universität wien

DIPLOMARBEIT / DIPLOMA THESIS

Titel der Diplomarbeit / Title of the Diploma Thesis

„Exploiting open data for individualized medicine“

verfasst von / submitted by

Astrid Kempf

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Magistra der Pharmazie (Mag.pharm.)

Wien, 2022 / Vienna, 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 449
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Diplomstudium Pharmazie
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Gerhard Ecker
Mitbetreut von / Co-Supervisor:	Dipl.-Ing. (FH) Dr. Daniela Digles

Abstract

ABC transporters are a superfamily of transmembrane proteins with more than 250 members. Through hydrolysis of ATP the generated energy is used to move specific substrates across the cell membranes. In humans 48 ABC transporters have been classified. Mutations cause several diseases affect the immune system and change phenotypes. They also play an important role in cancer therapy as they can lead to multidrug resistance of cancer cells. The aim of this thesis was to establish a KNIME workflow which allows the identification of disease relevant mutations at ABC transporters.

In the workflow all human ABC transporters where examined to detect their variants of the ClinVar database. 15.763 mutations were found for all transporters, 3.266 were cited as pathogenic.

Further, cystic fibrosis and multidrug resistance protein 1 were analyzed in more detail to verify the workflow. Over 2.000 mutations are known from the CFTR gene, about 300 of them are pathogenic. In the workflow 2.749 variants were found and 669 were cited as pathogenic. The total count of mutations found are nearly equal but the fact that the pathogenic variants are double as much could be explained as previous papers often recommend to the same sources, two databases, and the possible lack of information due to limited updates.

In literature about 50 SNP’s were found for the ABCB1 gene. The workflow presented 602 SNP’s. This could be explained as the papers were published in 2013 and 2020 but refer to sources from 2001, 2003 and 2010. Further, such studies often include just one distinct population group and would not include all SNP’s since these vary highly within an ethnic group.

Kurzzusammenfassung

ABC Transporter bilden eine Superfamilie mit mehr als 250 Mitglieder. Die durch Hydrolyse von ATP zu ADP generierte Energie wird verwendet um bestimmte Substrate durch die Zellmembran zu transportieren. Im Menschen sind 48 ABC-Transporter bekannt. Mutationen verursachen diverse Krankheiten, beeinflussen das Immunsystem und verändern Phänotypen. Sie spielen eine wichtige Rolle in der Krebstherapie, da sie zu einer Multiresistenz von Krebszellen führen können. Ziel der Diplomarbeit war daher, einen KNIME Workflow zu entwickeln der es erlaubt, Krankheitsrelevante Mutationen in ABC-Transportern zu idnetifizieren.

Im Workflow wurden alle 48 ABC Transporter auf ihre Anzahl an Mutationen in der ClinVar Datanbank untersucht. 15.763 Mutationen wurden gefunden, 3.266 davon waren pathogen.

Des Weiteren wurden Cystische Fibrose und Multidrug Resistentes Protein 1 näher bearbeitet um den den Workflow zu verifizieren. Über 2.000 Mutationen sind vom CFTR Gen bekannt, rund 300 davon sind pathogen. Im Workflow wurden 2.749 Variationen gefunden, 669 davon pathogen. Dass die Gesamtzahl an gefundenen Mutationen ähnlich ist, aber doppelt so viele pathogene Varianten gefunden wurden, könnte dadurch erklärt werden, dass sich Publikationen oft auf dieselben Quellen, zwei Datanbanken, beziehen und diese durch eingeschränkte Aktualisierungen möglicherweise ein Informationsdefizit aufweisen.

In der Literatur sind circa 50 SNP’s für das ABCB1 Gen bekannt. Der Workflow resultierte in 602 SNP’s. Die markante Abweichung könnte daher stammen, dass die Publikationen 2013 und 2020 veröffentlicht wurden, sich aber auf Quellen aus 2001, 2003 und 2010 beziehen. Weiters inkludieren diese Studien häufig nur bestimmte Personengruppen und enthalten daher nicht alle SNP’s, da diese stark innerhalb einer Ethnie variieren.

Acknowledgement

First, I want to thank Univ.-Prof. Mag. Dr. Gerhard Ecker, who initiated this diploma thesis and supported me with his advice and never ending patience.

I am grateful to Dipl.-Ing. (FH) Dr. Daniele Digles, who helped with my work in guiding me through the Knime nodes and discussing patiently the resulting data.

Furthermore, I want to thank Jana Gurinova who kept motivating me throughout this whole time.

Finally, my special thanks to my parents, Mag. Christine and Ing. Werner KEMPF, and my partner Recep Ekilmis. They always have shown great confidence in me. Without their spiritual, emotional and moral support I would never have come this far.

Abbreviations

ABC	ATP-binding cassette
ADP	Adenosine diphosphate
ATP	Adenosine triphosphate
BCRP	breast cancer resistance protein
CF	cystic fibrosis
CFTR	cystic fibrosis transmembrane conductance regulator
ClinVar	Database of variants
dbSNP	Database of Single Nucleotide Polymorphism
HGMD	Human Genomic Mutation Database
HuVarBase	Human Variant Database
KNIME	Konstanz Information Miner
M1	O-desmethyltramadol
MDR	multidrug resistance
mRNA	messenger RNA
MRP	multidrug resistance associated protein
NBD	nucleotide-binding domains
OMIM	Online Mendelian Inheritance in Man
P-gp	P-glycoprotein
RNA	Ribonucleic acid
rowID	Row identifier
SNP	single nucleotide polymorphism
SUR	sulphonylreceptor
TAP1	transporter associated with antigen processing 1
TAP2	transporter associated with antigen processing 2
TMD	transmembrane domains
URL	Uniform Resource Locators

Table of content

A. Introduction

A.1. ABC transporter

In living organisms four different classes of membrane-bound transport proteins are known: ion channels, transporters, aquaporins and ATP-powered pumps, like ATP-binding cassette (ABC) transporters [1]. ABC transporters exist in prokaryotic as well as in eukaryotic cells, in eukaryotic all are exporters [2].

Typically they consist of four core domains, two transmembrane domains (TMD) and two nucleotide-binding domains (NBD). The TMD are hydrophobic alpha-helices fixed in the membrane bilayer, normally six times and provide substrate specificity and translocation across the lipid membrane [1]. The NBD are located in the intracellular cytoplasm. By hydrolysing ATP to ADP the released energy is used to transport substrates across the membrane against their concentration gradient, thereby ABC transporters are active transporters [2]. The substrates vary from metal ions, peptides, amino acids, sugars, hydrophobic compounds and metabolites over drugs and toxins [1]. They are located in the plasma membrane as well as in cell organelles, such as peroxisomes, lysosomes and endosomes [2].

In humans so far 48 ABC transporters have been classified. Not all are active transporters, the cystic fibrosis transmembrane conductance regulator (CFTR) is a chloride ion channel and the sulphonylreceptor (SUR) is a regulator for an ion channel. They are encoded by 48 genes, 22 pseudogenes have been identified yet [3]. Further they are clustered in seven subfamilies termed ABCA through ABCG. Most of them are full transporters where their two TMD and two NBD are encoded by a single polypeptide. Half transporters consist of one TMD and NBD and are considered to homo- or heterodimerize to form a functional transporter [2]. For example the ABCG5 and ABCG8 are assembled into a heterodimer.

Mutations in the ABC transporter genes are known to cause genetic diseases, for example the cystic fibrosis or tangier disease. Further the immune system can be affected through transporter defects caused from polymorphism like TAP 1 and TAP 2. Moreover, mutations are associated with altered drug effectiveness or severe drug-induced adverse events as they are affecting the normal functioning of these transporter proteins resulting in an increased or decreased drug level [4]. Especially in cancer cells mutations of these transporters lead to decreased drug levels and further multidrug resistance (MDR) may result. Multidrug resistance describes the phenomenon of resistance against multiple, structurally unrelated compounds. For example, the P-glycoprotein transporter (P-gp) encoded from MDR1 or ABCB1 is a well known member and resistance against compounds used in chemotherapy is often developed [5]. Other important members are encoded from subfamily C, so called multidrug resistance associated proteins (MRP). Nowadays 9 of the 13 representatives are known to be multidrug resistance proteins (MRP's 1 - 9) [6]. Overcoming MDR is still a problem as

developing these transporters as therapeutic targets have been unsuccessful. Many inhibitors were identified, however, their effects could not be verified in patients yet [5].

A.2. Mutations

A mutation is a change in the nucleotide sequence of a genome and is the driving force of evolution. They happen naturally, often these spontaneous mutations occur while replication, as well as induced through exposure to mutagens, such as ultraviolet light. They range from single-basepair alterations to mega-basepair deletions, insertions, duplications and inversions [7]. Most of them are so called point mutations. Other mutation classes would be chromosomal mutations, where a region of a whole chromosome is duplicated, deleted, translocated or via inversion flipped and reinserted. A third class are copy number variations which are further distinguished in copy number gain or copy number loss. Thereby additional gene copies are inserted or decreased in the human genome.

With regard to point mutation the most common variation is substitution. Other point mutations are insertion, whereas a single or more nucleotides are added and deletion of one or more nucleotides. Both may result in frameshift mutations if not a whole codon is inserted or deleted. Substitution can be either through transition between two purine or pyrimidine bases or transversion between a purine base and a pyrimidine base [8].

If just one basepair is substituted it is a single nucleotide polymorphism (SNP). They are widely distributed throughout the human genome with an estimated frequency of about 1/1000 basepairs [9]. Therefore there are about 4 to 5 million SNP’s in a person’s genome. Three main effects can occur: First, a silent mutation where the altered codon corresponds to the same amino acid. Second, the base substitution can be a missense mutation where the altered codon corresponds to a different amino acid. Or third, the base substitution can be a nonsense mutation where the altered codon corresponds to a stop signal. Further, SNP may influence gene expression, stability of messenger RNA (mRNA) conformation and subcellular localization of mRNAs and/or proteins [9]. SNP’s can be used as biological markers for diseases or predict the risk for developing particular diseases, such as cystic fibrosis. They may also forecast the individual response rate to certain drugs and treatment.

A.3. Knime

The Konstanz Information Miner (KNIME) is free open source software for data analyzing. Its concept is based on the idea of building a visual workflow out of blocks by using various components for machine learning and data mining, so called nodes. The graphical interface works via drag and drop so even beginners can easily create workflows without the need of programming or coding [10]. Since it was released in 2006 it offers several tools for the analysis of chemical and pharmacoinformatics data [11] and it is continuously updated and integrating new developments. Additionally, it provides self-paced courses of learning to work with KNIME and a collection of guides for KNIME Analytics Platform and KNIME server [12]. Furthermore the so called KNIME-Hub community help provides fast help and guidance by other KNIME users.

Nowadays those data analyzing platforms are essential to search through publicly available data and make state of the art assessments as well as handle these to achieve new assumptions, because manually performed literature search to collect data is consuming a lot of time human resources.

A.4. Aim of the work

The aim of this thesis was to establish a KNIME workflow which allows the identification of disease relevant mutations at ABC transporters. Therefore open data was used. Open data is freely available data which can be used and republished without any copyright or patents restrictions. There are many kinds of open data such as culture, finance, statistics, weather, environment, government, science and many more.

Open science data is often collected in a large number of databases, preferably available by downloading over the internet for further analyzing. As mentioned above manually performed searches for collecting data is limited through human resources. Compared to this, a workflow allows advantages by shortening this step. The workflow, once established, performs independent search queries, collects data and represents it for further analyzing. This saves time, needs less human resources and search queries can be easily modified and reproduced for using other databases.

Open data is an important input source for achieving new assumptions and sum up knowledge of disease relevant SNP's. This knowledge can be further used in personalized medicine in screening for genetic diseases. Personalized medicine is based on each patient's unique genetic code and predicts how a specific mutation might affect a person's risk of getting a certain disease or treatment success. It holds the idea of overcoming limitations of traditional medicine by increasing health impact of existing treatments by improving the matching process between patients and different available treatments and moreover by making predictions about an individual patient risk of serious side effects.

B. Methods and Results

B.1. Prework for the Workflow

As mentioned above, in every of the 48 human ABC transporter genes mutations occur and can lead to severe diseases, phenotypes or mutate the transporter protein to form a MRP.

The whole workflow (Fig.1) consists of 31 nodes, whereas they are further separated in four metanodes for better visual presentation.

In this workflow all transporter genes were examined to detect their variations through a distinct database. To select a database with a high number of variations a brief manually prepared search of the following databases was performed: ClinVar, HuVarBase, UniProt, OMIM and HGMD. Therefore the total count of variants of the CFTR gene for every database was examined as it causes one of the most cited ABC transporter diseases. The following results were accessed on February 08 2022 and assembled into table 1.

dbSNP provided most variant entries with 65.132, the Ensemble Database reported 61.365 entries about the CFTR gene, the ClinVar Database hold 2.749 variants, the Human Variants Database (HuVarBase) showed 764 entries. In UniProt only 210 variants were found, due to only definitely pathogenic variants were disposed and the Online Mendelian Inheritance in Man (OMIM) comprised just 138 entries.

The public version of Human Genomic Mutation Database (HGMD) is freely available to registered users from academic institutions and non-profit organisations. It is updated twice annually [13] and as a registration is necessary, it may lead to an inoperable workflow over a longer period of time. Therefore it was excluded from my research.

Three databases offered an easy access to filter pathogenic variants, dBSNP included 753, ClinVar 669 and UniProt 210 pathogenic variants. Although dbSNP and Ensembl provides a much bigger amount of variant entries, the needed processing power and expenditure of time to perform the workflow would be tremendous. Therefore ClinVar was used for building the workflow as the number of pathogenic identified variants was almost as high as those displayed in dbSNP. ClinVar is a freely accessible, public archive of reports of the relationships among human variations and their phenotypes and is hosted by the National Center of Biotechnology (NCBI). The submissions contain human variants which were found in patient samples and assertions regarding their clinical significance are made if reviews are available [14]. The xml data is updated weekly and archived monthly [15]. The variants have unique identifier (UID) in this work shortened to the term ClinVar ID’s.

The workflow was made with KNIME Analytics Platform version 4.5.1 and performed on February 10 2022 last time.

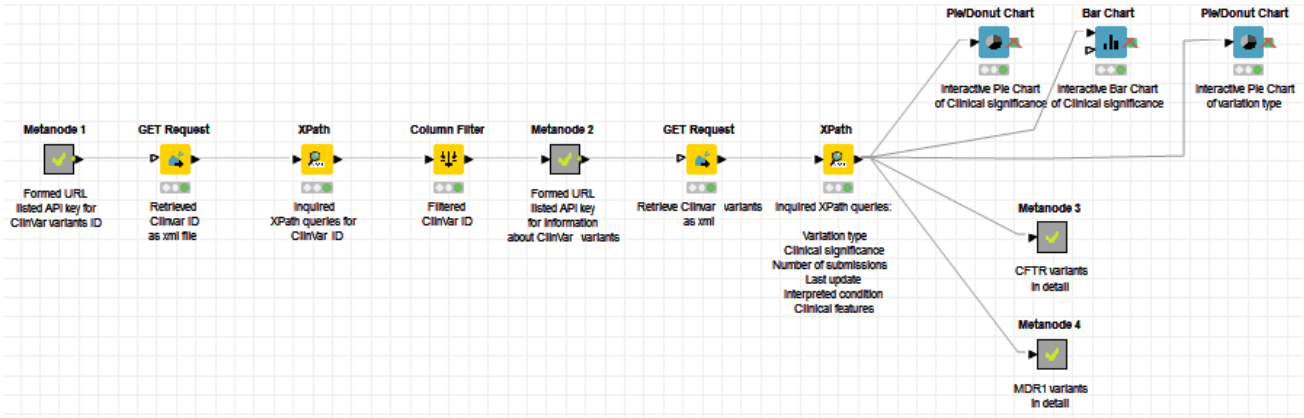

Fig. 1 shows the whole workflow.

Database	Amount of variants	
	Total count	Pathogenic
dbSNP	65.132	753
Ensembl	61.365	-
ClinVar	2.749	669
HuVarBase	764	-
Uniprot	210	210
OMIM	138	-
HGMD	-	-

Table 1 shows the located number of variants of CFTR through manually search.

B.2. Metanode 1: Forming API keys for searching ClinVar ID's

The workflow starts with a metanode named Metanode 1 to form the URL retrieving all ClinVar ID's of the 48 human ABC transporter genes listed in separate xml file formats (shown in Fig. 2). Xml is used due to the readability of both humans and machines programs such as used in KNIME. Therefore 3 terms were joined together to form the Uniform Resource Locators (URL). All of them were inserted in 3 *Table Creator* nodes. This node is used to create a data table manually.

The API keys for searching and downloading material of a definite database of NCBI are described in the Help guide "A General Introduction to the E-utilities" [16]. E-utilities (Entrez Programming Utilities) are public API providing the interface to the NCBI Entrez system and their databases [15],[17]. Basically first a set of unique identifiers of the requested data (e.g. ID of a gene or entry in PubMed) is listed and after that these are used to retrieve an overview of requested record entries (ESummary) or the complete records (EFetch).

The URL's which are used to search for and retrieve requested data are built the same way with defining parameters. These URL's starts with the base term *https://eutils.ncbi.nlm.nih.gov/entrez/eutils/* and further a command for searching or downloading data from a defined database. Other instructions can be added such as minimizing the number of requests. The output format can be specified as xml or json.

For NCBI databases such as ClinVar further guides are available [15]. In this guide the example for the requested URL to achieve ClinVar ID's of a single gene is listed and was changed to *https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearcch.fcgi?db=clinvar&term=XXXX[gene]&retmax=100000*. This URL was divided into three parts in three different *Table Creator* nodes. The XXXX-part represents the gene names.

The first node was the product of the base term and a term for text searching through the ClinVar database https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=clinvar&term=.

The second part consists of the different gene identifiers (gene ID) which are used in NCBI databases shown in table 2.

The last part was made of *[gene]&retmax=100000*. The term retmax describes the amount of variants which are shown in the xml file and can be set to any number. If it is not defined only the first 20 variant ID's are listed. Therefore the number was set to 100.000 as previous manual literature search about the CFTR gene showed fewer variants even in the dbSNP database.

The established columns of the *Table Creator* nodes were named “First Part URL”, “Gene ID” and “Third Part URL” and contained only one row each except the second *Table Creator* which hold 48 rows. No further parameters were set.

The first two *Table Creator´*s were joined by using the *Cross Joiner* node whereas each row of the top table is joined with each row of the bottom table. A second *Cross Joiner* node was used to connect the new generated table with the third part so that the established table consisted of the 3 columns “First Part URL”, “Gene ID” and “Third Part URL” in that sequence from left to right and altogether 48 rows.

In both *Cross Joiner’s* the parameters for chunk size were set to 1 which means only one row was read at once. The newly generated row identifiers (RowID) were a product of the anterior Row ID’s separated through a _. The bottom table’s column name suffix is not important and was not changed from the inherent (#1) as the bottom table wasn’t attached.

In the next step the *Column Aggregator* node merged the 3 columns “First Part URL”, “Gene ID” and “Third Part URL” together and formed the URL´s. This was adjusted as in the section “column” all three were selected to enforce inclusion through manual selection. Further in the section “option” the tack for removing aggregation columns was set, the generated column was named “Full combined URL 1” and aggregated by concatenation. The maximum unique values per row were not modified from the suggested value of 10000. The resulting table consisted of the column “Full combined URL 1” and 48 rows.

Because of the combinations through the *Cross Joiner’s* the Row ID’s expanded to three Row ID’s separated through a _, e.g. Row0_Row0_Row0. The rows were renamed via *RowID* node by setting a tack for replacing Row ID with selected column values or create a new one. Thereby each Row ID identifies one transporter gene whereas the first row starts with “Row0” and the other continued numerical till “Row47” as shown in table 2.

The output of Metanode 1 was the processed API keys of E-Utilities for searching up to 100.000 ClinVar ID’s for all 48 human ABC transporter genes.

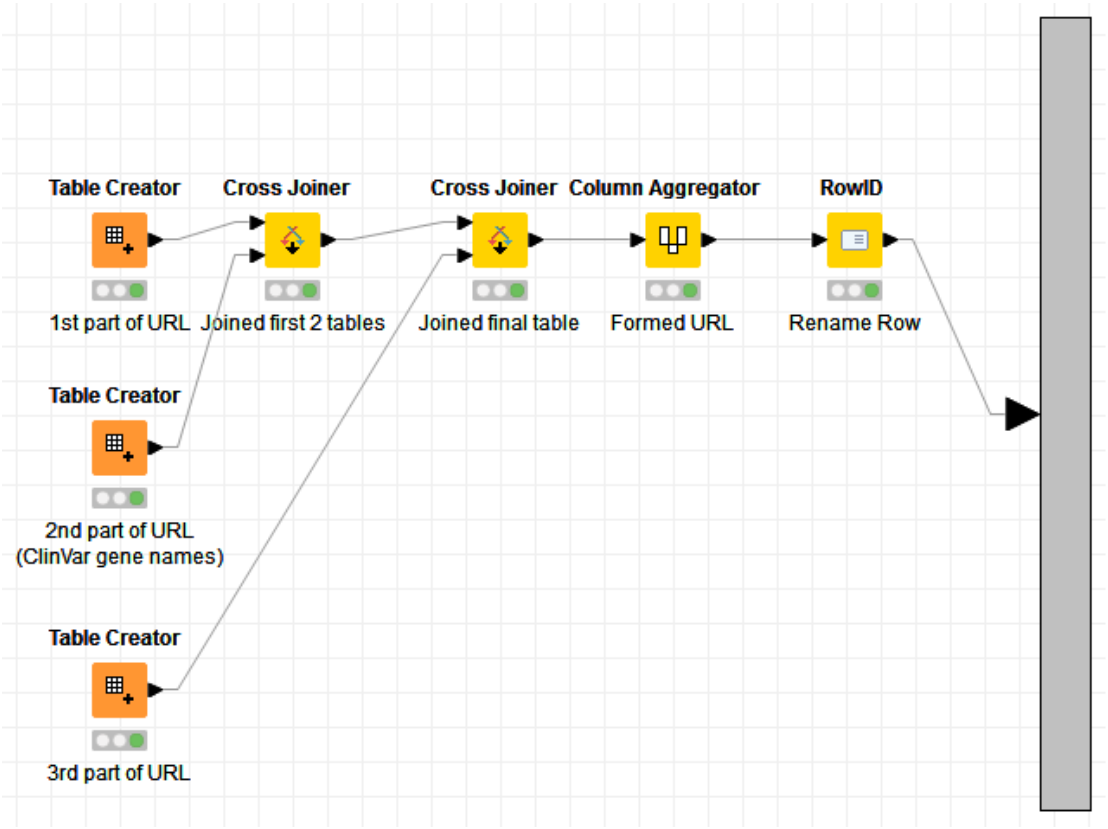

Fig. 2 shows Metanode 1 to form the URL retrieving all ClinVar ID's of the ABC transporter genes.

ABC Transporter	Gene ID used in workflow	Row ID for identification	ABC Transporter	Gene ID used in workflow	Row ID for identification
1	ABCA1	Row0	25	ABCC2	Row24
2	ABCA2	Row1	26	ABCC3	Row25
3	ABCA3	Row2	27	ABCC4	Row26
4	ABCA4	Row3	28	ABCC5	Row27
5	ABCA5	Row4	29	ABCC6	Row28
6	ABCA6	Row5	30	CFTR	Row29
7	ABCA7	Row6	31	ABCC8	Row30
8	ABCA8	Row7	32	ABCC9	Row31
9	ABCA9	Row8	33	ABCC10	Row32
10	ABCA10	Row9	34	ABCC11	Row33
11	ABCA12	Row10	35	ABCC12	Row34
12	ABCA13	Row11	36	ABCD1	Row35
13	ABCB1	Row12	37	ABCD2	Row36
14	TAP1	Row13	38	ABCD3	Row37
15	TAP2	Row14	39	ABCD4	Row38
16	ABCB4	Row15	40	ABCE1	Row39
17	ABCB5	Row16	41	ABCF1	Row40
18	ABCB6	Row17	42	ABCF2	Row41
19	ABCB7	Row18	43	ABCF3	Row42
20	ABCB8	Row19	44	ABCG1	Row43
21	ABCB9	Row20	45	ABCG2	Row44
22	ABCB10	Row21	46	ABCG4	Row45
23	ABCB11	Row22	47	ABCG5	Row46
24	ABCC1	Row23	48	ABCG8	Row47

Table 2 shows all 48 human ABC transporter with their gene name and resulting Row ID's.

CFTR, TAP1 and TAP2 do not follow the common nomenclature as the others because in case of CFTR this term is way more common and TAP1 and TAP2 are used in order to separate them from ABCB2 and ABCB3 transporters in other species.

B.3. Extracting ClinVar ID's

Afterwards a *GET Request* node was performed. This node is used to issue HTTP GET requests to retrieve data from a web service. The requests were sent to the 48 built "Full combined URL 1" from Metanode 1. Concurrency was set to 1, both "Follow redirects" and "Send large data to chunk" were checked. Timeout was leaved to recommended 4 seconds.

In the created table the columns "Status" and "Content type" were enclosed next to the existing "Full combined URL 1". The first column showed if GET requests were performed correctly and second one displayed if the text was xml format or not. The resulting 48 GET requests were received in single xml files in a new added column "body".

Further a *XPath* node was implemented. This node performed a XPath query to the previous processed 48 xml files to achieve the ClinVar ID's. Therefore the XPath query was named ClinVar ID, defined as */eSearchResult/IdList/Id* and the parameter "Multiple Tag options" was set to multiple rows.

The output table contained the three columns "Full combined URL 1", "Status" and "Content type" from the *GET Request* node and the resulting variant identifiers from XPath were listed in the new processed column "ClinVar ID's" with separated rows whereas the Row ID's (see table 2) where expanded by the separator _ and their number of variants, each Row ID starting with variant 1.

The outcome delivered a total number of 15.763 variants for all 48 ABC transporter genes as summarized in table 3. The most variants were found for CFTR and the numbers provided through the workflow corresponded to the previous ones found through manually search through the ClinVar database.

Due to the number of variants further outputs holding all variants are not presented in tables.

Afterwards a *Column Filter* node was inserted to exclude all columns by enforcing exclusion of all columns except the "ClinVar ID's" column which was used for further processing. The whole part is visualized in fig. 3.

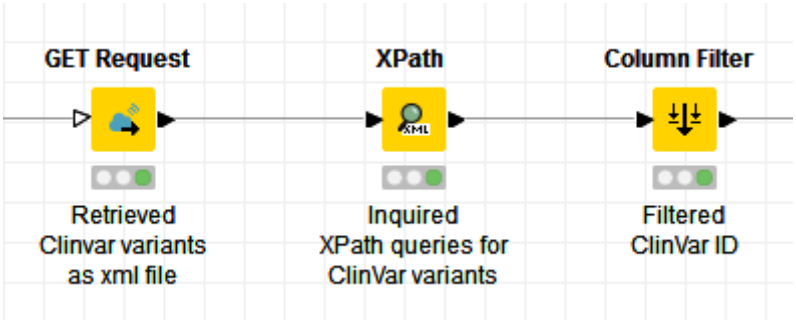

Fig. 3 shows the used nodes for retrieving ClinVar ID's and extracting them.

ABC Transporter	Total count of variants
ABCA1	578
ABCA2	135
ABCA3	421
ABCA4	2146
ABCA5	50
ABCA6	20
ABCA7	127
ABCA8	24
ABCA9	19
ABCA10	22
ABCA12	338
ABCA13	71
ABCB1	659
TAP1	178
TAP2	204
ABCB4	417

ABC Transporter	Total count of variants
ABCB5	42
ABCB6	81
ABCB7	248
ABCB8	77
ABCB9	20
ABCB10	36
ABCB11	765
ABCC1	382
ABCC2	422
ABCC3	23
ABCC4	95
ABCC5	41
ABCC6	916
CFTR	2749
ABCC8	1188
ABCC9	914

ABC Transporter	Total count of variants
ABCC10	12
ABCC11	41
ABCC12	30
ABCD1	973
ABCD2	21
ABCD3	28
ABCD4	262
ABCE1	38
ABCF1	13
ABCF2	69
ABCF3	38
ABCG1	87
ABCG2	37
ABCG4	36
ABCG5	280
ABCG8	390

Table 3 shows the total number of variants found in ClinVar for human ABC transporter.

B.4. Metanode 2: Forming URL to download whole ClinVar ID records

In the next step The ClinVar ID's were integrated in a second metanode, Metanode 2 (fig. 4). This was used to build API keys of E-Utilities for additional record downloads of each variant. Those were constructed similar to the first metanode.

Metanode 2 started with 2 *Table Creator* nodes and the extracted ClinVar ID's. The API key for performing the whole record download of the ClinVar variants was defined as *https://eutils.ncbi.nih.gov/entrez/eutils.efetch.fcgi?db=clinvar&rettype=vcv&is_variationid&id=XXXX&from_esearch=true*. The XXXX-part represents the ClinVar ID's.

The first *Table Creator* in Metanode 2 included the first part of the URL defined through the basic EUtils term, the EFetch term and the requested database [15]: *https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=clinvar&rettype=vcv&is_variationid&id=*.

The ClinVar ID's built the second part of the URL and last term was defined as *&from_esearch=true*.

According to the designations in Metanode 1 the established tables of the *Table Creator* nodes were named "First Part URL" and "Third Part URL" and each contained only one row. No further parameters were set.

They were joined by two *Cross Joiner* nodes and previous described parameters were set: Bottom table's column name suffix was not changed from (#1), Separator was defined as _ and Chunk size was 1.

The resulting table consisted of "First Part URL", "ClinVar ID's" and "Third Part URL" with 15.763 rows.

The Row ID's increased by attaching Row0_ as prefix and _Row0 as suffix (e.g. Row0_Row0_1_Row0 whereas the middle row term represented the first listed ClinVar variant of the ABCA1 gene, equal to Row 0 see table 2).

At the end the three columns "First Part URL", "ClinVar ID" and "Third Part URL" were merged together by a *Column Aggregator* node and formed the full combined URL´s. The settings were the same as in Metanode 1: in section "column" all three parts were enforced for inclusion through manual selection, and in section "option" the aggregated columns were removed by setting a tack. The built column was named "Full combined URL 2" and

aggregated also via concatenation. The maximum unique values per row were not modified from the suggested value of 10.000.

As the aggregated columns were excluded the output table consisted only of the column “Full combined URL 2”, holding the API keys for downloading the ClinVar ID records for all 15.763 variants.

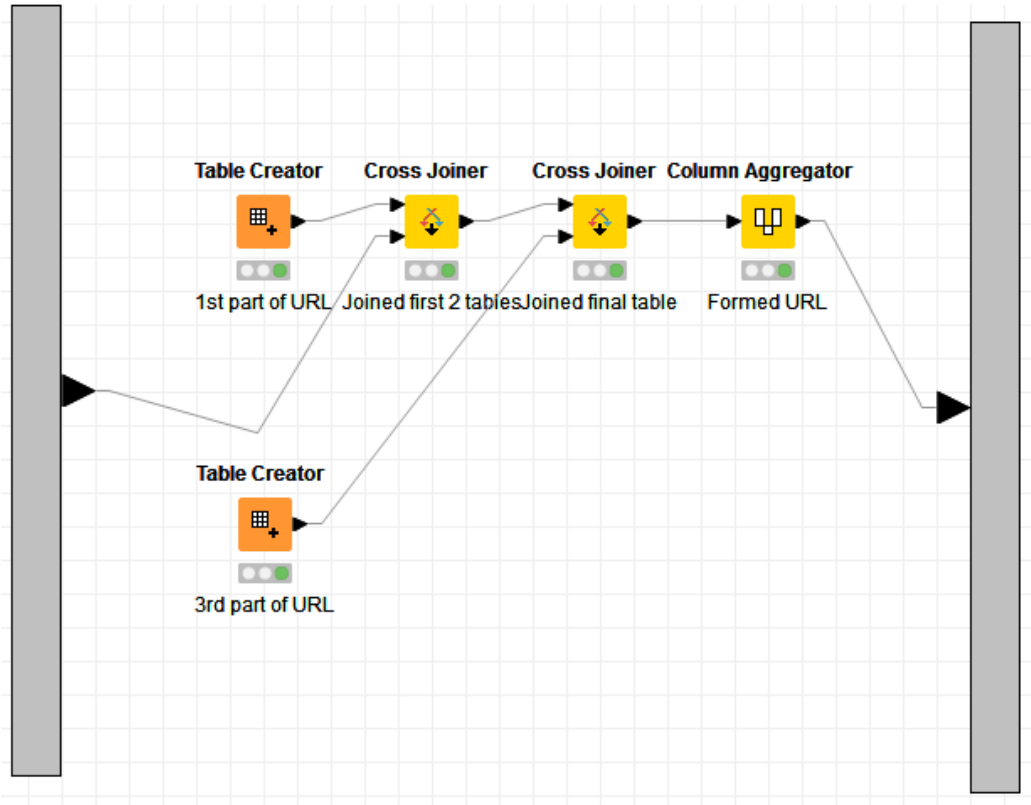

Fig. 4 shows Metanode 2 to generate URL retrieving whole records of all ClinVar ID’s of the ABC transporter genes.

B.5. Performing XPath queries retrieving information about variants and their visualization

Next to the built API keys a *GET Request* node was performed with similar settings as in the previous *GET Request* node (fig. 5).

In section “Connection settings” the source columns was “Full combined URL 2” with the proposed concurrency 1. Further the tacks for both “Follow redirects” and “Send large data to chunk” were checked. Timeout was increased up to 60 seconds as former trials resulted in single error requests from server side, eventually caused by limited processing power of the used equipment. Therefore in the section “Error handling” subitem “Retry on error” was tacked and number of retries was heightened to 5 times and retry delay was set to 3 seconds.

The output table contained the source column “Full combined URL 2”, “Status”, “Content type” and the 15.763 obtained GET requests as single xml files in the formed column “Body”, one per row.

Besides a *XPath* node was accomplished to acquire following information’s about the 15.763 variations with six XPath query parameters (displayed in table 4):

The requested terms are described in “The ClinVar variation report” [18]. “Variation type” describes the type of variation, whereas combinations of variations are described as Haplotype or Genotype. “Clinical significance” is adjusted through the sum of interpretation of all submitters. If no submission was performed, the cell was leaved empty apart from a question mark “?”.The term “Number of Submissions” represents the count of submissions for a variant, the expression “Last Update” shows the most recent update of the record. The “Interpreted Condition” describes the condition of the variant that was assumed from the submitters.

The output table contained the three columns “Full combined URL 2”, “Status” and “Content type” from the *GET Request* node and the resulting XPath queries about the variants in six new composed columns.

The outcome delivered 15.763 variations for all 48 human ABC transporter genes. 3.266 were cited as pathogenic and 985 as likely pathogenic. 1.800 were found to be benign and 2.480 likely benign. 4.720 variants have an uncertain significance. The remaining 2.512 mutations are divided into further categories like “not provided”, “pathogenic/benign and other”, or have “conflicts in their interpretations”. In this group the highest occurrence count is 920 with the term “conflicting interpretations of pathogenicity”.

For visualization a *Bar chart* node was implemented. Category column “Clinical significance” and aggregation method “Occurrence Count” were set. “Process table in memory” was tacked in order to fasten process speed.

Because this chart is an interactive chart from the workflow the total count of concurrences of the six biggest variation types were manually inserted in fig. 6.

The other received clinical significances were listed too but due to their low number their total counts were not inserted.

There are four entries, where no clinical significance was obtained as these variants were not submitted yet. These were listed with value 0 in the column “Number of Submissions” and thus no condition could be interpreted. The columns “Clinical Significance” and “Interpreted Condition” were left with question marks.

In all interactive charts from KNIME the visualization options can be adjusted via setting command which is placed in the upper right corner.

In addition, the variation type of the 15.763 processed variants was imagined in another *Pie/Donut Chart* node (fig. 7). Settings were the same except “category column” was defined as “Variation type” column. All 3 features of the chart were contemplated for the three biggest types. SNP’s were the largest representatives with 11.750 and made up 75% of all variant types. Besides 1.447 Copy number gains and 1.077 Deletions were registered.

The persisted 1.489 variation types were split into Copy number loss (793), Duplication (361), Microsatellite (179), Indel (81), Insertion (68), Haplotype (31), Inversion (17), Variation (7), Complex (3) and Protein only (2).

For verifying the workflow two genes were examined in more detail, the CFTR and MDR1 (ABCB1, P-gp).

Column name	Xpath query	Type
Variation typ	/ClinVarResult-Set/VariationArchive/@VariationType	String (SingleCell)
Clinical significance	/ClinVarResult-Set/VariationArchive/InterpretedRecord/Interpretations/Interpretation/Description	String (SingleCell)
DateCreated	/ClinVarResult-Set/VariationArchive/@DateCreated	String (SingleCell)
Number of submissions	/ClinVarResult-Set/VariationArchive/InterpretedRecord/Interpretations/Interpretation/@NumberOfSubmissions	String (SingleCell)
Last Update	/ClinVarResult-Set/VariationArchive/@DateLastUpdated	String (SingleCell)
Interpreted condition	/ClinVarResult-Set/VariationArchive/InterpretedRecord/RCVList/RCVAccession/InterpretedConditionList/InterpretedCondition	String (CollectionCell)

Table 4 shows the Xpath queries and adjusted parameters.

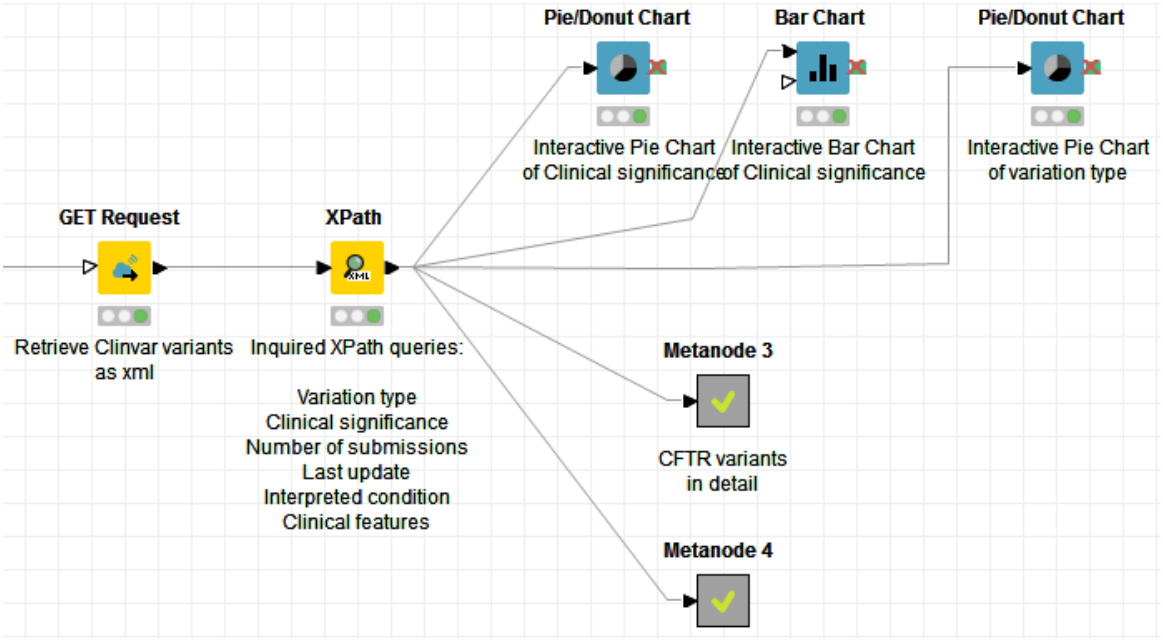

Fig. 5 shows the *used nodes for retrieving information about the variant entries, extracting them and their visualization.*

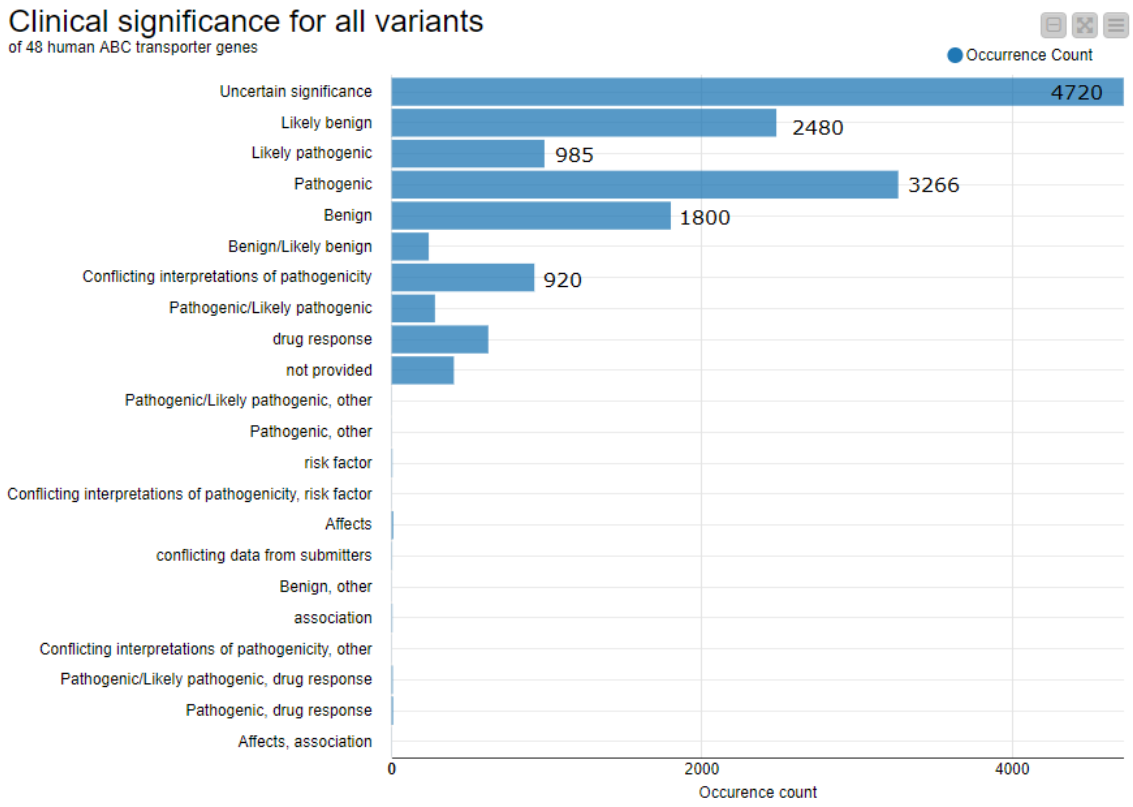

Fig. 6 shows the clinical significance of the variants; the total numbers of the 6 biggest groups are shown.

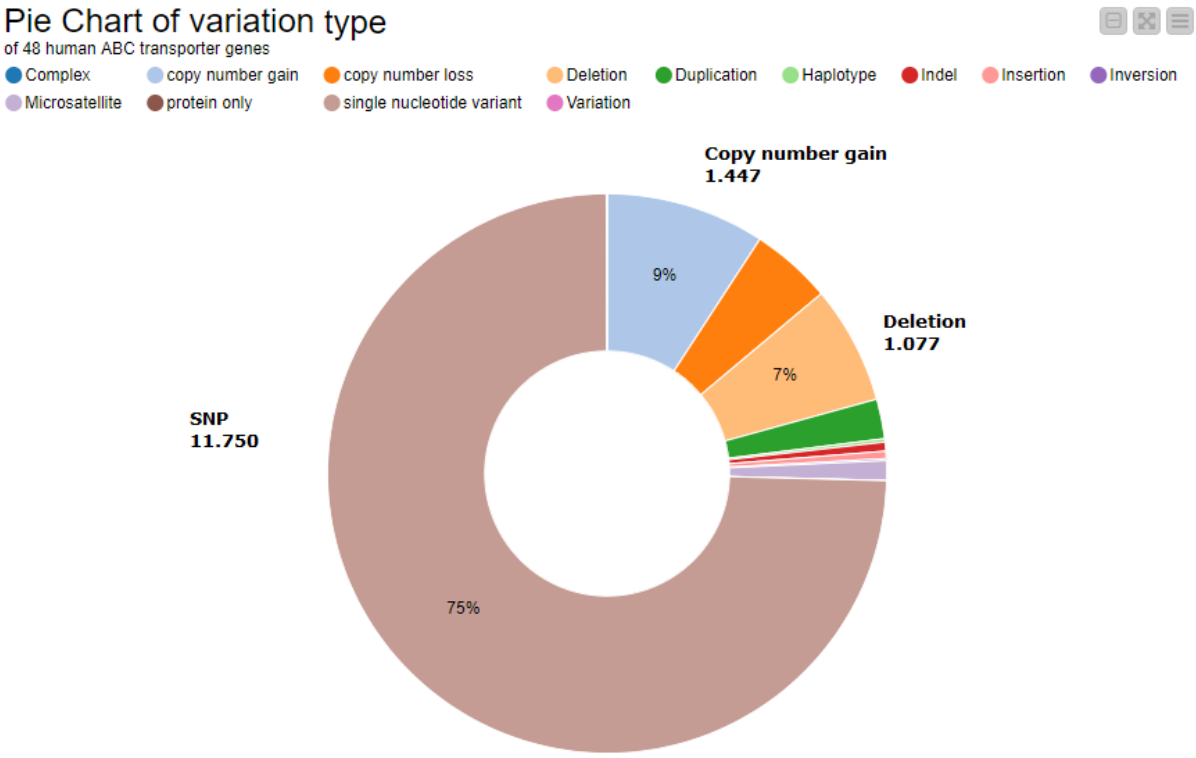

Fig. 7 shows the pie chart of variation types for the three largest groups retrieved from all 48 human ABC transporters.

B.6. Metanode 3: Cystic fibrosis

Cystic fibrosis (CF) is a complex autosomal recessive disease with a large amount of different phenotypes. The most common are pancreatic insufficiency, chronic pulmonary disease and salt loss in sweat [19]. CF is derived through variants affecting the cystic fibrosis transmembrane conductance regulator (CFTR) gene. It encodes a chloride ion channel, the only known of the human ABC transporters. The protein controls ion and water secretion and absorption in respiratory, digestive, reproductive and sweat glands epithelium [20]. Variations cause thicker mucus in respiratory and digestive systems and also malabsorption of chloride and sodium [21]. Since cystic fibrosis was first described in 1938, over 2000 mutations are known from the CFTR gene, about 300 of them are pathogenic [9], [21].

For further data editing the previous resulted XPath queries were inserted in a Metanode 3 (fig. 8). They were divided through a *Row Splitter* node as the filter criteria was set to “Include rows by row ID” and the regular expression was defined as Row0_Row29 which was the composed Row ID for CFTR from Table 1. The tack for “Row ID must start with this expression” was set. Thus, 2749 variants remained. Their variation type was pictured by using another *Pie/Donut Chart* node, the source column was expounded as “Variation type” and aggregated by its occurrence count (imagined in fig. 9). The outcome showed that also SNP built the largest group of variation types with a total count of 2.117 (77%) and far afield Deletion with just 370 (13%). The remaining 262 were Duplication (112), Microsatellite (51), Indel (28), Insertion (24), Haplotype and Copy number loss (each 15), Copy number gain (11), Variation (3) and Protein only (2).

In the next step the interpreted conditions were visualized in a *Bar Chart* node. Adjusted parameters were “Interpreted column” as Source column, Aggregation via Occurrence count was taken. The resulting chart showed that cystic fibrosis made 1.947 of the variants, 402 were not provided and 248 not specified (fig. 10). The term “not provided” indicates that no assumption for the condition was made through submitters. Whereas “not specified” is used when a variant is asserted to be benign, likely benign, or of uncertain significance for conditions that have not been specified. The remaining cases were not further specified as they vary widely due to different publications.

For proving the accuracy of the workflow two *Row Filter* nodes were inserted with the following parameters: Column to test was in both nodes “Clinical significance” and by tagging “Include rows by attribute value” and as matching criteria the pattern match terms were defined as first “Pathogenic” and in the second node “Cystic fibrosis”.

669 variants were cited as pathogenic with the interpreted condition cystic fibrosis. The fact that both the total count of found mutations and pathogenic variants were higher than these

accessed through manual search could possibly be explained as previous papers often recommend to the same sources, the cystic fibrosis mutation database (CFTR1 database) or the CFTR2 database. Both provide a collection of expert-reviewed functional and clinical information on CFTR mutations. But they possibly lack information due to limited updates, CFTR2 was last updated September 24, 2021. And 667 of the 669 gained variants from the ClinVar database were last updated through a submitter after that date.

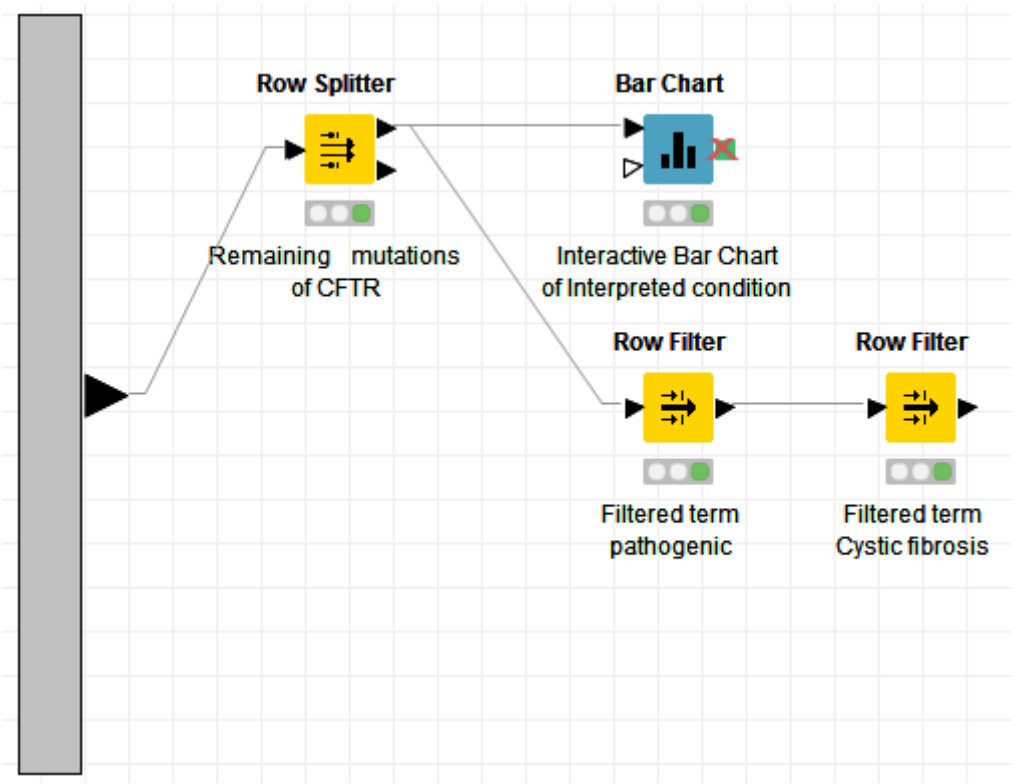

Fig. 8 shows the used nodes of Metanode 3

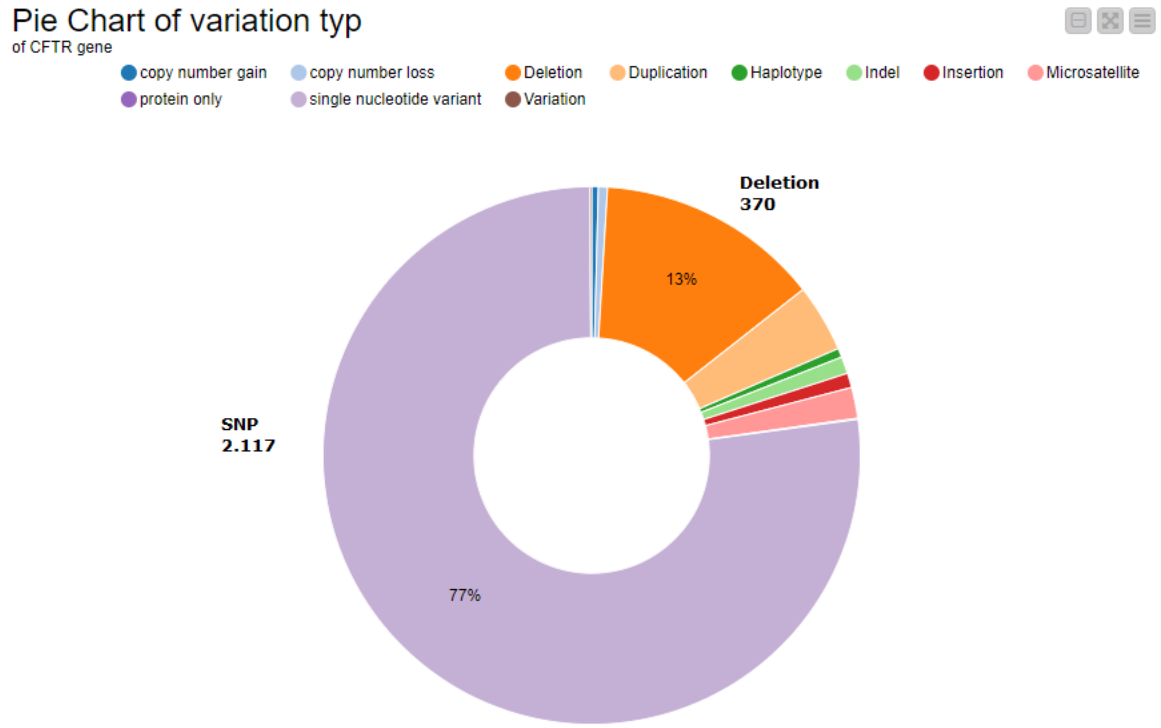

Fig. 9 shows the pie chart of variation types for the two largest groups retrieved from CFTR

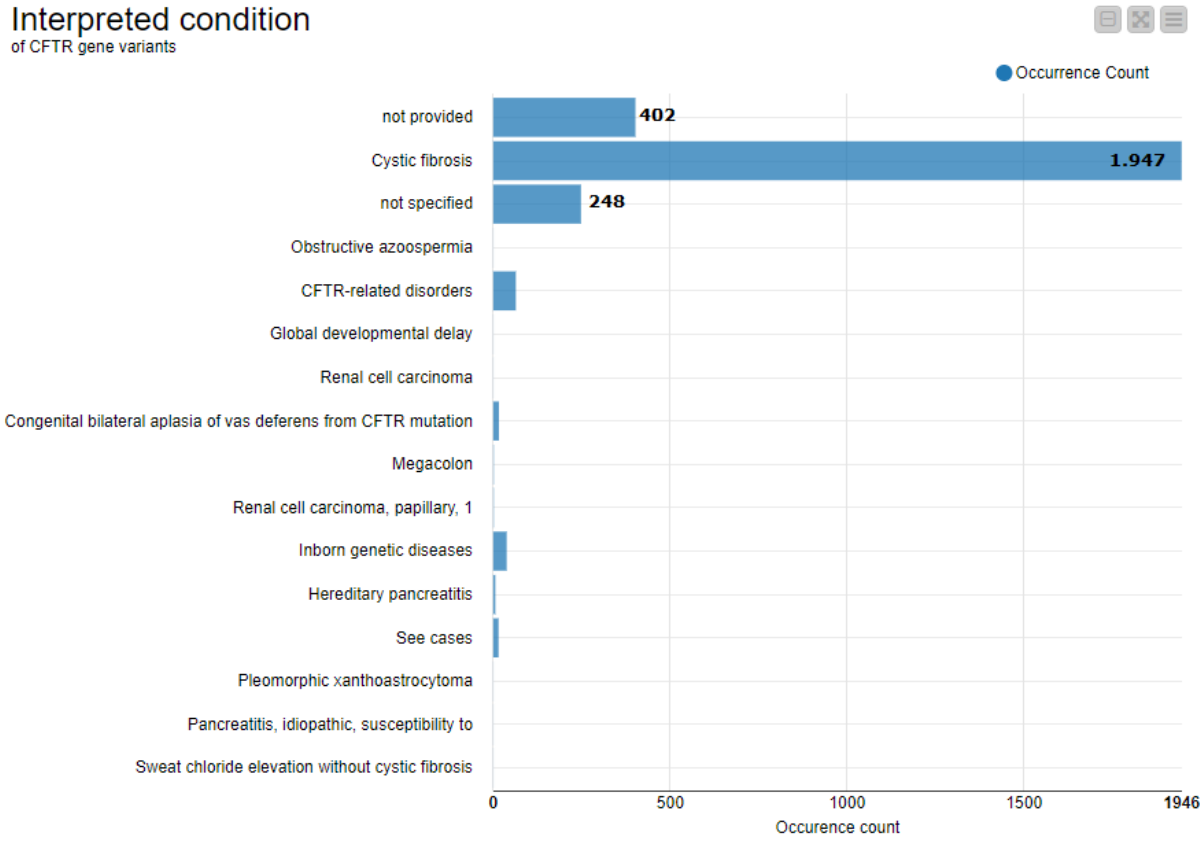

Fig. 10 shows the interpreted condition of the CFTR variants, for the three first the values were inserted.

B.7. Metanode 4: Multidrug resistance protein 1

MDR 1 also known as p-glycoprotein 1, is an ABC transporter encoded from the ABCB1 gene. It was the first discovered in the row of ABC transporters which cause multidrug resistance (MDR) [11]. Other important members of MDR are MRP (multidrug resistance associated protein, encoded from the C family of ABC transporters) and BCRP (breast cancer resistance protein, encoded from ABCG2).

MDR 1 is widely distributed and expressed in many organs, especially the apical membrane of intestinal epithelium or enterocytes [22], proximal tubule of the kidneys [23] and capillary endothelium cells in microvessels in the blood brain barrier [24]. Normal excretion of p-glycoprotein preserves these organs from damage as it exports a large variety of substrates, such as toxins or drugs, outside of the cells [23]. An example would be the export of analgesics such as opioids (e.g. tramadol) and mutations of the gene may be affecting opioid therapy [25]. Due to its overexpression in the membranes of cancer cells it renders these cells multidrug resistant and therefore it is the mainly cause for treatment failure in cancer therapy [26].

In literature about 50 SNP's [29], [28] were found for the ABCB1 gene. In the workflow the resulting XPath queries were used to form Metanode 4 (fig. 11). As in Metanode 3 a *Row Splitter* node was adjusted with similar filter criteria: "Include rows by row ID", the regular expression was defined as Row0_Row12 (see Table 2) and the tack for "Row ID must start with this expression" was set.

659 variants of MDR1 were delivered and their variation type was visualized through a *Pie/Donut Chart* node, the source column was expounded as "Variation type" and aggregated by its occurrence count (fig. 12). The outcome showed that once more SNP was the biggest group of variation types with a total count of 602 (91%) and Deletion with just 35 (5%). The remaining 22 were Copy number loss (10), Copy number gain (8), Microsatellite (3) and Inversion (1).

Next the interpreted conditions were visualized in a *Bar Chart* node. Adjusted parameters were "Interpreted column" as Source column, Aggregation via Occurrence count was taken. The resulting chart showed that with a count of 605 "tramadol response" was the main condition, 34 were "not provided" (fig. 13). The other 20 conditions were not further determined due to their low number of interpretations through submitters on ClinVar.

The term tramadol response is defined through ClinVar as a changed enzyme activity of CYP2D6. It can be increased, decreased or absent [29].

This enzyme converts tramadol into its six time more potent metabolite O-desmethyltramadol (M1). Therefore patients with lower CYP2D6 activity or even absent activity are named “intermediate metabolizer” and “poor metabolizer” and their pain relief would not be accomplished because of lower M1 serum levels. These “poor metabolizers” cells are therefore resistant to tramadol. Patients with increased enzyme activity are named “ultrarapid metabolizer” and as they are exposured by higher M1 serum levels they may achieve a higher risk of side effects. Unfortunately it is not detailed defined in which way the enzyme activity is changed.

For accuracy proving of the workflow a *Row Filter* node was inserted to filter the number of SNP’s from MDR1 and to comprise the results with those found in literature. The parameters in the *Row Filter* were defined as Column to test was “Variation type”, by tagging “Include rows by attribute value” and for matching criteria the pattern match term was defined as “single nucleotide*?” and further “wild card expression” must be tagged. Both the term “*?” and wild card expression include all variants that contain “single nucleotide” as SNP’s are often named single nucleotide polymorphism but in ClinVar they are named single nucleotide variants.

602 variants were listed as SNP’s. The found number of SNP’s for MDR1 gene is twelve times higher than the ones named in literature, which could be explained by the fact, that these papers were published in 2013 and 2020 but they refer to sources from 2001, 2003 and 2010. Further, studies often include just a distinct population group such as Japanese or Caucasian and would not include all SNP’s since these vary highly within an ethnic group.

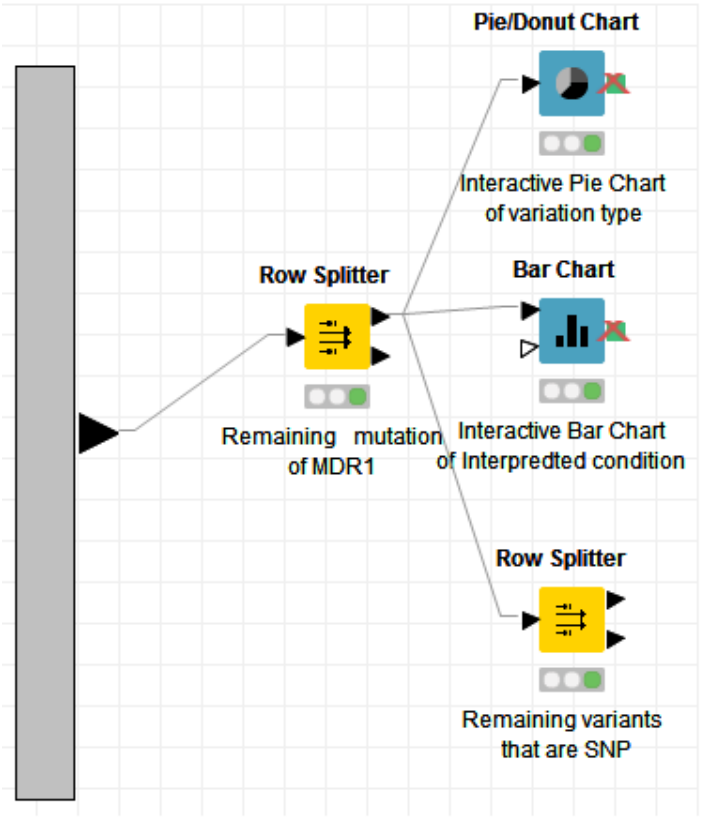

Fig. 11 shows Metanode 4.

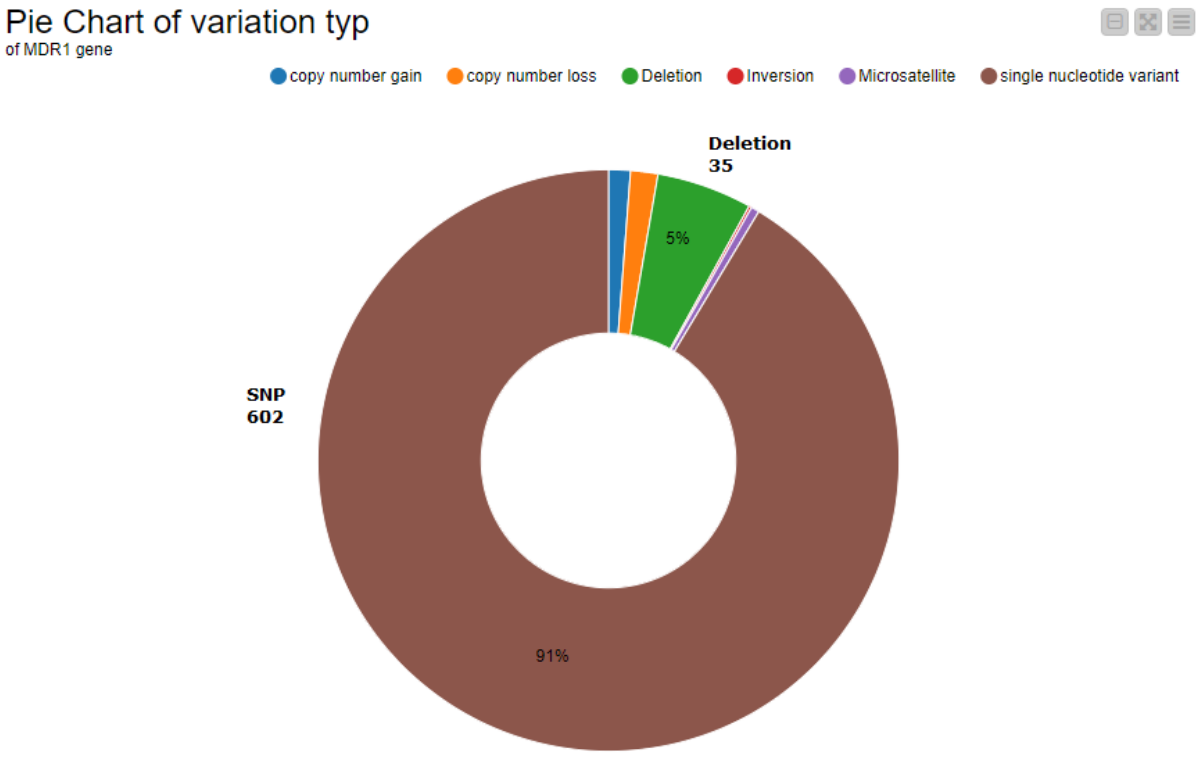

Fig. 12 shows the interactive Pie/Donut chart of the variation types from MDR.1

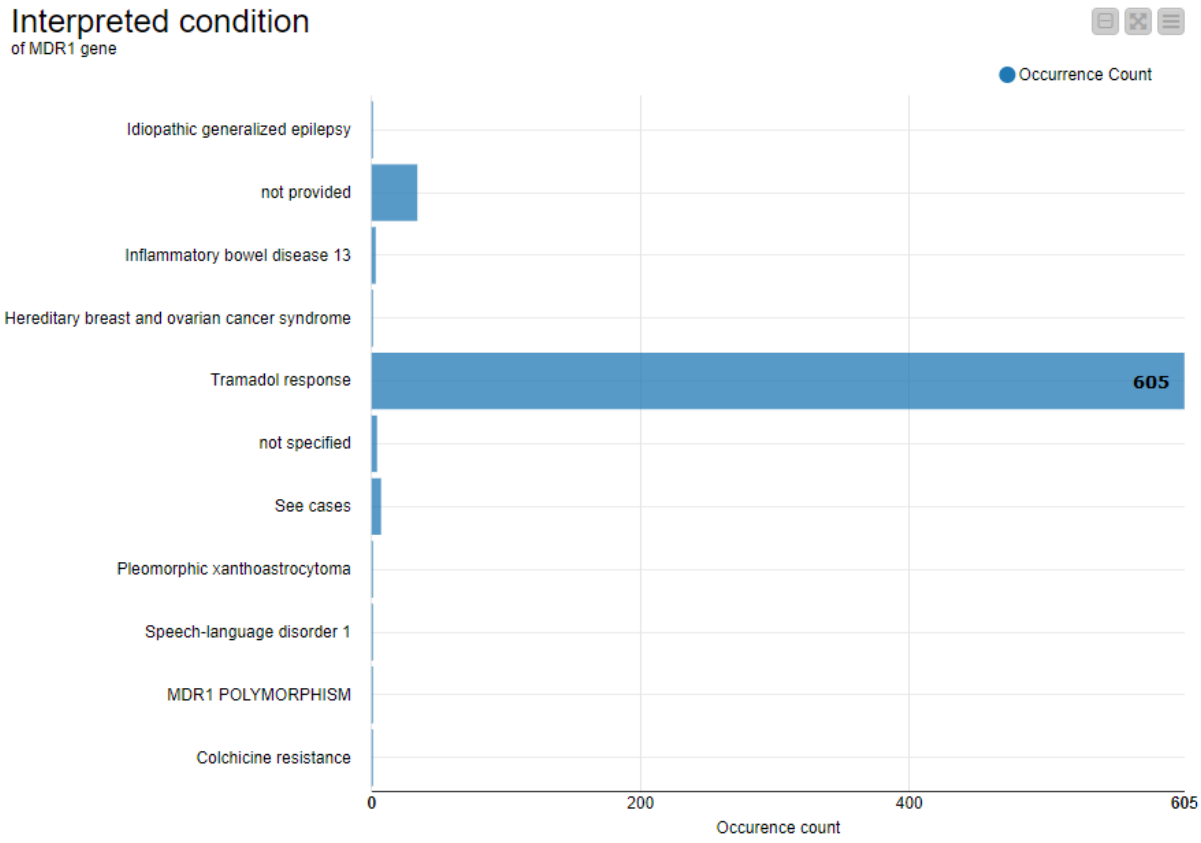

Fig. 13 shows the interpreted condition of the MDR1 variants.

B.8. Conclusion

ABC transporter are common in all living organism and are important membrane bound transport proteins. Knowledge about ABC transporter and their involvement in the development of diseases and MDR effects has increased significantly in last two decades.

Numerous mutations have already been identified and were reviewed in various papers since the expansion of genome wide association studies. Especially SNP play an important key role as they can be used as biological markers for defining the individuals risk for pathogenesis and response to drugs.

The workflow developed in this thesis, listed many of these known disease causing SNP's. But also a lot which were not further classified by submitters and needs further analyzes.

Even if their influence regarding to their clinical significances is well established and pointed out in the databases, the in-depth knowledge about phenotypes still needs updates and further reviews.

Since all data is distributed in different databases and reviewed in many papers, future workflows could use pathogenic SNP's as inputs in disease computational studies as pattern searching in classification model to combine known effects of these variants from different databases to expand knowledge of diseases and phenotypes.

Therefore personalized medicine needs not just further research but also open access of data.

Moreover, variants which are linked to MDR effects can be used as inputs to perform reevaluation of present available data. This may be important as classic ABC transporter inhibitors of first generations were inefficient or to less efficient in patients and thereby research was intensified. Especially in the last years new strategies were performed and reviewed to overcome MDR and deal with it to accomplish a successful treatment for cancer patients.

C. Literature

C.1. List of references

[1]: Vasiliou V., Vasiliou K., Nebert D.W. Human ATP-binding cassette (ABC) transporter family. Hum. Genom. 2009; 3:281–290. doi: 10.1186/1479-7364-3-3-281. [PMC: 2752038]

[2]: Tarling E. J., de Aguir Vallim T. Q., Edwards P. A. Role of ABC transporters in lipid transport and human disease. Trends Endocrinal Metab. 2013; 24:342-350. Doi:10.1016/j.tem.2013.01.006. [PubMed: 23415156]

[3]: Piehler A. P., Hellum M., Wenzel J. J., Kaminski E., Haug K., Kierulf P., Kaminski W. E. 2008. The human ABC transporter pseudogene family: evidence for transcription and gene-pseudogene interference. BMC Genomics 9:165. 10.1186/1471-2164-9-165. [PMC: 2329642]

[4]: Lopez-Fernandez L.A. ATP-binding cassette transporters in the clinical implementation of pharmacogenetics. J. Pers. Med. 2018; 8, 40 10.3390/jpm8040040. [PMC: 6313650]

[5]: Robey R. W., Pluchino K. M., Hall M. D., Fojo A. T., Bates S. E., Gottesman M. M. Revisiting the role of ABC transporters in multidrug-resistant cancer. Nature Reviews. Cancer. 2018; 18(7):452–464. doi: 10.1038/s41568-018-0005-8. [PMC 6622180]

[6]: Wang J.-Q., Yang Y., Cai C.-Y., Teng Q.-X., Cui Q., Lin J., Assaraf Y. G., Chen Z.-S. Multidrug resistance proteins (MRPs): Structure, function and the overcoming of cancer multidrug resistance. Drug Resist. Updat. 2021; 54:100743. doi: 10.1016/j.drup.2021.100743 [PubMed 33513557]

[7]: Fitzgerald DM, Rosenberg SM. What is mutation? A chapter in the series: How microbes “jeopardize” the modern synthesis. PLoS Genet. 2019;15(4):e1007995. doi: 10.1371/journal.pgen.1007995. [PMC 6443146]

[8]: Siegel GJ, Agranoff BW, Albers RW, et al. Basic Neurochemistry: Molecular, Cellular and Medical Aspects. 6th edition. Lippincott-Raven. 1999

[9]: Shastry B. S. SNPs: impact on gene function and phenotype. Methods Mol Biol. 2009. 578:3-22. doi: 10.1007/978-1-60327-411-1_1. [PubMed 19768584]

[10]: Online in the Internet: “https://www.knime.com/knime-analytics-platform”. February 09 2022

[11]: Online in the Internet: “https://www.knime.com/knime-open-source-story“. February 09 2022

[12]: Online in the Internet: “https://www.knime.com/getting-started-guide?”. February 09 2022

[13]: Stenson P. D., Mort M., Ball E. V., Evans K., Hayden M., Heywood S., Hussain M. Philips A. D. Cooper D. N. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet. 2017; 136:665–77. doi: 10.1007/s00439-017-1779-6. [PMC 5429360]

[14]: Online in the Internet: “https://www.ncbi.nlm.nih.gov/clinvar/intro”. February 08 2022

[15]: Online in the Internet: “https://www.ncbi.nlm.nih.gov/clinvar/docs/maintenance_use/“. February 10 2022

[16]: Sayers E. Bethesda (MD): NCBI: A General Introduction to the E-utilities. 2010. Online in the Internet: https://www.ncbi.nlm.nih.gov/books/NBK25497/. February 10 2022

[17]: Online in the Internet: “https://www.ncbi.nlm.nih.gov/tools/”. February 10 2022

[18]: Online in the Internet: “https://www.ncbi.nlm.nih.gov/clinvar/docs/variation_report/”. February 10 2022

[19]: Wiencek J. R., Lo S. F. Advances in the Diagnosis and Management of Cystic Fibrosis in the Genomic Era. Clin Chem. 2018. 64:898-908. doi: 10.1373/clinchem.2017.274670. [PubMed 29436379]

[20]: Radlovic N. Cystic fibrosis. Srp Arh Celok Lek. 2012. 140:244-9. [PubMed 22650116]

[21]: Pereira S. V.-N., Ribeiro J. D., Ribeiro A. F., Bertuzzo C. S., Lima Marson F. A. Novel, rare and common pathogenic variants in the CFTR gene screened by high-throughput sequencing technology and predicted by in silico tools. Sci Rep 9, 6234. 2019. doi: 10.1038/s41598-019-42404-6

[22]: Atkinson S. Huang S.-M., Lertora J. J. L. Markey S. P. Principles of Clinical Pharmacology (Third edition). Academic Press 2012. doi: 10.1016/B978-0-12-385471-1.01002-3.

[23]: Sosnik A., Bendayan R. Drug Efflux Pumps in Cancer Resistance Pathways: From Molecular Recognition and Characterization to Possible Inhibition Strategies in Chemotherapy (vol. 7). Academic Press. 2019. doi: 10.1016/C2017-0-00891-1.

[24]: Kokki H., Kokki M. Neuropathology of Drug Addictions and Substance Misuse. Chp Central Nervous System Penetration of the Opiod Oxycodone. Sci Rep. 2016. doi:10.1016/B978-0-12-800634-4.00045-7.

[25]: Lötsch J., Skarke C., Liefhold J., Geisslinger G. Genetic predictors of the clinical response to opioid analgesics: clinical utility and future perspectives. Clin Pharmacokinet. 2004;43:983-1013. doi: 10.2165/00003088-200443140-00003. [PubMed 15530129]

[26]: Dong. J. et al. Medicinal chemistry strategies to discover P-glycoprotein inhibitors: An update. Drug Resist. Updat. 2020. 49:100681. doi: 10.1016/j.drup.2020.100681. [PubMed 32014648]

[27]: Zawadzka I. et.al. The impact of ABCB1 gene polymorphis and its expression on non-small-cell lung cancer development, progressuin and therapy – preliminary report. Sci Rep 10 (6188). 2020. doi: 10.1038/s41598-020-63265-4.

[28]: Wang L-H, Song Y-B, Zheng W-L, Jiang L, Ma W-L. The association between polymorphisms in the *MDR1* gene and risk of cancer: a systematic review and pooled analysis of 52 casecontrol studies. Cancer Cell Int. 2013;13:46. doi: 10.1186/1475-2867-13-46. [PMC 3669001]

[29]: Online in the Internet: “https://www.ncbi.nlm.nih.gov/medgen/CN078023/”. February 09 2022