) (**Figure S1**).

Using LOOCV analysis, we achieved an accuracy of 100% in differentiating patients with CIPN from those without CIPN with an aggregate 110 SNPs which had a FR > 3 (**Table S1**). Whereas the single most discriminatory SNP (rs969768_A; C in controls and T in patients) only provided a classification accuracy of 58%, adding the second most discriminatory SNP (rs4909801_B) increased LOOCV accuracy to 81.06%. Adding the other 3 first SNPs (adding rs9872883_A, rs3109154_A and rs2762927_A) resulted in a jump in accuracy to 95%. Most of the substitutions in these SNPs consisted replacement of C (no CIPN) by T or by A (NP) (**Table S2**).

We then performed hold out sampling (75% for training and 25% for validation) selecting those genes with a FR>2.2 in the training and having a LOOCV validation accuracy higher than 95%. Using this set of genes. Posterior frequency analysis of this SNP set (**Table S3**) to identify the most frequently sampled genes in 500 holdouts; 75% of the samples were randomly selected for training and the rest (25%) for blind validation to establish the accuracy of the discriminatory SNPs found in the training. Of the SNPs identified in this fashion, the differences in frequency between the sampled genes was low, suggesting that there exists many combinations with high discriminatory power.

HOQ1 Dataset

The HOQ1 data set relied on outputs from a HumanOmni1-Quad1 array. Of the 65,880 SNPs (alleles) which had a discriminatory power > 0.5 by the FR, only 2.9% were noted. The median CDF of the FR of the most discriminatory SNPs was 1.99 with a low interquartile range (0.3) (**Figure S2**). In contrast to the first data set, the maximum FR was only 2.37. Only 102 SNPs had an FR > 2.2.

Using LOOCV, we achieved an accuracy of 95.7% when the first 96 SNPs were collectively tested (**Table S4**). In contrast, we could only achieve an LOOCV accuracy of 53.2% when we only included the three most discriminatory SNPs (rs17585279_A, rs1893040_A, and rs3737336_A). However, expanding this cluster to include the top 15 SNPs, LOOCV accuracy jumped to 92.1%. Thus the contribution of the remaining 81 SNPs only had an incremental impact on accuracy of 3.6% (**Table S4**). It can be observed that almost all the discriminatory SNPs refer to the allele A.

We subsequently performed holdout sampling, selecting those genes with a FR>2.2 in the training set and having a LOOCV validation accuracy higher than 95%. Using this set of genes, we performed posterior frequency analysis. Table 5 shows the most frequently sampled genes in 500 hold outs. (**Table S5**).

Merged data

While we recognized the likelihood of microarray-associated inconsistencies [20], we elected to merge the two datasets to enhance the comprehensiveness of the predictive SNP-cluster. The merged dataset had 903 samples (55 no CIPN and 848 NP) and 1,403,967 common SNPs, that were monitored in both datasets. Within the set of SNPs with FR greater than 0.5, only 10,364 are in common to both datasets (**Figures 1**). The maximum accuracy (96.1%) was obtained with 267 SNPs (**Table S6**). **Table 1** shows the list of most frequently sampled genes by the holdout sampler.

We next evaluated the significance of the most discriminatory SNPs (n=267) derived from the merged SNP set. (**Figure 2**) shows the Fisher's ratio of the 267 discriminatory SNPs as a function of $-\log_{10}(p \text{ value})$. The equation of the regression line found was

$$-\log_{10}(p \text{ value}) = 1.9926 + 0.0386 * FR,$$

that is, a higher FR implies a lower p-value, as it should be expected. The SNPS with a FR>1 would have a p-value lower than 0.0093 (**Figure 2**).

Discussion

CIPN is a common side effect of many chemotherapy regimens with an aggregate prevalence of 68.1% [21] Severe CIPN is especially dramatic among patients receiving taxanes for adjuvant treatment of early breast cancer [22].

Genomics' contribution to CIPN risk has been studied using candidate gene and GWAS [6-14] approaches of which one of the most comprehensive was reported by Schneider et al using data from over 3,000 breast cancer patients [17]. Here, we applied a novel analytical algorithm to identify the SNP cluster (rather than individual SNPs) predictive of CIPN risk hypothesizing that such an approach reflected the biological interactions associated with CIPN pathogenesis.

To increase the likelihood that we would identify a distinctive SNP/gene set associated with CIPN risk, we only studied patients who represented two phenotypic extremes – those with no evidence of CIPN and those with moderate to severe manifestations of the condition. We first analyzed each of the two datasets identified by Schneider based on their array platform and described predictive clusters of 110 and 96 SNPs respectively. Within each dataset, we created a hierarchy of SNPs such that the addition of additional SNPs did not favorably impact predictive accuracy. We tested the validity of both clusters by testing against a withheld cohort consisted of 25% of subjects in each database.

We reasoned that conglomerating the two databases might enhance our prediction and functionality objectives. The merged dataset also provided the advantage that the number of non-CIPN controls was increased (n=55) so we were able to create a 267 SNP cluster which predicted CIPN risk with an accuracy of 96.1% ($p<0.0093$).

Our findings demonstrate the potential value of an undirected analytical approach to determine CIPN risk. An association between SNPs and CIPN is not new. Nor is the use of SNPs to assess the risk of other cancer regimen-related toxicities [23]. Typically associations between SNPs and phenotypes have been done using GWAS or candidate gene approaches with the goal of identifying *the* most predictive SNP or gene. Replication of findings has been problematic and critical reviews of these approaches have been reported [15, 24]. Our process was driven by the concept that risk is the consequence of multiple genes simultaneously interacting to impact phenotype. In this case, genes were defined by attribution from SNPs, an approach that excludes non-SNP-related genes. While these may not be uniquely associated with risk, the assumption that they did not contribute globally undermines a systems-based hypothesis. While SNP arrays provide huge potential value, a whole genome array could be more comprehensive. The results of the network generation exercise included in this paper is illustrative of that potential. Likewise, to assume that risk is solely attributable to genomic influencers is naive. More likely, it is the sum of metabolic, epigenetic, proteomic, and microbiome elements among others. Analyses of all these inputs will require the application of newer methods such a multiplex networks [25, 26].

External validation of the SNP set described here will be required to confirm its clinical utility. The mechanistic information provided by our analyses is reassuring in that it confirms the value of genomic-based assessments to describe potential CIPN pathogenesis.

Conclusion

We demonstrated that SNP analysis strategy can reliably predict the cancer regimen-related toxicities, where the risk identification is improved by using a multiple-SNP signature. We applied a novel analytic to the comprehensive GWAS outputs from ECOG 5103 (Doxorubicin hydrochloride, cyclophosphamide, and paclitaxel with or without bevacizumab in treating patients with lymph node-positive or high-risk, lymph node-negative breast cancer; NCT00433511), to

recognize a comprehensive, hierarchical single nucleotide polymorphism (SNP) cluster that robustly predicts the CIPN risk. By using the GWAS aggregate data, we identified the SNP cluster which was associated with a CIPN+ phenotype with an accuracy of 96.1%. Thus, the present work contributes in accurately determining the CIPN risk; however, validation studies using an independent patient cohort should be performed.

Declaration

Abbreviations

CIPN: Chemotherapy-induced peripheral neuropathy; GWAS: standard genome-wide association studies; SNP: single nucleotide polymorphisms; HOQ1: HumanOmni1-Quad1; HOE: hold out experiment; LOOCV: Leave-one-out cross validation; CDF: cumulative distribution function.

Ethics approval and consent to participate

This is a secondary analysis of a completed clinical trial ECOG 5103. It was previously approved by ethics and patients were consented for participation and analysis of correlative data.

Consent for publication

Yes.

Availability of data and material

GWAS Data were provided by ECOG 5103.

Competing interests

Dr. Sonis reports personal fees from Biomodels, LLC, personal fees from Primary Endpoint Solutions, LLC, outside of the submitted work. As an employee of Biomodels and PES, he is involved in assisting industry, government and academics in the study and enablement of drugs, biologicals and devices to treat patients for a broad range of indications including cancer and

toxicities of cancer therapy. He does not have equity or receive payment from any of the companies' clients. The other authors have no relevant disclosures.

Funding

This work was supported by grants from Division of Cancer Prevention, National Cancer Institute (R01CA238946-02).

Authors' contributions

SS and MBL conceived of study concept and wrote manuscript, and edited manuscript. All others authors, analyzed data reviewed and edited manuscript.

Acknowledgement

We thank Dr. Antonio Wolff (Johns Hopkins University), Dr. Kathy Miller (Indiana University), and all the patients who participated in ECOG 5103 for providing access to the data that were used to analyze the reported results in this manuscript.

References:

1. Cavaletti, G., et al., *Chemotherapy-induced peripheral neurotoxicity: A multifaceted, still unsolved issue*. J Peripher Nerv Syst, 2019. **24 Suppl 2**: p. S6-S12.
2. Autissier, E., *Chemotherapy-Induced Peripheral Neuropathy: Association With Increased Risk of Falls and Injuries*. Clin J Oncol Nurs, 2019. **23**(4): p. 405-410.
3. Shah, A., et al., *Incidence and disease burden of chemotherapy-induced peripheral neuropathy in a population-based cohort*. J Neurol Neurosurg Psychiatry, 2018. **89**(6): p. 636-641.
4. Hu, S., et al., *Recent Developments of Novel Pharmacologic Therapeutics for Prevention of Chemotherapy-Induced Peripheral Neuropathy*. Clin Cancer Res, 2019. **25**(21): p. 6295-6301.
5. Jordan, B., et al., *Prevention and Management of Chemotherapy-Induced Polyneuropathy*. Breast Care (Basel), 2019. **14**(2): p. 79-84.

- 299 6. Cliff, J., et al., *The molecular genetics of chemotherapy-induced peripheral neuropathy:*
300 *A systematic review and meta-analysis.* Crit Rev Oncol Hematol, 2017. **120**: p. 127-140.
- 301 7. Dolan, M.E., et al., *Clinical and Genome-Wide Analysis of Cisplatin-Induced Peripheral*
302 *Neuropathy in Survivors of Adult-Onset Cancer.* Clin Cancer Res, 2017. **23**(19): p. 5757-
303 5768.
- 304 8. Leibovici, A., R. Sharon, and D. Azoulay, *BDNF Val66Met is Associated with Pre-*
305 *existing but not with Paclitaxel-induced Peripheral Neuropathy in an Israeli Cohort of*
306 *Breast Cancer Patients.* Isr Med Assoc J, 2018. **20**(12): p. 746-748.
- 307 9. Mahmoudpour, S.H., et al., *Chemotherapy-induced peripheral neuropathy: evidence*
308 *from genome-wide association studies and replication within multiple myeloma patients.*
309 BMC Cancer, 2018. **18**(1): p. 820.
- 310 10. Park, S.B., et al., *Clinical and genetic predictors of paclitaxel neurotoxicity based on*
311 *patient- versus clinician-reported incidence and severity of neurotoxicity in the ICON7*
312 *trial.* Ann Oncol, 2017. **28**(11): p. 2733-2740.
- 313 11. Terrazzino, S., et al., *Genetic determinants of chronic oxaliplatin-induced peripheral*
314 *neurotoxicity: a genome-wide study replication and meta-analysis.* J Peripher Nerv Syst,
315 2015. **20**(1): p. 15-23.
- 316 12. Abraham, J.E., et al., *Replication of genetic polymorphisms reported to be associated*
317 *with taxane-related sensory neuropathy in patients with early breast cancer treated with*
318 *Paclitaxel.* Clin Cancer Res, 2014. **20**(9): p. 2466-75.
- 319 13. Reyes-Gibby, C.C., et al., *Genome-wide association study identifies genes associated*
320 *with neuropathy in patients with head and neck cancer.* Sci Rep, 2018. **8**(1): p. 8789.
- 321 14. Boora, G.K., et al., *Testing of candidate single nucleotide variants associated with*
322 *paclitaxel neuropathy in the trial NCCTG N08C1 (Alliance).* Cancer Med, 2016. **5**(4): p.
323 631-9.

- 324 15. Marigorta, U.M., et al., *Replicability and Prediction: Lessons and Challenges from*
325 *GWAS*. Trends Genet, 2018. **34**(7): p. 504-517.
- 326 16. Reinbolt, R.E., et al., *Genomic risk prediction of aromatase inhibitor-related arthralgia in*
327 *patients with breast cancer using a novel machine-learning algorithm*. Cancer Med,
328 2018. **7**(1): p. 240-253.
- 329 17. Schneider, B.P., et al., *Genome-Wide Association Studies for Taxane-Induced*
330 *Peripheral Neuropathy in ECOG-5103 and ECOG-1199*. Clin Cancer Res, 2015. **21**(22):
331 p. 5082-5091.
- 332 18. Saligan, L.N., et al., *Supervised classification by filter methods and recursive feature*
333 *elimination predicts risk of radiotherapy-related fatigue in patients with prostate cancer*.
334 Cancer Inform, 2014. **13**: p. 141-52.
- 335 19. Fernández-Martínez, J.L., et al. *Sampling Defective Pathways in Phenotype Prediction*
336 *Problems via the Holdout Sampler*. 2018. Cham: Springer International Publishing.
- 337 20. Liu, F., et al., *Performance comparison of multiple microarray platforms for gene*
338 *expression profiling*. Methods Mol Biol, 2012. **802**: p. 141-55.
- 339 21. Brewer, J.R., et al., *Chemotherapy-induced peripheral neuropathy: Current status and*
340 *progress*. Gynecol Oncol, 2016. **140**(1): p. 176-83.
- 341 22. Willson, M.L., et al., *Taxanes for adjuvant treatment of early breast cancer*. Cochrane
342 Database Syst Rev, 2019. **9**: p. CD004421.
- 343 23. Palmirotta, R., et al., *SNPs in predicting clinical efficacy and toxicity of chemotherapy:*
344 *walking through the quicksand*. Oncotarget, 2018. **9**(38): p. 25355-25382.
- 345 24. Tam, V., et al., *Benefits and limitations of genome-wide association studies*. Nat Rev
346 Genet, 2019. **20**(8): p. 467-484.
- 347 25. Chung, R.-H. and C.-Y. Kang, *A multi-omics data simulator for complex disease studies*
348 *and its application to evaluate multi-omics data analysis methods for disease*
349 *classification*. GigaScience, 2019. **8**(5).

26. Kivelä, M., et al., *Multilayer networks*. Journal of Complex Networks, 2014. **2**(3): p. 203-271.

Figure Legends

Figure 1: Merged dataset. Cdf of the Fisher's ratio in the set of the discriminatory SNPs

Figure 2: Merged data set. Fisher's ratio of the 267 discriminatory SNPs as a function of $-\log_{10}(\text{p-value})$.

Additional file:

File name: CIPN1 supplemental tablesandfigures_1.20.2021_NV1.docx

Title of data:

Figure S1: HOE Dataset. Cumulative probability distribution function (cdf) of the Fisher's ratio in the set of discriminatory SNPs.

Table S1: HOE Dataset. List of the most discriminatory SNPs (C1: No CIPN; C2 CIPN). The minimum scale signature contains 30 SNPs to achieve 100% LOOCV accuracy.

Table S2: Nucleotide substitutions in the five most discriminating SNPs.

Table S3: HOE Dataset. Most frequently samples SNPs with frequency > 49% (C1: No CIPN; C2 CIPN).

Figure S2: HOQ1: cdf (cumulative probability) of the Fisher's ratio in the set of discriminatory P1 SNPs. The median, lower and upper quartiles are close to a $FR=2$.

Table S4: HOQ1 Dataset. List of the most discriminating SNPs (C1 no CIPN; C2 CIPN). The signature with the first 15 SNPs provides an LOOCV accuracy of 92.08%. The incremental gain with the addition of each SNP to the signature is small, but achieves an overall accuracy of 95.1% when the 96 most discriminatory SNPs are included. All the SNPs are in the allele A.

376 **Table S5.** List of the most frequently sampled SNPs in the HOQ1 dataset with a frequency >
377 3%.

378 **Table S6:** List of the most discriminatory SNPs in the merged dataset. Incremental gains in
379 accuracy are noted in the far right column. (C1 – no CIPN; C2 CIPN).

380

381

382

Figures

Figure 1

Merged dataset. Cdf of the Fisher’s ratio in the set of the discriminatory SNPs

Figure 2

Merged data set. Fisher's ratio of the 267 discriminatory SNPs as a function of $-\log_{10}(\text{p-value})$.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [CIPN1supplementaltablesandfigures1.20.2021NV1.docx](#)