Oncology

Oncology and Informatics – Research Article

Oncology 2024;102:1029–1040
DOI: 10.1159/000540395

Received: January 12, 2024
Accepted: July 9, 2024
Published online: July 18, 2024

Artificial Intelligence-Driven Prediction Revealed CFTR Associated with Therapy Outcome of Breast Cancer: A Feasibility Study

Mária Kováčová[a] Viktor Hlaváč[b, c] Renata Koževnikovová[d] Karel Rauš[e] Jiří Gatěk[f] Pavel Souček[b, c]

[a]Third Faculty of Medicine, Charles University, Prague, Czech Republic; [b]Laboratory of Pharmacogenomics, Biomedical Center, Faculty of Medicine in Pilsen, Charles University, Pilsen, Czech Republic; [c]Toxicogenomics Unit, National Institute of Public Health, Prague, Czech Republic; [d]Department of Oncosurgery, MEDICON, Prague, Czech Republic; [e]Institute for the Care for Mother and Child, Prague, Czech Republic; [f]Department of Surgery, EUC Hospital and University of Tomas Bata in Zlin, Zlin, Czech Republic

Keywords
Machine learning · Breast cancer · Gene prioritisation · Survival · Cystic fibrosis transmembrane conductance regulator

Abstract
Introduction: In silico tools capable of predicting the functional consequences of genomic differences between individuals, many of which are AI-driven, have been the most effective over the past two decades for non-synonymous single nucleotide variants (nsSNVs). When appropriately selected for the purpose of the study, a high predictive performance can be expected. In this feasibility study, we investigate the distribution of nsSNVs with an allele frequency below 5%. To classify the putative functional consequence, a tier-based filtration led by AI-driven predictors and scoring system was implemented to the overall decision-making process, resulting in a list of prioritised genes. ***Methods:*** The study has been conducted on breast cancer patients of homogeneous ethnicity. Germline rare variants have been sequenced in genes that influence pharmacokinetic parameters of anticancer drugs or molecular signalling pathways in cancer. After AI-driven functional pathogenicity classification and data mining in pharmacogenomic (PGx) databases, variants were collapsed to the gene level and ranked according to their putative deleterious role. ***Results:*** In breast cancer patients, seven of the twelve genes prioritised based on the predictions were found to be associated with response to oncotherapy, histological grade, and tumour subtype. Most importantly, we showed that the group of patients with at least one rare nsSNVs in cystic fibrosis transmembrane conductance regulator (*CFTR*) had significantly reduced disease-free (log rank, $p = 0.002$) and overall survival (log rank, $p = 0.006$). ***Conclusion:*** AI-driven in silico analysis with PGx data mining provided an effective approach navigating for functional consequences across germline genetic background, which can be easily integrated into the overall decision-making process for future studies. The study revealed a statistically significant association with numerous clinicopathological parameters, including treatment response. Our study indicates that *CFTR* may be involved in the processes influencing the effectiveness of oncotherapy or in the malignant progression of the disease itself.

karger@karger.com
www.karger.com/ocl

Correspondence to:
Mária Kováčová, maria.kovacova@lf3.cuni.cz

Introduction

Breast cancer is a complex disease with profound genetic heterogeneity. Globally, it continues to be the primary cause of death for women [1]. Germline variants in the background of the tumour genome can have a detrimental effect on onset and overall prognosis of the disease. The likelihood of harbouring germline tumour background variants that may affect the efficacy of oncotherapy is as high as 11 out of 25 patients [2]. Rare non-synonymous single nucleotide variants (nsSNVs) with an allele frequency of less than 5% are the most prevalent among SNVs in drug-related genes and show geographical association to specific populations [3, 4]. Due to their functional impact on drug metabolism, transport, or target interactions, they may have a profound effect on pharmacokinetic or pharmacodynamic properties of drug substances, leading to an increased systemic toxicity or failure of the therapy. Therefore, it is imperative to adopt a rapid and effective approach to interrogate the germline genetic portfolio that may be responsible for treatment without adequate response or may contribute to cancer progression.

In silico methods predicting potentially pathogenic (i.e., functional or deleterious) consequences of variations in the genome, have been developed early since the discovery of the human genome. To date, the best predictive tools have been demonstrated mainly for nsSNVs, which result in a change of one amino acid in the protein sequence. Yet, no gold standard tools have been developed for drug response-related variants, also known as pharmacogenomic (PGx) variants [5, 6]. Although these tools are primarily developed using datasets with variants linked to a disease, machine learning (ML) approaches, a crucial aspect of artificial intelligence, are the most commonly used along with the sequence conservation-based tools, e.g., MutationAssessor [7–9]. In terms of cancer progression prediction, the most recent studies show that the best prediction performance on functional cancer-related variant datasets, including breast cancer, was demonstrated by MetaSVM tool, but primarily by REVEL, which was, indeed, specifically developed for rare variants [10, 11].

Here, we report the results of a feasibility study conducted on a homogeneous ethnic group of breast cancer patients. In relation to the treatment prognosis, we examined the proportion of rare nsSNVs per gene guided by AI-driven prediction for pathogenicity.

Methods

Gene Panel

Gene panel selection followed criteria described in [12] using PGx databases and bibliography data in association to breast cancer and/or to drug substances like anthracyclines (adriamycin, daunorubicin, epirubicin), 5-fluorouracil (5-FU), cyclophosphamide, taxanes (paclitaxel or docetaxel) including their pharmacokinetic pathways (see online suppl. Table S1; for all online suppl. material, see https://doi.org/10.1159/000540395). PGx significant genes designated by PharmGKB as very important pharmacogenes (VIP) account for 8% (n = 40) out of 509 genes in the panel.

Patient Data and Pharmacotherapy

Clinical characteristics and drug therapy of 105 breast cancer patients, representing the Czech European Caucasian population, enrolled in the study are in online supplementary Table S2. Retrieval of clinical data, collection of blood samples, DNA extraction and samples handling procedures were followed as described previously [12]. A subset of the patients (n = 68, 65%) underwent neoadjuvant cytotoxic therapy (NACT), i.e., chemotherapy before surgery. The rest of the patients received adjuvant chemotherapy. Chemotherapy administered to all patients under study included combination regimens based on 5-FU, anthracycline either adriamycin (doxorubicin) or epirubicin and cyclophosphamide (FAC, FEC). Another combination regime included docetaxel, adriamycin, and cyclophosphamide (TAC). In adjuvant chemotherapy, adriamycin or epirubicin in combination with cyclophosphamide (AC or EC) was administered. Adjuvant hormonal therapy in postmenopausal patients included aromatase inhibitors (anastrozole, letrozole) and selective oestrogen receptor modulators such as tamoxifen. In premenopausal patients, gonadotropin-releasing hormone (GnRH) agonist goserelin was administered in the course of adjuvant therapy. Response to the NACT was evaluated based on routine imaging methods and the Response Evaluation Criteria in Solid Tumours (RECIST) v1.1 [13].

Targeted Sequencing and Bioinformatic Variant Annotation

Our in-house pipeline for processing and quality control of raw sequencing data was followed [12]. Reads were mapped on the reference sequence the Genome Reference Consortium Human Build 37 (GRCh37). Variants were annotated using Annovar with dbSNP release 151 and the bioinformatic functional prediction was processed via dbNSFP3.2. Allelic frequencies for patients were calculated to prevent missing data for variants without assigned frequency in Exome Aggregation Consortium-non-Finnish European (ExAC-NFE) population and filtered by the MAF ≤0.05 cut off. Allelic frequencies in the ethnically matched population from Czech National Center for Medical Genomics (NCMG) and from ExAC-NFE population [14] were used in the next step for comparison. For PGx variants acquired through data mining, the European allele frequency was updated by using the Allele Frequency Aggregator (ALFA) [15].

Bioinformatic Functional Prediction

Computational prediction was derived for all nsSNVs and collapsed by the number of nsSNVs to a gene level. The selection was limited to nsSNVs apart from PGx validated variants where filtration was extended to whole exonic consequence and (MAF ≤0.05). Genes with the highest number of putatively

pathogenic variants were further investigated for clinical association in breast cancer patients of the study (see decision-making process to prioritise final set of genes). For prediction of pathogenicity, we have selected three ML-based tools (REVEL [16], MetaSVM [17], and CADD [18]) and one, sequence conservation-based tool (MutationAssessor [19]) with threshold optimised for PGx variants as published elsewhere [9]. Specifically, REVEL, the ensemble method uses supervised machine learning with Random Forest algorithm and integrates 13 separate tools. Some of these tools are based on naive Bayes classifier, Hidden Markov Models, and Random Forest algorithms. Although MutationAssessor is also integrated to REVEL, the separate prediction was run with a more stringent and PGx specific threshold. Further selected AI-driven tools MetaSVM and CADD provide ensemble scores. MetaSVM integrates nine independent scores, allele frequency for true negative common and rare variants and uses a support vector machine (SVM) algorithm. CADD integrates various genomic annotations and features using a machine learning approach with logistic regression algorithm.

With intention to decrease the number of false positive predictions by these tools run separately [20], a specific set (i.e., Benign set) of AI-driven tools (REVEL, VEST3 3.0 [21] and MetaSVM) with concordant prediction of benign consequence was integrated into the tier-based filtration process. Absence of prediction by any of the tools in the combination discarded affected variants from this consensus assessment.

Threshold for prediction (binary or numerical) followed the recommended values balanced for optimal specificity and sensitivity as defined by tool developers. For pathogenic prediction: REVEL (score >0.5), MetaSVM (D: Deleterious), CADDphred (score ≥19), and MutationAssessor (score >2.0566). Threshold values for benign prediction: REVEL (score ≤0.5), VEST3 (score ≤0.8), MetaSVM (T: Tolerant).

Gene length data were acquired via Galaxy (version 0.1.2) [22]. Galaxy counts the number of bases in all exons of a gene, after merging any overlapping exons from different transcripts. Results from predictors along with gene length were integrated into an in-house database using MySQL (version 6.3.10).

Filtration Process

Pathogenic variants as predicted by REVEL, MetaSVM, CADD, and MutationAssessor were collapsed to gene level and further processed. Genes were ranked by the number of pathogenic variants with an arbitrary set threshold for inclusion of minimum 3 pathogenic nsSNVs per gene. Further, we have limited this process up to ten genes (arbitrary threshold). In cases where genes had the same number of pathogenic variants per gene at the last position of the set (n = 10), an additional condition was applied. This allowed for inclusion of more than 10 genes with an identical rank. Exclusion process was divided into Tier 1 and Tier 2. In a Tier 1, consensus prediction for benign variants (i.e., Benign set) was called and collapsed to a gene level (≥3 nsSNVs/gene, up to 10 [+] genes of identical rank). The resulting sets of pathogenic genes for each tool and the Benign set were compared. This step overruled prediction for pathogenicity in all four predictors and eliminated overlapping genes. In a Tier 2, all unique genes across four predictors were discarded from the final gene selection process due to presumably low predictive power.

Data Normalisation

The intention was to adjust results from REVEL, MetaSVM, CADD, and MutationAssessor and to acquire a set of genes with the highest "gene pathogenic ratio" (GPR) for a given algorithm. Considering that the study is narrowed to protein-coding regions, the number of nsSNVs per gene length (kbp) was used in this normalization step. GPR specific for each tool (nsSNVs/kbp) was calculated and genes were ranked by the GPR values. Conditions for inclusion and exclusion criteria were followed as described above.

PGx Evidence (PGx Data)

Data mining for PGx-related variants in PharmVar and ClinVar databases as described below was conducted. Many of these variants were in vitro or in vivo validated and thus are superior to predictions from in silico tools built on disease-related variants.

Genes and exonic variants (GRCh37), available in the PharmVar (version 5.2.14) database were downloaded [23] and compared with our gene panel. In addition, web-based information which defines the functional variants by Clinical Pharmacogenetics Implementation Consortium (CPIC) by evidence for affecting gene function as "decreased function," "severely decreased," or "increased function" was matched to downloaded genes and variants. Genes with a limited/absent or literature only data were omitted. PharmGKB database (clinicalAnnotations file) provided information on VIP genes.

ClinVar (version 20180603) VCF data integrated into an in-house database were searched for PGx clinical significance term "drug response" (a general term for a variant that affects a drug response, not a disease) [24]. In addition, for prioritised genes, data were supplemented by updated search at National Center for Biotechnology Information (NCBI).

Decision-Making Process to Prioritise Final set of Genes (in silico Prediction and PGx Data)

Comparison of results across four predictors with the addition of PGx data served as a point of departure for the final set of genes intended for statistical analysis. Score = 1 was assigned for concordantly called genes to each predictor including the predictors with normalised results (n = 8). The classification for the total score was as follows: very high (score ≥8–7 and PGx evidence), high (score ≥6–5 methods and PGx evidence), medium (score ≥4–3 and PGx strong evidence), or low (score ≤2 and/or PGx strong evidence). "PGx" evidence was assigned to genes included in the PharmVar and ClinVar database (see PGx data) and "PGx strong evidence" throughout the additional inclusion of the gene to the VIP PharmGKB group. The AI-driven decision encompassed the highest ratio in the scoring system (i.e., 6/8 score). Prediction called by two tools only (n = 2) was considered as insufficient for conducting statistical analysis and these genes were excluded.

Statistical Analysis

Although we are aware that for study of rare variants a larger sample size would be needed, this is a feasibility study and as such lacks a formal analysis for sample size calculation [25]. Ethnic homogeneity and available phenotype records including the response to therapy of patients in our study, however, provide sound data for studying germline variants [26].

To determine whether there is a significant association between two categorical variables in the final set of prioritised genes

(defined by presence of nsSNV) and clinical data (nominal variables), including therapy response, the Pearson's χ^2 test (degree of freedom, df = 1) was used. In the case of small sample size in compared groups, the Fisher's exact test was followed. Odds ratios (ORs) were expressed with 95% confidence intervals (CIs). Tested variables were as follows: tumour size (pTis/pT1 vs. pT2-4), lymph node metastasis (pN0 vs. pN1-2), pathology stage (SI vs. SII–III), oestrogen, progesterone, and ERBB2 receptor status (positive vs. negative), Ki-67 expression (positive vs. negative based on 13.25% cut off [27]), intrinsic molecular subtype luminal A versus luminal B and triple negative, and response to NACT (non-responders, i.e., stable or progressive disease vs. responders, i.e., partial response). Disease-free survival (DFS) was defined as the time elapsed between surgery and relapse. Overall survival (OS) was considered as the time elapsed from surgery to the patient's death due to any cause. The study follow-up was 120 months and DFS data were censored to this value. Survival functions of patients stratified by carriage of nsSNVs were evaluated by the log-rank test. Kaplan-Meier plots were generated by R version 4.2.3. Cox regression analysis for DFS and OS was adjusted to tumour grade and disease stage. All tests were two-sided and p values <0.05 were considered statistically significant. Analyses were conducted using the statistical program SPSS v16.0 (SPSS, Chicago, IL, USA).

Results

The most pathogenic genes defined by AI-driven decision and filtration were *CFTR, ABCC1, SLC22A1, TUBB1*, *ABCB6*, and *SLC22A18*. Overall, we have detected 1,080 rare nsSNVs for 365 out of 509 sequenced genes. Out of these genes, 34.5% (n = 126) were affected by only one rare nsSNV, i.e., singleton. In total, 7.8% (n = 84) of rare nsSNVs were novel, i.e., lacking the reference number in dbSNP151. The allele frequency for 13.1% of detected nsSNVs (n = 141) was absent in the closely ethnically matched population (ExAC-NFE) and for 12.3% variants (n = 133) in the Czech presumably healthy population (NCMG). The highest number of variations per chromosome and gene was observed for chromosome 13 with 25 nsSNVs in four genes (ratio 6.3), chromosome 16 with 64 nsSNVs in 13 genes (4.9) and chromosome 7 with 126 nsSNVs in 30 genes (4.2) (see online supplementary Table S3). At gene level, the highest numbers of nsSNVs were in *ABCA13* (n = 34), *ABCA4* (n = 21), and *ABCC1* (n = 14) (see online supplementary Table S4).

Benign Set

In the Tier 1 exclusion step (see Filtration Process), 300 genes were predicted as benign (gene length 3,785 ± 2,553 bp, mean ± SD). In total, 737 nsSNVs (novel, n = 50, 6.8%) were distributed in a range of 7–23 nsSNVs per gene in the ten most frequently mutated genes. Herein 95 genes (4,533 bp ± 3,143, mean ± SD) were identified with at least 3 nsSNVs per gene. Six genes (*ABCA13, ABCA12, NCOR2, PIK3C2B*, *CIT*, and *ABCA4*) were shared with pathogenic high-rank genes and two genes (*ABCA4* and *ABCC2*) were shared with normalised sets of genes and thus were excluded (see online supplementary Table 5).

Pathogenic Sets

The highest pathogenic score was acquired by AI-driven algorithms (REVEL, MetaSVM and CADD) and MutationAssessor after Tier 1 (Benign set) and Tier 2 (unique) filtration for five (36%) genes *CFTR, ABCC1, SLC22A1*, *TUBB1*, and *ABCA1* (see Fig. 1a; online suppl. Table 4). After adjusting the results for gene length, only *SLC22A1* (5%) remained classified as the most pathogenic by all tools (see Fig. 1b; online suppl. Table 5). Final set of prioritised genes and total score per gene are listed in Tables 1 and 2.

PGx Databases

PGx Related Variants Were Identified in More than 26% of Breast Cancer Patients of the Study

PharmVar currently deposits 25 genes, all but one (*NUDT15*) are shared with the gene panel of the study. Exonic variants appointed for decreased function of *DPYD* (n = 2) and *CYP2D6* (n = 3) were identified for 16% of patients in our dataset, see Table 3. Variants severely decreasing or increasing gene function were not identified. Data mining in the ClinVar database revealed additional variants in *CYP2D6* (n = 2) and *ABCC1* (n = 1) (see Table 3) for 12 patients. Final set of prioritised genes and total score per gene from predictions are listed in Tables 1 and 2. Allele frequency for PGx nsSNVs of breast cancer patients in the prioritised genes was comparable in terms of the set threshold value, i.e., below the MAF ≤0.05 with data from ALFA, ExAC-NFE and NCMG.

Associations with Clinical Data and Therapy Prognosis

Disease-Free-Survival and OS Was Significantly Reduced in Patients with CFTR nsSNVs

The study revealed *CFTR* as a potentially important gene significantly associated with an oncotherapy prognosis in breast cancer. Patients (n = 16, 15.2%) carrying any of rare germline nsSNVs (n = 12) in *CFTR* had significantly reduced DFS (log rank, Mantel-Cox, p = 0.002) and OS (log rank, Mantel-Cox, p = 0.006) in comparison to patients without such variants (see Fig. 2). Multivariate analysis using Cox regression adjusted to tumour grade and disease stage confirmed that nsSNVs in *CFTR* provide independent prognostic information for DFS (HR = 3.03, 95% CI: [1.32–7.00], p = 0.009) and OS (HR = 3.38, 95% CI: [1.37–8.37], p = 0.008), respectively.

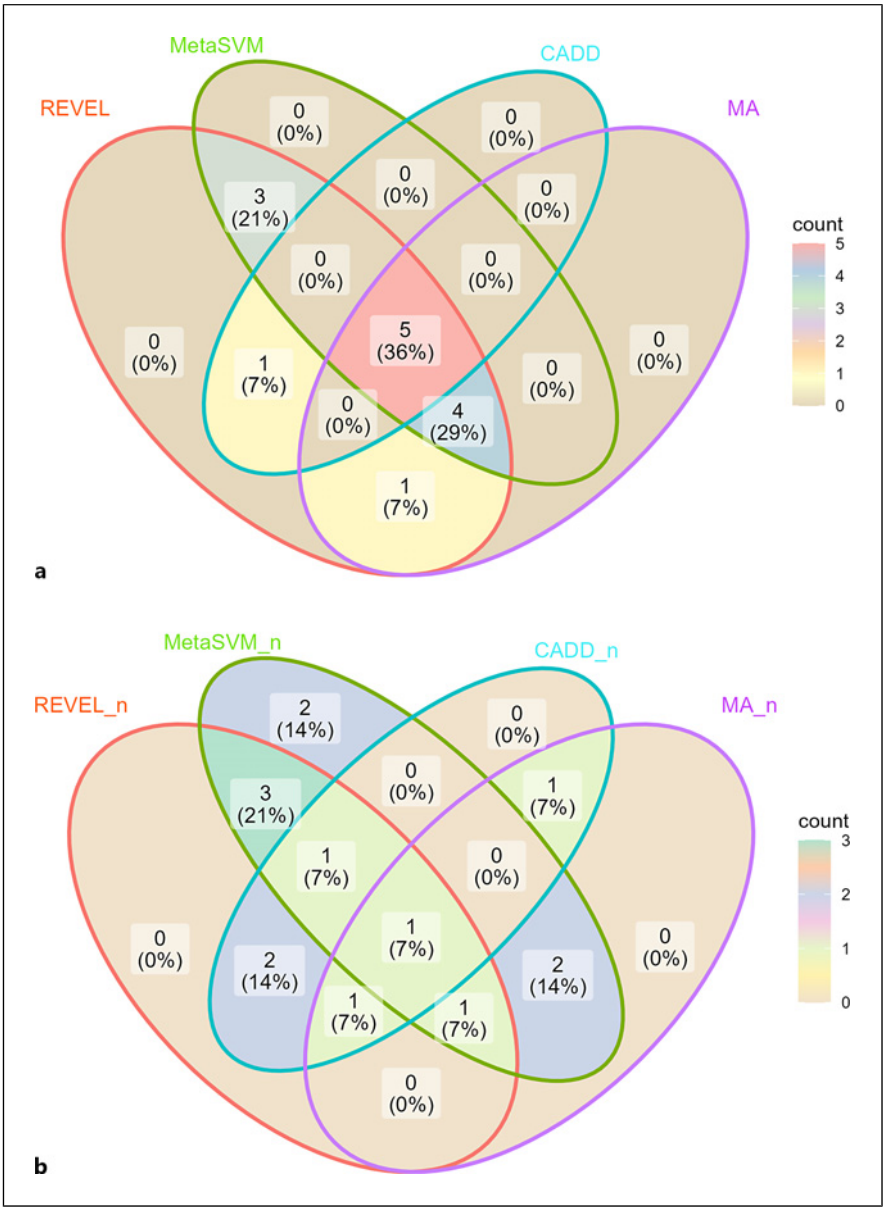

Fig. 1. Venn diagram of genes concordantly predicted after Tier 1 and 2 filtration process. **a** AI-driven predictors, i.e., REVEL, MetaSVM, and CADD in comparison to MutationAssessor (MA). **b** AI-driven predictors and MutationAssessor ranked and prioritised by gene pathogenic ratio (GPR).

Rare nsSNVs in ABCC1, ABCC4, ABCB6, ATP7B, CYP2D6, *and* CYP4F3 *Significantly Associated with Clinical Data, Including Poor Response to Therapy*

In *ABCC1*, fourteen unique rare nsSNVs (n = 8 singletons, i.e., variants observed in 1 patient in the dataset) were detected in 26 patients (24.5%). In 13 patients (12.4%), eleven nsSNVs (n = 7 singletons) were identified for *ABCC4* gene. At maximum, 1 patient carried two nsSNVs per gene for *ABCC1* and *ABCC4*. Patients harbouring any of a rare frequency variant in *ABCC1* had at least three times higher chance to develop a grade 3 tumours (OR = 3.1, 95% CI: [1.2–8.3], p = 0.021, Pearson χ^2), see Table 4.

On the contrary, variants in *ABCC4*, showed association suggesting their protective role in terms of a low tumour grade (G1 or G2) and other than invasive ductal carcinoma (IDC) histological tumour type in comparison to patients without alterations in *ABCC4* (p = 0.046, Pearson χ^2 and p = 0.045, Fisher's exact test, respectively). However, more than nine times higher risk of a poor response to therapy (OR = 9.64, 95% CI: [1.75–53.26], p = 0.007, Fisher's exact test, see Table 4) was observed in patients harbouring nsSNV in *ABCC4*.

For *ATP7B*, seven nsSNVs (n = 5 singletons) were detected in 10 patients (9.5%) and for *CYP2D6*, three

Table 1. Final set of top-ranked genes and number of variants (nsSNVs) in comparison to gene variant ratio (GVR) and gene pathogenic ratio (GPR) for each tool

Gene (total number of nsSNVs; novel nsSNVs)	Patients, *n*	GVR	GPR REVEL	GPR MetaSVM	GPR CADD	GPR MutationAssessor
ABCC1 (14;0)	26	2.15	1.54	1.54	1.85	1.54
ABCA1 (11;1)	19	1.05	0.29	0.48	0.67	0.38
TUBB1 (7;1)	13	2.00	0.86	1.43	2.00	2.00
CFTR (12;1)	16	1.96	1.79	1.79	1.63	1.14
CYP4F3 (6;1)	6	2.01	1.00	1.34	1.00	1.34
SLC22A1 (8;0)	26	4.20	1.57	2.62	3.67	3.15
ABCB6 (7;1)	13	2.41	2.06	2.06	2.06	1.03
ABCC4 (11;0)	13	3.86	1.40	1.05	NA	1.05
ATP7B (7;1)	10	1.16	1.00	0.83	0.83	0.66
SLC22A18 (5;1)	9	3.24	1.95	NA	2.59	2.59
ABCA9 (7;1)	7	1.10	0.63	0.63	0.63	0.63

Gene variant ratio (GVR) was calculated at gene level for the total number of nsSNVs detected in breast cancer patients of the study per gene and gene length. Gene pathogenic ratio (GPR) was calculated as the total number of pathogenic nsSNVs as predicted by indicated tool per gene and gene length. NA, not available.

Table 2. Total score for in silico prediction and PGx prioritised genes (PharmVar, ClinVar and PharmGKB)

Very High (++++)	High (+++)	Medium (++)	Low (+)
score: 8–7	score: 6–5	score: 4–3	score: ≤2 and/or PGx strong evidence
SLC22A1 Score: 8	*TUBB1* Score: 6 PGx	*CYP4F3* Score: 4	*DPYD* PGx strong evidence *VIP* gene
ABCC1 Score: 7 PGx ClinVar	*CFTR* Score: 6 PGx *VIP* gene	*ABCC4* Score: 4	Score: 2 *CYP2D6* PGx strong evidence *VIP* gene Score: 0
	ABCB6 Score: 5	*ABCA1* Score: 4	
	SLC22A18 Score: 5	*ATP7B* Score: 3	
		ABCA9 Score: 3	

nsSNVs (n = 2 singletons) were detected in 5 patients (4.8%). There was no more than 1 variant per patient in any of these genes. The presence of rare frequency nsSNVs in pharmacogene *CYP2D6* indicated a protective role as no variants were detected in patients with high grade (G3) tumours (p = 0.023, Fisher's exact test).

Patients with wild type *ATP7B* had significantly more frequently luminal A or luminal B subtype and additionally also a lower chance for a high grade G3 tumours than carriers of nsSNVs in *ATP7B* (p = 0.040 and p = 0.013, respectively, Fisher's exact Test). Overall, in 6 patients (5.7%) only singletons were detected for *CYP4F3* and seven nsSNVs (n = 4 singletons) in twelve (11.43%) patients were identified in *ABCB6* (see Table 4). For *CYP4F3* gene, nsSNVs were more frequent in patients with LB subtype of tumour in comparison to TNBC (p = 0.042, Fisher's exact test). Patients with premenopausal

Table 3. PharmVar and ClinVar "drug response" data

Gene	rsID/exonic function	HGVS (GRCh37)	ALFA[1]	CPIC clinical function	Ref	Alt	Patients, *n*
PharmVar							
DPYD	rs67376798*/nsSNV	NC_000001.10:g.97547947T>A	T = 0.995 A = 0.005	Decreased function	T	A	1
	rs56038477/ synonymous SNV	NC_000001.10:g.98039419C>T	C = 0.980 T = 0.019	Decreased function	C	T	4**
CYP2D6	rs1081003/ synonymous SNV	NC_000022.10:g.42525756G>A	G = 0.972 A = 0.028	Decreased function	G	A	5**
	rs5030656/in-frame deletion	NC_000022.10: g.42524178_42524180del	CTT = 0.022	Decreased function	CTT	–	2
	rs79292917/ synonymous SNV	NC_000022.10:g.42523854C>T	C = 0.995 T = 0.005	Decreased function	C	T	5
ClinVar							
CYP2D6	rs5030655/frameshift deletion	NC_000022.11:g.42129084del	A = 0.999 = 0.0017	NA	A	–	2
	rs35742686/ frameshift deletion	NC_000022.11:g.42128242del	T = 0.986 = 0.014	NA	T	–	4**
ABCC1	rs45511401/nsSNV	NC_000016.10:g.16079375G>T	G = 0.954 T = 0.0484	NA	G	T	7**

NA, not available. [1]ALPHA (allele frequency) release version: 20201027095038. *Clinical significance in "drug response" also by ClinVar. **2 patients identified with combination of variants rs56038477-rs35742686 and rs1081003-rs45511401.

status had more frequently nsSNVs in *ABCB6* ($p = 0.03$, Pearson χ^2) than postmenopausal ones. Moreover, carriage of nsSNVs in *ABCB6* was significantly associated with almost five times higher risk of poor response to neoadjuvant cytotoxic chemotherapy (OR = 4.89, 95% CI [1.04–22.96], $p = 0.046$, Fisher's exact test).

No association with DFS or OS was observed for these genes. Consequently, Cox regression analysis for DFS and OS adjusted to tumour grade (G3) and disease stage, showed no statistical significance.

No Association Has Been Found for Any of PGx Associated Exonic Variants in DPYD *and* CYP2D6

In analysis focused on all exonic variants (i.e., non-synonymous, synonymous, frameshift and inframe indels) in *DPYD* ($n = 5$) and *CYP2D6* ($n = 18$), no association has been found for any of the tested variables in comparison to patients without these variants.

Discussion

In the last two decades, a wide array of in silico tools have been developed to predict the functional implications of variants, utilising various approaches such as sequence and structure analysis, evolutionary conservation and machine learning. However, the in silico-driven classification for drug response, role of rare genetic variants, and even prognosis in cancer remains a significant challenge [2, 6]. In our study, the approach which integrates the AI-driven prediction and data mining for estimation of functional consequences of rare nsSNVs to the overall decision-making process have been explored. To our knowledge, this is the first time when gene pathogenic ratio and exclusion criteria for assumed benign genes with combination of tools dominated by AI-driven prediction were used as point of departure in prioritising pathogenic genes in Czech breast cancer patients.

When reviewing the in silico results, one of the top-ranked genes for both, pathogenic and benign functional consequence were *ABCA13, NCOR2*, and *ABCA4*. These were excluded in the Tier 1 process due to presumably false positive predictions with all four algorithms (REVEL, MetaSVM, CADD, and MutationAssessor). The high number of patients with variants in *ABCA13* and *ABCA4* (35 and 45 in 105 patients, respectively) further suggested a non-pathogenic nature of these genes. In our previously published pilot study [12] *NCOR2*, *ABCA13*, *RPTOR*, *ABCA4*, and *CIT* were, indeed, identified as the top five polymorphic genes in breast cancer patients.

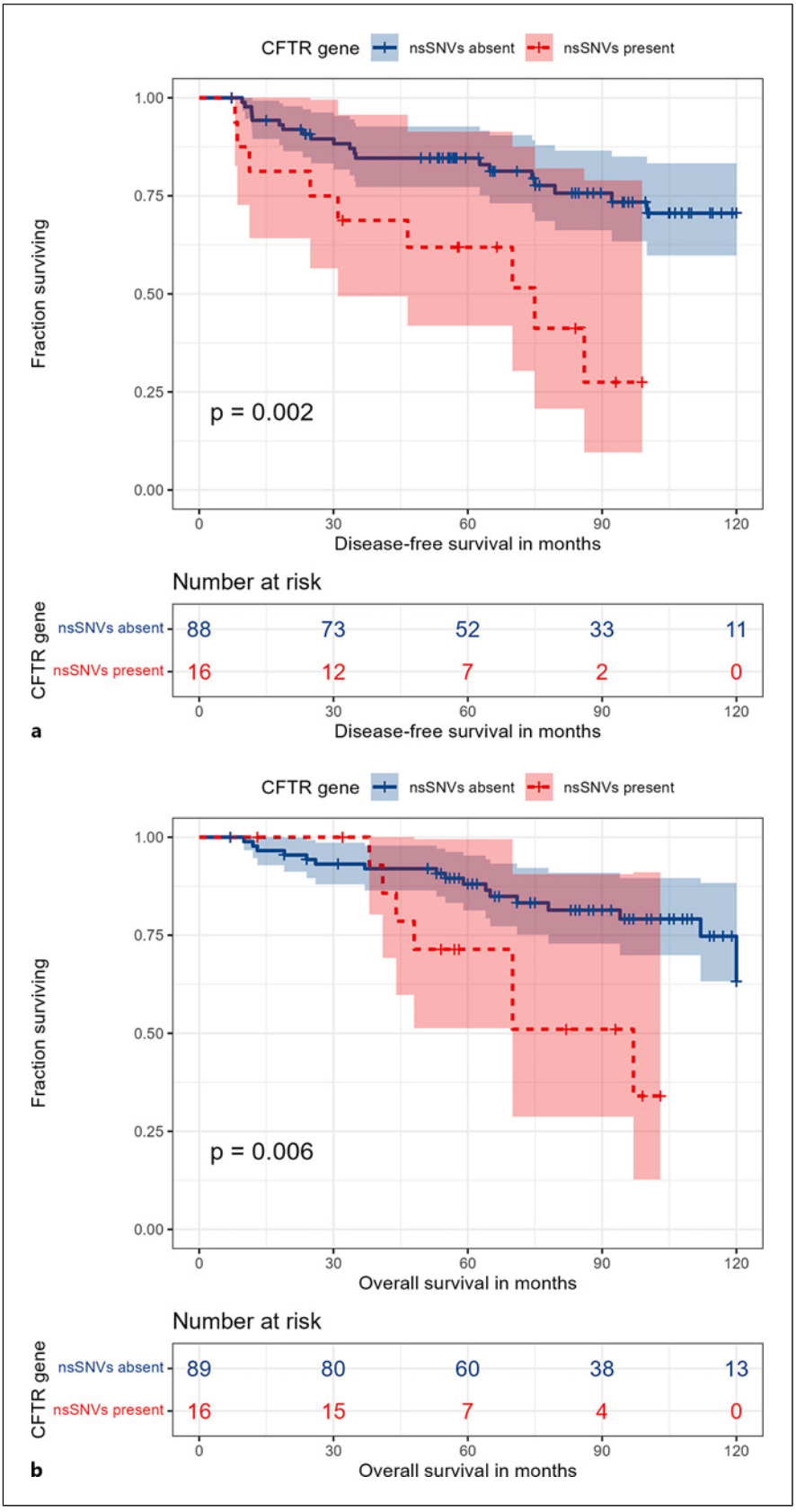

Fig. 2. Kaplan-Meier plots for associations of *CFTR* nsSNVs with DFS (**a**) and OS (**b**) of patients. Red line represents patients ($n = 16$) carrying nsSNVs in *CFTR* while the blue line represents patients without such variants. The shaded areas denote 95% confidence intervals. Significance was evaluated by the log-rank test at 5% level.

Table 4. Association of rare nsSNVs to clinicopathological data of breast cancer patients

Gene	nsSNV	Associated variable (patients, *N*)		OR	95% (CI)	*p* value
ABCC1		pGrade 1–2	pGrade 3			
	Absent	41	34			
	Present	7	18	3.10	(1.16–8.30)	0.021
ABCC4		pGrade 1–2	pGrade 3			
	Absent	39	49			
	Present	9	3	0.27	(0.07–1.05)	0.046
		Type (other)	IDC			
	Absent	13	79			
	Present	5	8	0.26	(0.07–0.93)	0.045
		Poor response	Good response			
	Absent	14	45			
	Present	6	2	9.64	(1.75–3.26)	0.007
CYP2D6		pGrade 1–2	pGrade 3			
	Absent	43	52			
	Present	5	0	0.45	(0.36–0.56)	0.023
ATP7B		pGrade 1–2	pGrade 3			
	Absent	40	51			
	Present	8	1	0.10	(0.01–0.82)	0.013
		LA or LB	TNBC			
	absent	39	57			
	present	7	2	0.20	(0.04–0.99)	0.040
CYP4F3		LB	TNBC			
	Absent	26	58			
	Present	4	1	0.11	(0.01–1.05)	0.042
ABCB6		Poor response	Good response			
	Absent	15	44			
	Present	5	3	4.89	(1.04–22.96)	0.045
		Pre-menopause	Post-menopause			
	Absent	39	54			
	Present	9	3	0.24	(0.06–0.95)	0.030

nsSNV, non-synonymous single nucleotide variant; OR, odds ratio; CI, confidential interval; IDC, intraductal histological type; LA/LB, luminal A or B molecular subtype; TNBC, triple negative breast carcinoma molecular subtype.

Adjustment of predictions by gene length further refined the filtration process and allowed for inclusion of *DPYD* – VIP pharmacogene to the final set of genes. Data mining in PGx databases also revealed a significant number of patients in our study (n = 28, 26.6%) harbouring PGx relevant exonic variants, *in DPYD*, *CYP2D6*, and *ABCC1*. In silico functional predictions by the four tools were available only for two nsSNVs (rs67376798 in *DPYD* and rs45511401 in *ABCC1*) out of eight identified PGx variants and these were considered as pathogenic in accordance with the databases. *DPYD* encodes for the dihydropyrimidine dehydrogenase enzyme, responsible for pharmacokinetics and associated toxicity of fluoropyrimidines. In heterozygous patients, a fluoropyrimidine-related grade 3 or 4 toxicity was reported for rs67376798 or rs56038477 demanding for a 50% reduction of a starting dose [28, 29]. In our study, no association to any of tested clinical parameters was found for nsSNVs or exonic variants in *DPYD*. On the contrary, while no association has been found for exonic variants in *CYP2D6* either, for a group of patients with presence of any nsSNVs, an association with tumour grade, i.e., 1 and 2 (p = 0.023) was demonstrated. The study also showed that breast cancer patients who had one of the rare variants in the *ABCC1* were at least three times more likely to develop highly aggressive grade 3 tumours. Published data documents that nsSNV rs45511401 in *ABCC1* was significantly (p = 0.012) associated with increased risk of cardiotoxicity in patients with Non-Hodgkin lymphoma treated with adriamycin [30]. Anthracyclines such as adriamycin belong to a common oncotherapy treatment in breast cancer

patients and occurrence of serious adverse side effects can lead to a discontinuation and failure of the therapy.

Overall, our approach provided fast and effective way to prioritise seven (*ABCC1, ABCC4, ABCB6, ATP7B, CFTR, CYP2D6*, and *CYP4F3*) out of twelve genes showing statistical significance to clinical characteristics or to oncotherapy prognosis (*CFTR*) in breast cancer patients. Mutations in *CFTR* were as of late connected to an elevated susceptibility to diverse forms of malignancies, indicating that *CFTR* function is even beyond that of an ion channel, and instead has ability to regulate numerous signalling pathways. The association of *CFTR* expression levels with patient survival and the possible use of *CFTR* as a prognostic biomarker in human malignancies such as breast cancer have been reported and reviewed by several authors [31–34]. Generally, germline mutations in *CFTR* cause cystic fibrosis (CF) – the most common life-limiting recessive genetic disease among Caucasians and *CFTR* is mainly known for its PGx value in CF. Of the total 12 *CFTR* variants observed in our patients with breast cancer, eight variants are also deposited in the Clinical and Functional TRanslation of *CFTR* (*CFTR*2) database. These complex variants are regarded either as non-causing the CF or possessing varying clinical consequences (e.g., pancreatic insufficiency) depending upon the co-occurring variant(s) situated in *cis/trans* position and affecting CFTR protein production. For the present study, it suggests potential for a functional activity of detected variants. From the perspective of gene pathogenic ratio, *CFTR* was in a range for top 10 pathogenic genes and several other studies showed that the length of the transcript correlates with functional activity of co-expressed genes or proteins related to cancer [35, 36]. *CFTR* having a transcript length of 6,132 bp and protein size of 1,480 amino acids can be considered a longer gene, denoting that mechanism of action in drug resistance or the malignant process may be also dependable on other genes. We have hitherto demonstrated that *CFTR*, previously known as *ABCC7/MRP7* prevailed among the most dysregulated genes on the transcript level in breast, colorectal and pancreatic cancers [33].

ATP binding cassette transporters (ABCs) constitute a large family of active transporters translocating substances, including anticancer drugs, across extracellular and intracellular membranes affecting drug efflux related to the development of cancer cell chemoresistance [37, 38]. Here, we identified four out of seven clinically associated genes, which belong to the ABC family. As an example, transporters from the subfamily *ABCC* reduce the intracellular concentration of monophosphorylated nucleoside oncotherapy drugs, e.g., 5-FU [34]. For *ABCC4* and in addition for *ABCB6*, we demonstrated a significantly enhanced risk of poor response to NACT in patients carrying rare nsSNVs.

The other prioritised and clinically associated genes, *ATP7B* and later *CYP4F3* or *CYP2D6* are known to be involved in metabolism of oncotherapy drugs and in the resistance to cisplatin-based chemotherapy in several cancers [39, 40]. Our findings suggest that clinical importance of these genes may not be limited to platinum derivatives, but also other structurally and mechanistically diverse drugs. The link between genetic variability and breast cancer, however, awaits confirmation in independent studies on larger populations and follow-up functional analyses.

Notwithstanding all above observations, our study has some limitations that need to be considered. At first, the biological role of *CFTR* in cancer progression and/or drug response and clinical associations with breast cancer characteristics observed for *ABCC1, ABCC4, ABCB6, ATP7B, CYP2D6*, and *CYP4F3* require attention by follow-up studies using robust clinical data. We are aware that our present results may not be independent of other unknown clinical or molecular factors associating with patient's prognosis even more strikingly. Among others, confounding somatic changes in the target tissue, both genetic and epigenetic may serve as an example, which could not be controlled by our study. Apart from the modest sample size, another limitation of our study is the gene panel with target enrichment of ABC transporters, which may have compromised the advantage of a calculated gene pathogenic ratio. On the other hand, our study has clear benefits in the focus of the most relevant pharmacogenes and oncogenes and assessment of ethnically homogeneous population, which contributes to the concept of diversity in medicine and specifically precision oncology. Moreover, the study included patients with long-term (120 months) and complete clinical follow-up, which makes the findings robust and suitable for subsequent meta-analyses and hypothesis-generating screens.

Taken together, our present study shows that the AI-driven in silico analysis, connected with PGx data mining, and robust clinical covariates deliver statistically significant associations of rare nsSNVs in pharmacogenes with important parameters for patient prognosis e.g., treatment response and survival. Our study also indicates that *CFTR* may be connected with key processes influencing resistance to oncotherapy or progression of the disease.

Acknowledgments

We thank the National Center for Medical Genomics (LM2015091) for providing allelic frequencies in the ethnically matched population for comparison (project CZ.02.1.01/0.0/0.0/16_013/0001634).

Statement of Ethics

This study protocol was reviewed and approved by the Ethical Commission of the National Institute of Public Health in Prague (Approval Nos. 9799-4, 15-25618A and 17-28470A). All patients were informed about the study, and only those who agreed and signed an informed consent form were recruited.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

The study was supported by the Czech Health Research Council Grant No. NV22-08-00281 to V.H.

Author Contributions

Conceptualisation for this manuscript was undertaken by M.K. and P.S. The methodology of the study was prepared by M.K. and V.H. The use of chosen software was carried out by M.K. Formal analysis of the study data was conducted by M.K., V.H. and P.S. Resources were provided by R.K., K.R., and J.G. Data curation was performed by M.K., R.K., K.R., and J.G. The original draft was written by M.K., V.H., and P.S. All authors participated in writing review and editing. Visualisation of data was performed by M.K. The research was conducted under the supervision of P.S. The project administration was managed by P.S. and M.K. Funding acquisition was attributed to V.H.

Data Availability Statement

The sequencing data that support the findings of this study are openly available in Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra), under accession No. PRJNA510917 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA510917). The code developed to generate the results in this study is available upon request. Further enquiries can be directed to the corresponding author.

References

1 Malhotra GK, Zhao X, Band H, Band V. Histological, molecular and functional subtypes of breast cancers. Cancer Biol Ther. 201010(10):955–60. https://doi.org/10.4161/cbt.10.10.13879
2 Schärfe CPI, Tremmel R, Schwab M, Kohlbacher O, Marks DS. Genetic variation in human drug-related genes. Genome Med. 2017;9(1):117. Available from: https://doi.org/10.1186/s13073-017-0502-5
3 Kozyra M, Ingelman-Sundberg M, Lauschke VM. Rare genetic variants in cellular transporters, metabolic enzymes, and nuclear receptors can be important determinants of interindividual differences in drug response. Genet Med. 2017;19(1):20–9. https://doi.org/10.1038/gim.2016.33
4 Nelson MR, Wegmann D, Ehm MG, Kessner D, St. Jean P, Verzilli C, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science. 2012;337(6090):100–4. https://doi.org/10.1126/science.1217876
5 Gerek NZ, Liu L, Gerold K, Biparva P, Thomas ED, Kumar S. Evolutionary Diagnosis of non-synonymous variants involved in differential drug response. BMC Med Genomics. 2015;8(Suppl 1):S6. https://doi.org/10.1186/1755-8794-8-S1-S6
6 Tremmel R, Pirmann S, Zhou Y, Lauschke VM. Translating pharmacogenomic sequencing data into drug response predictions: how to interpret variants of unknown significance. Br J Clin Pharmacol [Internet]. [cited 2023 Oct 22];n/a(n/a). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/bcp.15915
7 Gunning AC, Fryer V, Fasham J, Crosby AH, Ellard S, Baple EL, et al. Assessing performance of pathogenicity predictors using clinically relevant variant datasets. J Med Genet. 2021;58(8):547–55. https://doi.org/10.1136/jmedgenet-2020-107003
8 Tian Y, Pesaran T, Chamberlin A, Fenwick RB, Li S, Gau CL, et al. REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification. Sci Rep. 2019;9(1):12752. https://doi.org/10.1038/s41598-019-49224-8
9 Zhou Y, Mkrtchian S, Kumondai M, Hiratsuka M, Lauschke VM. An optimized prediction framework to assess the functional impact of pharmacogenetic variants. Pharmacogenomics J. 2019;19(2):115–126. https://doi.org/10.1038/s41397-018-0044-2
10 Yazar M, Ozbek P. Assessment of 13 in silico pathogenicity methods on cancer-related variants. Comput Biol Med [Internet]. 2022;145:105434. https://doi.org/10.1016/j.compbiomed.2022.105434
11 Cubuk C, Garrett A, Choi S, King L, Loveday C, Torr B, et al. Clinical likelihood ratios and balanced accuracy for 44 in silico tools against multiple large-scale functional assays of cancer susceptibility genes. Genet Med. 2021;23(11):2096–104. https://doi.org/10.1038/s41436-021-01265-z
12 Hlavac V, Kovacova M, Elsnerova K, Brynychova V, Kozevnikovova R, Raus K, et al. Use of germline genetic variability for prediction of chemoresistance and prognosis of breast cancer patients. Cancers. 2018;10(12):511. https://doi.org/10.3390/cancers10120511
13 Schwartz LH, Litière S, de Vries E, Ford R, Gwyther S, Mandrekar S, et al. RECIST 1.1-Update and clarification: from the RECIST committee. Eur J Cancer. 2016;62:132–7. https://doi.org/10.1016/j.ejca.2016.03.081
14 Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–91. https://doi.org/10.1038/nature19057
15 ALFA: allele frequency aggregator [Internet]. [cited 2023 Apr 23]. Available from: https://www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/
16 Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet. 2016;99(4):877–85. https://doi.org/10.1016/j.ajhg.2016.08.016
17 Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24(8):2125–37. https://doi.org/10.1093/hmg/ddu733

18 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. Nat Genet. 2014; 46(3):310–5. Available from: http://search.ebscohost.com/login.aspx?authtype=shib&custid=s1240919&profile=eds
19 Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39(17):e118. Available from: https://doi.org/10.1093/nar/gkr407
20 Ghosh R, Oak N, Plon SE. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Biol. 2017;18(1):225. https://doi.org/10.1186/s13059-017-1353-5
21 Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14(Suppl 3):S3:https://doi.org/10.1186/1471-2164-14-S3-S3
22 Goecks J, Nekrutenko A, Taylor J; Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11(8):R86. https://doi.org/10.1186/gb-2010-11-8-r86
23 PharmVar [Internet]. [cited 2023 Apr 22]. Available from: https://www.pharmvar.org/
24 Representation of clinical significance in ClinVar and other variation resources at NCBI [Internet]. [cited 2023 Apr 22]. Available from: https://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/
25 Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014;95(1): 5–23. https://doi.org/10.1016/j.ajhg.2014.06.009
26 Lee NY, Hum M, Zihara S, Wang L, Myint MK, Lim DWT, et al. Landscape of germline pathogenic variants in patients with dual primary breast and lung cancer. Hum Genomics. 2023;17(1):66. https://doi.org/10.1186/s40246-023-00510-7
27 Cheang MCU, Chia SK, Voduc D, Gao D, Leung S, Snider J, et al. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. J Natl Cancer Inst. 2009; 101(10):736–50. https://doi.org/10.1093/jnci/djp082
28 Knikman JE, Gelderblom H, Beijnen JH, Cats A, Guchelaar H, Henricks LM. Individualized dosing of fluoropyrimidine-based chemotherapy to prevent severe fluoropyrimidine-related toxicity: what are the options? Clin Pharmacol Ther. 2021;109(3):591–604. https://doi.org/10.1002/cpt.2069
29 Carr DF, Turner RM, Pirmohamed M. Pharmacogenomics of anticancer drugs: personalising the choice and dose to manage drug response. Br J Clin Pharmacol. 2021; 87(2):237–55. https://doi.org/10.1111/bcp.14407
30 Wojnowski L, Kulle B, Schirmer M, Schlüter G, Schmidt A, Rosenberger A, et al. NAD(P)H oxidase and multidrug resistance protein genetic polymorphisms are associated with doxorubicin-induced cardiotoxicity. Circulation. 2005;112(24):3754–62. https://doi.org/10.1161/CIRCULATIONAHA.105.576850
31 Zhang J, Wang Y, Jiang X, Chan HC. Cystic fibrosis transmembrane conductance regulator: emerging regulator of cancer. Cell Mol Life Sci. 2018;75(10):1737–56. https://doi.org/10.1007/s00018-018-2755-6
32 FitzMaurice TS, Nazareth DS. Incidence of breast cancer in people with cystic fibrosis: a cause for concern? J Cyst Fibros [Internet]. 2021. [cited 2022 Aug 6];0(0). Available from: https://www.cysticfibrosisjournal.com/article/S1569-1993(21)02158-5/fulltext#relatedArticles
33 Dvorak P, Pesta M, Soucek P. ABC gene expression profiles have clinical importance and possibly form a new hallmark of cancer. Tumour Biol. 2017;39(5): 1010428317699800. https://doi.org/10.1177/1010428317699800
34 Fukuda Y, Schuetz JD. ABC transporters and their role in nucleoside and nucleotide drug resistance. Biochem Pharmacol. 2012;83(8): 1073–83. https://doi.org/10.1016/j.bcp.2011.12.042
35 Lopes I, Altab G, Raina P, de Magalhães JP. Gene size matters: an analysis of gene length in the human genome. Front Genet. 2021;12: 559998. https://doi.org/10.3389/fgene.2021.559998
36 Sahakyan AB, Balasubramanian S. Long genes and genes with multiple splice variants are enriched in pathways linked to cancer and other multigenic diseases. BMC Genomics. 2016; 17(1):225. https://doi.org/10.1186/s12864-016-2582-9
37 Fletcher JI, Williams RT, Henderson MJ, Norris MD, Haber M. ABC transporters as mediators of drug resistance and contributors to cancer cell biology. Drug Resist Updat. 2016;26:1–9. https://doi.org/10.1016/j.drup.2016.03.001
38 Chen Z, Shi T, Zhang L, Zhu P, Deng M, Huang C, et al. Mammalian drug efflux transporters of the ATP binding cassette (ABC) family in multidrug resistance: a review of the past decade. Cancer Lett. 2016;370(1):153–64. https://doi.org/10.1016/j.canlet.2015.10.010
39 Kanzaki A, Toi M, Neamati N, Miyashita H, Oubu M, Nakayama K, et al. Copper-transporting P-type adenosine triphosphatase (ATP7B) is expressed in human breast carcinoma. Jpn J Cancer Res. 2002;93(1): 70–7. https://doi.org/10.1111/j.1349-7006.2002.tb01202.x
40 Nakayama K, Kanzaki A, Ogawa K, Miyazaki K, Neamati N, Takebayashi Y. Copper-transporting P-type adenosine triphosphatase (ATP7B) as a cisplatin based chemoresistance marker in ovarian carcinoma: comparative analysis with expression of MDR1, MRP1, MRP2, LRP and BCRP. Int J Cancer. 2002;101(5):488–95. https://doi.org/10.1002/ijc.10608