Paper DH02

Biomarkers Unleashed: A Programming Pipeline from Chaos to Clarity in Genomic Data

Lea Zittlau, Chrestos Concept GmbH & Co. KG, Essen, Germany

Dr. Saskia Ständler, Chrestos Concept GmbH & Co. KG, Essen, Germany

ABSTRACT

Biomarker analyses are becoming increasingly important in diagnostics and clinical trials, not only in oncology but also in other fields, as the gained insights pave the way to personalized therapies. The generic term “biomarker” includes a great variety of biological molecules, and this diversity also applies to the type of their analysis methods. One promising case is the use of genetic data which can be obtained via next generation sequencing (NGS). Working in the genomics field, statistical programmers face the challenge of handling a huge amount of primarily unstructured raw data. Here, we show an exemplary pipeline which details the processing of NGS data to create a final analysis set. Thereby, we address questions of data standardization, harmonization, and annotation of genomic data by using public databases. Sharing insights about such a workflow hopefully increases the awareness of the challenges in biomarker analyses and their solutions.

INTRODUCTION

In recent years, patient care has been revolutionized by the introduction of big data applications within clinical trials. Genomic sequencing, for example, has provided insights into high inter-individual variations between individuals, leading to the conclusion that it may be possible to tailor a disease to individual characteristics.[1] As a result, personalized medicine - also known as precision medicine - has found its way into the field of clinical trials. While both terms are generally considered to be analogous, the term “personalized” has raised concerns as it could be misinterpreted to being uniquely developed for each individual.

WHAT’S THE GOAL OF PRECISION MEDICINE?

Precision medicine aims to use biomarkers - individual characteristics - to classify individuals into different subpopulations in order to tailor their medical treatment. Classifications may be based on the susceptibility to a particular disease, the underlying biology of a disease, or the response to a specific treatment. Consequently, personalized medicine differs from conventional therapies in that it focuses on groups of individuals who are most likely to benefit from this kind of treatment, sparing expenses and side effects for these people which might arise through conventional therapies.[2] Based on these classifications, it is possible to develop targeted approaches that are predominantly highly selective in combination with low grade toxicities.[3] A prominent example is the small-molecule inhibitor crizotinib, which is used in a targeted therapy approach for NSCLC patients with ALK or ROS1 gene fusions.[4–6] In this case, both fusions cause tumor development and thus can function as genomic biomarkers. By using them in diagnostics, it is possible to narrow down small subpopulations of up to 7%[4–6] of NSCLC patients with these types of fusions who are sensitive to crizotinib. However, such targeted approaches can also have some pitfalls, as there may be a certain number of individuals from the subpopulation who do not respond to the therapy at all or who stop responding after some time, which either induces further progression or relapse of the disease. This is based on the common phenomenon of drug resistance, another hot topic in precision medicine. While pre-existing resistance is referred to as "primary resistance"[7], mechanisms that only develop during, or as a result of, therapy are referred to as "acquired resistance".[8] Reasons for drug resistance can be manifold, yet the predominant mechanisms are mutations in the target genes[3], also known as on-target mutations[9], in which cancer cells try to evade the given therapy. In addition to other mechanisms, such as epigenetic changes, another common cause of drug resistance is mutations in downstream signaling genes that the tumor cell uses to provide sustained signaling. These kinds of genetic alterations are often referred to as off-target mutations[9] and can occur in different gene subtypes, which are generally divided into three categories: Oncogenes, tumor suppressor genes (TSGs), and genes with ambiguous effects. Oncogenes generally mediate signaling for cell growth and cell division, whereas TSGs slow down or even inhibit this signaling and mediate apoptosis. Mutational changes in these subtypes can lead to an increased activation of oncogenes or an inhibition of TSGs, which can lead to massive and uncontrolled cell growth, for example of tumor cells. Ambiguous genes can function as either oncogenes or TSGs, depending on the respective cell signaling. All kinds of genetic abnormalities mentioned can be analyzed using genetic data where such resistance mutations, either on- or off-target, are considered biomarkers.

WHAT ARE BIOMARKERS IN GENERAL AND HOW CAN WE USE GENOMIC DATA AS BIOMARKERS?

Biomarker is a generic term for “a defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or biological responses to an exposure or intervention, including therapeutic interventions”.[10] This term refers not only to a great variety of biological molecules, such as proteins, DNA, RNA, or cellular characteristics, which are measured by different technologies, but also to physiological-, behavioral-, imaging-, electrophysiological- and digital characteristics. Given this high level of complexity, there are also many rationales for the use of biomarkers in clinical trials based on the type of biomarker, the time of measurement and the study design. In general, biomarkers can be used in diagnostics, as mentioned in the example of crizotinib, but they are also useful in predictive, prognostic, and pharmacodynamic evaluations. Within the scope of this paper, we will focus on genomic data as biomarkers that can be obtained using next-generation sequencing (NGS) and how this data can be used in clinical trials. Referring to the example of crizotinib, we will present an exemplary pipeline of processing NGS data into a final ADaM dataset for TLF creation.

SIMULATED CASE STUDY FOR PROGRAMMING PIPELINE

NGS sequencing provides a large number of so-called short reads, which are mapped to a reference genome to obtain a final aligned sequence. The final aligned sequence is then compared to the reference genome to reveal any kind of genetic changes in the sequenced sample, such as single nucleotide variants (SNVs). Because of its massively parallel sequencing approach, NGS produces a large amount of data, which can be overwhelming when viewed for the first time. One of the challenges of working with genomic data is identifying those SNVs or other kinds of genetic changes that are relevant to the respective research question. In our simulated case study, we are targeting the research question of whether there are individuals with resistance mutations that ultimately lead to crizotinib resistance.
In the following sections, we will present a programming pipeline example which, in addition to answering our research question, will illustrate the issues faced by a statistical programmer in data standardization, harmonization, and annotation of genomic data.

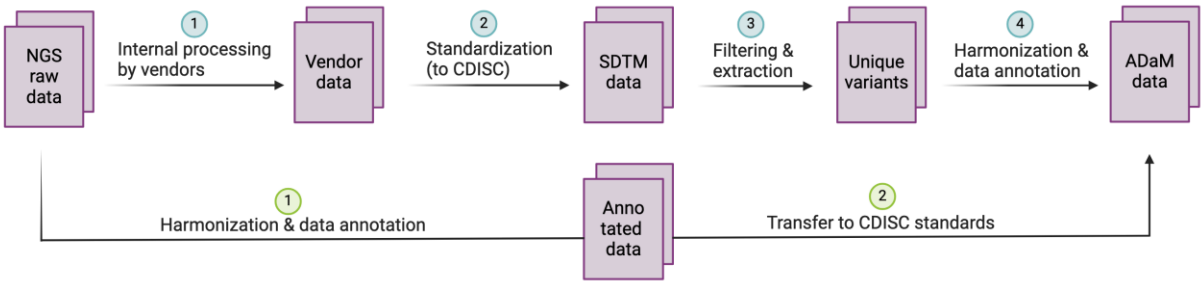

Figure 1: Schematic presentation of the programming pipeline

The programming pipeline (**Figure 1**) consists of four steps that must be completed to obtain a final ADaM dataset. First, raw NGS data is generated by sequencing in the vendor's lab, which is internally translated into a readable format before the sequencing results are delivered to the sponsor of a study. This readable format shows, for example, the name of the gene in which an alteration is observed and the associated genetic position. Given the fact that a sponsor may receive data from multiple vendors, it is essential to integrate all the data received into one combined, standardized data set. This is the starting point for the second step. Here, the statistical programmers' task is to map all the given information into one data standard, since for example, variable names may be different for the same type of information, or even the way in which a particular information is given may vary. Guidance on how to map such non-standardized vendor data into proper SDTM data (GF domain) is given by the SDTM Implementation Guide for Human Clinical Trials (Version 3.4) published by CDISC in 2021. Before carrying out the final step of data annotation (step 4), which is most likely done by a bioinformatician, it is advisable to extract the most important parts of the data which are so-called unique variants (step 3). As the parallel sequencing approach in NGS generates a large amount of data, and depending on the type of assay used, the quality of the samples and the number of individuals enrolled, this amount of data can quickly reach hundreds of thousands of data points. Annotating all these data points would potentially take a lot of time and computing power, but this can be circumvented by using only unique variants. Such unique variants may be present multiple times in one subject or even occur in more than one subject. Identification of unique variants can be done through the extraction of all emerging combinations of the genomic position, nucleotide change and amino acid change. Once extracted, the process of data annotation aids to harmonize the data and even more important, to give meaning to the respective sequences. Typical information most likely to be obtained from database queries is the classification of gene types, the degree of pathogenicity or the type of disease with which the respective alteration is associated. Based on the information gained, it is possible to properly interpret sequencing data and to narrow down the records to those entries showing relevance in the respective study. Overall, this pipeline could also work the other way around, if a sponsor were to receive NGS raw data

directly. If so, this raw data could immediately be annotated (blue; step 1) and afterwards transferred into CDISC standards (blue; step 2). However, in the simulated case study we will not focus on this alternative way.
To demonstrate the programming pipeline, we have generated small sample vendor datasets (see **Tables 1&2**) including four non-existent individuals treated with crizotinib, based on ALK or ROS1 gene fusions, but showing limited responses due to potential resistance mechanisms. To make the pipeline easy to understand, we will only show the most important information such as the name of the gene, the genomic position or the underlying nucleotide change, which is obtained by sequencing, as well as a limited set of mutations. Although focusing on a small number of records, the sample data also includes some challenges, such as standardization problems. These kinds of problems are the first ones to target with our pipeline.

Table 1: Data from Vendor A

SUBJID	**SPEC**	**GENE**	**SOMSFI**	**AMACCH**	**NUCLCH**	**CHROM**	**CHRPOS**
Subject Identifier	Specimen of Biomarker	Gene	Somatic Status/Functional Impact	Amino Acid Change	Nucleotide Change	Chromosome	Chromosome Position
1	TUMOR BIOPSY	ALK	MISSENSE	Leu1196Met	C>A	2	29220765
1	TUMOR BIOPSY	BRCA1	SYNONYMOUS	Leu771=	T>C	17	41245237
3	TUMOR BIOPSY	KIT	MISSENSE	Asp816Gly	A>G	4	54733155
3	TUMOR BIOPSY	EGFR	MISSENSE	Pro546Ser	C>T	7	55163737
3	TUMOR BIOPSY	ROS1	MISSENSE	Gly2032Arg	G>A	6	117317184

Table 2: Data from Vendor B

SUBJID	**SPEC**	**GENE**	**SOMSTF**	**AMACCH**	**NUCLCH**	**GENPOS**
Subject Identifier	Specimen of Biomarker	Gene	Somatic Status/Functional Impact	Amino Acid Change	Nucleotide Change	Genome Position
2	PLASMA	CD246	MISSENSE_VARIANT	L1196M	C>A	Chr2:29220765
2	PLASMA	ROS1	SYNONYMOUS_VARIANT	L110L	A>T	Chr6:117725578
4	PLASMA	TP53	MISSENSE_VARIANT	C229R	A>G	Chr17:7577596
4	PLASMA	APEX1	MISSENSE_VARIANT	I146V	A>G	Chr14:20925016
4	PLASMA	ROS1	SYNONYMOUS_VARIANT	P160P	G>T	Chr6:117724399

As can be seen in **Tables 1&2,** our simulated case study received data from two different vendors, named Vendor A and B, since different specimens needed to be analyzed. Although providing the same kind of information, both datasets show differences in variable names (SOMSFI vs. SOMSTF) and the display of information (AMACCH: three-letter vs. one-letter code, CHROM & CHRPOS vs. GENPOS) due to varying internal vendor standards. To avoid differences in subsequent processes, both datasets are combined and aligned to CDISCs' GF domain standards (see **code below,** alignment of amino acid changes and the variables SOMSTF and GENPOS not shown).

```
DATA SDTM1;
    SET VENDORA VENDORB;
RUN;

PROC SORT DATA=SDTM1;
    BY SUBJID GENE AMACCH NUCLCH SOMSFI CHROM CHRPOS;
RUN;

PROC TRANSPOSE DATA=SDTM1 (DROP= SPEC) OUT=SDTM2;
    BY SUBJID GENE AMACCH NUCLCH SOMSFI CHROM CHRPOS;
    VAR AMACCH NUCLCH SOMSFI;
RUN;

DATA SDTM3;
    SET SDTM2;
    ATTRIB GFTSTDTL LENGTH=$50;
    *Create variable GFTSTDTL;
    IF _NAME_="AMACCH" THEN GFTSTDTL="Predicted Amino Acid Change";
    ELSE IF _NAME_="NUCLCH" THEN GFTSTDTL="Predicted Coding Sequence Change";
    ELSE IF _NAME_="SOMSFI" THEN GFTSTDTL="Variant Impact Classification";

    *Create other GF domain variables GFORRES, GFSYM, GFCHROM and GFGENLOC;
    RENAME COL1=GFORRES  GFSYM=GENE  GFCHROM=CHROM  GFGENLOC=CHRPOS;

    DROP GENE AMACCH NUCLCH SOMSFI CHROM CHRPOS _NAME_;
RUN;
```

```
PROC SORT DATA=SDTM3;
    BY SUBJID GFSYM GFTSTDTL;
RUN;

DATA SDTM4;
    RETAIN SUBJID GFSEQ GFGRPID GFTSTDTL GFCHROM GFSYM GFGENLOC;
    SET SDTM3;
    BY SUBJID GFSYM GFTSTDTL;

    *Count for GFSEQ;
    RETAIN GFSEQ;
    GFSEQ+1;
    IF FIRST.SUBJID THEN GFSEQ=1;

    *Count for GFGRPID;
    RETAIN GFGRPID 0;
    IF FIRST.GFSYM THEN GFGRPID+1;
    IF FIRST.SUBJID THEN GFGRPID=1;
RUN;

*Create format for SDTM dataset;

PROC SORT DATA=SDTM4 OUT=SDTM;
    BY SUBJID GFGRPID GFSEQ;
RUN;
```

The following **Table 3** displays the standardized SDTM data including one record per finding per sample within a subject. Variables such as GFNAM (Laboratory/Vendor Name) or GFSPEC (Specimen Material Type) are not presented in this table as those are irrelevant for the upcoming steps. Furthermore, GFTEST (Name of Genomic Measurement) is also not shown as we focus on single nucleotide variations (SNVs) only and would therefore show the same in each record. Besides the alignment into GF domain standards, amino acid changes in GFORRES have also been aligned to the same notation of the amino acid one-letter code.

Table 3: SDTM data

SUBJID	**GFSEQ**	**GFGRPID**	**GFTSTDTL**	**GFORRES**	**GFCHROM**	**GFSYM**	**GFGENLOC**
Subject Identifier	Seqeunce Number	Group ID	Measurement, Test, or Examination Detail	Result or Finding in Original Units	Chromosome Identifier	Genomic Symbol	Genetic Location
1	1	1	Predicted Amino Acid Change	L1196M	2	ALK	29220765
1	2	1	Predicted Coding Sequence Change	C>A	2	ALK	29220765
1	3	1	Variant Impact Classification	Missense variant	2	ALK	29220765
1	1	2	Predicted Amino Acid Change	L771L	17	BRCA1	41245237
1	2	2	Predicted Coding Sequence Change	T>C	17	BRCA1	41245237
1	3	2	Variant Impact Classification	Synonymous variant	17	BRCA1	41245237
2	1	1	Predicted Amino Acid Change	L1196M	2	CD246	29220765
2	2	1	Predicted Coding Sequence Change	C>A	2	CD246	29220765
2	3	1	Variant Impact Classification	Missense Variant	2	CD246	29220765
2	1	2	Predicted Amino Acid Change	L110L	6	ROS1	117725578
2	2	2	Predicted Coding Sequence Change	A>T	6	ROS1	117725578
2	3	2	Variant Impact Classification	Synonymous variant	6	ROS1	117725578
3	1	1	Predicted Amino Acid Change	D816G	4	KIT	54733155
3	2	1	Predicted Coding Sequence Change	A>G	4	KIT	54733155
3	3	1	Variant Impact Classification	Missense Variant	4	KIT	54733155
3	1	2	Predicted Amino Acid Change	P546S	7	EGFR	55163737
3	2	2	Predicted Coding Sequence Change	C>T	7	EGFR	55163737
3	3	2	Variant Impact Classification	Missense Variant	7	EGFR	55163737
3	1	3	Predicted Amino Acid Change	G2032R	6	ROS1	117317184
3	2	3	Predicted Coding Sequence Change	G>A	6	ROS1	117317184
3	3	3	Variant Impact Classification	Missense Variant	6	ROS1	117317184
4	1	1	Predicted Amino Acid Change	C229R	17	TP53	7577596

4	2	1	Predicted Coding Sequence Change	A>G	17	TP53	7577596
4	3	1	Variant Impact Classification	Missense Variant	17	TP53	7577596
4	1	2	Predicted Amino Acid Change	I146V	14	APEX1	20925016
4	2	2	Predicted Coding Sequence Change	A>G	14	APEX1	20925016
4	3	2	Variant Impact Classification	Missense Variant	14	APEX1	20925016
4	1	3	Predicted Amino Acid Change	P160P	6	ROS1	117724399
4	2	3	Predicted Coding Sequence Change	G>T	6	ROS1	117724399
4	3	3	Variant Impact Classification	Synonymous Variant	6	ROS1	117724399

As explained above, the next step of the pipeline is to extract unique variants from SDTM data as a basis for data annotation. Having unique variants means to retain one line per unique SNV change with gene name, amino acid as well as nucleotide change and the genome position. If a variant is present several times, it is only displayed once. Furthermore, in the variable GFTSTDTL one can see an entry called “Variant Impact Classification". This observation emphasizes whether a variant is synonymous, missense, or another kind of genetic variant. A missense variant or missense mutation is a mutation that introduces a new amino acid into the protein chain as a result of a point mutation, resulting in a different protein being expressed when the genetic change is in the coding region. If a variant is called synonymous, it means that there has also been a point mutation, but the new codon still codes for the same amino acid and the same protein is expressed. Consequently, synonymous variants do not have any impact on our research question, hence these entries can be filtered out before extracting the unique variants. Aligning variables into one row, deleting repeated variants, and filtering out synonymous variants, leaves 7 unique entries based on the simulated data (**Table 4**, program code not shown), which are potentially relevant in context of the research question. Here, we have a lucid number of entries, but in a real dataset the extent of data is likely still overwhelming. However, not all single nucleotide variants are cancer driver mutations and there are ways to identify them.

Table 4: Unique variants

GENE	**AMINO**	**NUCCH**	**CHROM**	**GENLOC**
Genomic Symbol	Amino Acid Change	Nucleotide Change	Chromosome Identifier	Genetic Location
ALK	L1196M	C>A	2	29220765
CD246	L1196M	C>A	2	29220765
KIT	D816G	A>G	4	54733155
EGFR	P546S	C>T	7	55163737
ROS1	G2032R	G>A	6	117317184
TP53	C229R	A>G	17	7577596
APEX1	I146V	A>G	14	20925016

Determination of cancer driver mutations and consequently narrowing down the number of data entries is done by data annotation and harmonization, the last part of the programming pipeline. During data annotation, various public databases can be accessed to obtain information that assigns meaning to both the genes and the individual mutations. One of these databases is the COSMIC (*Catalogue of Somatic Mutations in Cancer*) database, founded by Wellcome Trust Sanger Institute, which curates details about somatically acquired mutations in human cancer. As part of these details, COSMIC implemented the Cancer Gene Census (CGC) and the Cancer Mutation Census (CMC) as a standard description of cancer driving genes and their mutations across basic research, pharmaceutical development, and medical reporting.[11,12] While the CGC gives information about whether a gene is an oncogene or TSG, it also differentiates genes into groups or so-called tiers based on the scientific strength supporting the participation of each gene in oncogenesis. Further, the CMC differentiates mutations into tiers acknowledging their documented pathogenicity. Hence, these classifications can be used to filter only for pathogenic variants identified within a study. In this context, the CMC Tier classification (https://cancer.sanger.ac.uk/cmc/help) for example contains four categories: Tier 1 mutations are those classified as pathogenic and associated with the color red in the database; Tier 2 mutations (orange) are identified as likely pathogenic and Tier 3 as possibly pathogenic (yellow). All other mutations that are not reported to be pathogenic, are grouped in Tier 4 (grey).

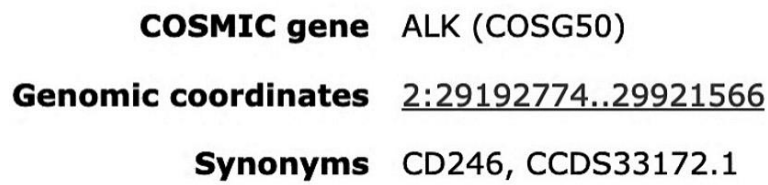

Figure 3: COSMIC search for gene CD246

Furthermore, the COSMIC database is also of great help when it comes to data harmonization. As we have already seen, if labs have different standards in reporting, say, amino acid changes, then there is a high probability that there are also other differences present. In fact, genes also have several synonyms that also need to be harmonized while doing data annotation. For example, when searching for the gene CD246 (see line 4 **Table 4**), COSMIC provides the information that it is the same as ALK (**Figure 3**). For such cases a new variable can be added giving approved gene names or symbols as it was done in **Table 5** (GENSYM). In fact, this harmonization is a prerequisite for doing a proper annotation as, for example, ALK is listed as the COSMIC gene name and not CD246. For all further annotation steps, such as the identification of the gene type or the CMC Tiers, ALK or any other COSMIC gene synonym needs to be used for querying. After harmonization of the underlying information is done, one can start extracting information from the database. As an example, below, we will extract some information for our simulated unique variants by hand. Normally there are significantly more unique entries in a data set, so all needed information is obtained via a programmatic query.

```
ALK   ALK receptor tyrosine kinaseCOSG18163   2     29415640    30144432
      2p23.2      y     y     ALCL, NSCLC, neuroblastoma, inflammatory
myofibroblastic tumour, Spitzoid tumour neuroblastoma     familial
neuroblastoma     L, E, M     Dom   oncogene, fusion T, Mis, A   NPM1, TPM3,
TFG, TPM4, ATIC, CLTC, MSN, RNF213, CARS, EML4, KIF5B, C2orf22, DCTN1,
HIP1, TPR, RANBP2, PPFIBP1, SEC31A, STRN, VCL, C2orf44, KLC1  n
      1     ALK,ENSG00000171094.11,Q9UM73,238,CD246
```

Figure 4: Cosmic Cancer Gene Census Data (CGC) for gene ALK

As previously mentioned, genes can be differentiated into three gene types: oncogenes, TSGs, or ambiguous ones that can function in both ways. The knowledge of the associated gene type is one key factor in the interpretation of the underlying biology of mutational variants. To extract the associated gene type, data from the COSMIC database can be downloaded with a choice of selecting the Genome Reference Consortium Human Build 37 (GRCh37) or 38 (GRCh38). Here, it is crucial to select the same build that was used by the vendor. This kind of information is usually given in the data provided by the vendor. For the simulated case study, GRCh37 was used, and the gene type information can be found in the TSV file "COSMIC_CancerGeneCensus_v100_GRCh37". The mentioned file is only a very limited open-source document, but luckily includes data for the gene ALK (**Figure 4**). As shown in **Figure 4,** ALK is classified as an oncogene and further identified to be a potential fusion gene, which is the case in our simulated data. Doing this for all genes in our sample data, only the gene APEX1 could not be linked to any gene type and can thus be considered as irrelevant (**Table 5**).

```
EGFR  ENST00000275493.2      oncogene    1
      https://cancer.sanger.ac.uk/cosmic/mutation/overview?id=22307044
      COSM1238072 c.1636C>T   p.P546S     546   546   2     C     T     P
      S     Substitution      Substitution - Missense      SO:0001583
      COSV51837871      7:55231430-55231430    7:55163737-55163737
      166248      1

                                Likely pathogenic      Head and neck
neoplasm    5.84  0.095 T       3
```

Figure 5: COSMIC Cancer Mutation Census (CMC) Data for EGFR mutation P546S

The same kind of information extraction can also be done for the CMC Tier classifications using the freely available test TSV file "CancerMutationCensus_AllData_v100_GRCh37". **Figure 5** shows example data for the EGFR mutation P546S, which is classified as TIER 3 correlated with the group of genes of being possibly pathogenic**.** The Tier classification was extracted for all mutations and included into the sample data (**Table 5**). Please note that not all variants shown are mentioned in the open-source documents and thus have been annotated through a licensed account.

Table 5: Unique variants with COSMIC annotation

GENE	AMINO	NUCCH	CHROM	GENLOC	GENSYM	COGENTYP	COMUTTYP
Genomic Symbol	Amino Acid Change	Nucleotide Change	Chromosome Identifier	Genetic Location	Approved gene symbol	Cosmic Gene Type	Cosmic Mutation Type
ALK	L1196M	C>A	2	29220765	ALK	oncogene, fusion	Red: pathogenic
CD246	L1196M	C>A	2	29220765	ALK	oncogene, fusion	Red: pathogenic
KIT	D816G	A>G	4	54733155	KIT	oncogene	Grey: unknown significance
EGFR	P546S	C>T	7	55163737	EGFR	oncogene	Yellow: possibly pathogenic
ROS1	G2032R	G>A	6	117317184	ROS1	oncogene, fusion	Grey: unknown significance
TP53	C229R	A>G	17	7577596	TP53	oncogene, TSG, fusion	Yellow: possibly pathogenic
APEX1	I146V	A>G	14	20925016	APEX1	/	/

After annotation of the unique variants, it is necessary to merge the new information to the SDTM data to create a final ADaM data set (see **code below).** This can be done in two different merging steps. First, the annotated data (**Table 5**) are merged back to a reduced SDTM dataset which contains all entries with the predicted amino acid change via the variables GFSYM:GENE and GFORRES:AMINO. Afterwards, one can extract all given annotation variables together with the SUBJID and GFGRPID and merge it back to the whole SDTM dataset (**Table 3**) resulting in the final ADaM dataset (**Table 6**). For visualization purposes the variables GENLOC, NUCCH and CHROM were removed in the table below, but they are still part of the dataset.

```
*extract all records from SDTM with predicted amino acid changes;
PROC SORT DATA=SDTM (WHERE=(GFSEQ=1)) OUT=SDTM_RED;
    BY GFSYM GFORRES;
RUN;

PROC SORT DATA=ANNO(RENAME=(GENE=GFSYM AMINO=GFORRES)) OUT=ANNO_SORT;
    BY GFSYM GFORRES;
RUN;

*merge of reduced SDTM data with annotated data, extract annotated variables with ID variables;
DATA HELP;
    MERGE SDTM_RED(IN=A) ANNO_SORT(IN=B);
    BY GFSYM GFORRES;
    IF A;

    KEEP SUBJID GFGRPID GENSYM COGENTYP COMUTTYP;
RUN;

PROC SORT DATA=HELP;
    BY SUBJID GFGRPID;
RUN;

*merge of annotated data via ID variables to complete SDTM dataset;
DATA ADAM;
    MERGE SDTM(IN=A) HELP(IN=B);
    BY SUBJID GFGRPID;
RUN;
```

As seen in the final ADaM dataset, not all records have information in the annotation variables. This is because the synonymous variants were not annotated and additionally for the APEX1 mutation no entry was found in the COSMIC database.

Table 6: ADaM data

SUBJID	GFSEQ	GFGRPID	GFTSTDTL	GFORRES	GFSYM	GENSYM	COGENTYP	COMUTTYP	MUTFL
Subject Identifier	Seqeunce Number	Group ID	Measurement, Test, or Examination Detail	Result or Finding in Original Units	Genomic Symbol	Approved gene symbol	Cosmic GeneType	Cosmic Mutation Type	Flag for relevant mutations
1	1	1	Predicted Amino Acid Change	L1196M	ALK	ALK	oncogene, fusion	Red: pathogenic	YES
1	2	1	Predicted Coding Sequence Change	C>A	ALK	ALK	oncogene, fusion	Red: pathogenic	YES
1	3	1	Variant Impact Classification	Missense variant	ALK	ALK	oncogene, fusion	Red: pathogenic	YES
1	1	2	Predicted Amino Acid Change	L771L	BRCA1				
1	2	2	Predicted Coding Sequence Change	T>C	BRCA1				
1	3	2	Variant Impact Classification	Synonymous variant	BRCA1				
2	1	1	Predicted Amino Acid Change	L1196M	CD246	ALK	oncogene, fusion	Red: pathogenic	YES
2	2	1	Predicted Coding Sequence Change	C>A	CD246	ALK	oncogene, fusion	Red: pathogenic	YES
2	3	1	Variant Impact Classification	Missense Variant	CD246	ALK	oncogene, fusion	Red: pathogenic	YES
2	1	2	Predicted Amino Acid Change	L110L	ROS1				
2	2	2	Predicted Coding Sequence Change	A>T	ROS1				
2	3	2	Variant Impact Classification	Synonymous variant	ROS1				
3	1	1	Predicted Amino Acid Change	D816G	KIT	KIT	oncogene	Grey: unknown significance	
3	2	1	Predicted Coding Sequence Change	A>G	KIT	KIT	oncogene	Grey: unknown significance	
3	3	1	Variant Impact Classification	Missense Variant	KIT	KIT	oncogene	Grey: unknown significance	
3	1	2	Predicted Amino Acid Change	P546S	EGFR	EGFR	oncogene	Yellow: possibly pathogenic	
3	2	2	Predicted Coding Sequence Change	C>T	EGFR	EGFR	oncogene	Yellow: possibly pathogenic	
3	3	2	Variant Impact Classification	Missense Variant	EGFR	EGFR	oncogene	Yellow: possibly pathogenic	
3	1	3	Predicted Amino Acid Change	G2032R	ROS1	ROS1	oncogene, fusion	Grey: unknown significance	
3	2	3	Predicted Coding Sequence Change	G>A	ROS1	ROS1	oncogene, fusion	Grey: unknown significance	
3	3	3	Variant Impact Classification	Missense Variant	ROS1	ROS1	oncogene, fusion	Grey: unknown significance	
4	1	1	Predicted Amino Acid Change	C229R	TP53	TP53	oncogene, TSG, fusion	Yellow: possibly pathogenic	
4	2	1	Predicted Coding Sequence Change	A>G	TP53	TP53	oncogene, TSG, fusion	Yellow: possibly pathogenic	
4	3	1	Variant Impact Classification	Missense Variant	TP53	TP53	oncogene, TSG, fusion	Yellow: possibly pathogenic	
4	1	2	Predicted Amino Acid Change	I146V	APEX1	APEX1			
4	2	2	Predicted Coding Sequence Change	A>G	APEX1	APEX1			
4	3	2	Variant Impact Classification	Missense Variant	APEX1	APEX1			
4	1	3	Predicted Amino Acid Change	P160P	ROS1				
4	2	3	Predicted Coding Sequence Change	G>T	ROS1				
4	3	3	Variant Impact Classification	Synonymous Variant	ROS1				

Returning to the initial theme of this paper, the overall aim was to identify those individuals harboring an acquired resistance mutation that leads to crizotinib resistance and therefore the pipeline should narrow down the variant entries to those that are relevant to answering this question. Now that the COSMIC gene types and mutation classifications have been mapped to the SDTM data to create the final ADaM dataset (**Table 6**), a new variable can be created to flag those variants that are relevant. In this scenario the new flag (MUTFL) is set to "YES" if a gene type could be assigned to a gene and if the mutation is classified as red ("pathogenic") or orange ("likely pathogenic"). **Table 6** shows that there are two individuals (Subject 1 and 2) harboring a relevant mutation. In both cases it is the missense mutation L1196M which was already identified in previous studies as to induce resistance in patients with crizotinib treatment.[13,14]

CHALLENGES

As shown through the programmatic pipeline, working with genomics data is fraught with many challenges. One of the biggest challenges during this process is that vendors have different internal standards and therefore combining all delivered datasets can be difficult. Besides the issues already addressed in this pipeline, there are far more things that can complicate data processing and especially data annotation. For example, one vendor may display nucleotide changes from the forward strand and another vendor may use the reverse strand without displaying this type of information in their data transfers. During annotation, this can cause problems in the alignment process. Other problems may include truncation of data, for example when complex nucleotide changes can't be represented within the constraints of a SAS® data file. For a statistical programmer without data annotation know-how, it is hard to identify such issues. For such purposes, it is best to have a bioinformatician in the study team that can help with data annotation. In general, it is important to stay in regular contact and discussions with the whole study team including statisticians, statistical programmers, and bioinformaticians. Depending on the number of samples, specimens, and assays, the amount of data might be overwhelming making it important to create rules on how data should be analyzed and presented. One way to target such questions is to narrow down the number of entries to only those relevant for the research question. In this pipeline it was decided to just identify mutations as being relevant if they are classified as Tier 1 and 2 by the CMC of the COSMIC database, but in case of exploratory purposes, including Tier 3 might also be valuable. Such decisions should be based on the type of study, the study design and above all the underlying research questions and its audience. Another way of narrowing down the relevant mutations could be to focus only on specific genes such as the target gene itself or genes which are part of the downstream signaling. For this, it would be an advantage to have someone in the team who has a background in biology, be it a statistician, bioinformatician, or statistical programmer. Besides focusing on the genetic findings itself it is also challenging to clarify if an identified mutation meets the criteria of being an acquired mutation or not. For this, one needs to take into account the point int time of when the mutation was identified and the detection capacity of the used assay, but this goes beyond the scope of this paper. In addition, SDTM and ADaM programming as well as working on TLFs can be challenging too as there are no established standards yet, as the whole field of biomarker analysis and genomic findings is relatively new in pharmaceutical studies. However, how to present information in a meaningful way, or even perform exploratory analysis, is its own topic that could be addressed in the future.

CONCLUSION

Recently, precision medicine and genomic data as biomarkers have become hot topics in the pharmaceutical industry. In this paper, a programming pipeline has been presented to give insights into this challenging area of work from a statistical programmer's point of view. Beginning with vendor NGS data, it was shown how to gain a final ADaM data set going through steps like data standardization in accordance with the SDTM GF domain, identifying unique small nucleotide variants, data harmonization and in the end, data annotation by identifying gene types and mutation classes that help interpreting genomic data. As a statistical programmer, it can be challenging to work in the field of genomics, especially when having a deficiency in biological background or bioinformatics skills. Luckily, programming of datasets got clearer based on the implementation of CDISC standards for SDTM and ADaM programming. In general, a tight cooperation between the statistical programmer and a bioinformatician is needed to bypass a potential lack of biological knowledge. Although working with genomic data is not easy, operating in this field is a great way for statistical programmers to prove themselves and learn more through all its challenges.

REFERENCES

1. Goetz, L. H. & Schork, N. J. Personalized medicine: motivation, challenges, and progress. *Fertil Steril* **109**, 952–963 (2018).

2. President´s Council of Advisors on Science and Technology. Priorities for Personalized Medicine. 19–19 Preprint at (2008).

3. Li, X. *et al.* The multi-molecular mechanisms of tumor-targeted drug resistance in precision medicine. *Biomedicine & Pharmacotherapy* **150**, 113064 (2022).

4. Soda, M. *et al.* Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer. *Nature* **448**, 561–566 (2007).

5. Rotow, J. & Bivona, T. G. Understanding and targeting resistance mechanisms in NSCLC. *Nat Rev Cancer* **17**, 637–658 (2017).

6. Morris, T. A., Khoo, C. & Solomon, B. J. Targeting ROS1 Rearrangements in Non-small Cell Lung Cancer: Crizotinib and Newer Generation Tyrosine Kinase Inhibitors. *Drugs* **79**, 1277–1286 (2019).

7. Lin, J. J., Riely, G. J. & Shaw, A. T. Targeting ALK: Precision Medicine Takes on Drug Resistance. *Cancer Discov* **7**, 137–155 (2017).

8. Mansoori, B., Mohammadi, A., Davudian, S., Shirjang, S. & Baradaran, B. The Different Mechanisms of Cancer Drug Resistance: A Brief Review. *Adv Pharm Bull* **7**, 339–348 (2017).

9. Redaelli, S. *et al.* Lorlatinib Treatment Elicits Multiple On- and Off-Target Mechanisms of Resistance in ALK-Driven Cancer. *Cancer Res* **78**, 6866–6880 (2018).

10. FDA-NIH Biomarker Working Group. *BEST (Biomarkers, EndpointS, and Other Tools) Resource*. (2016).

11. Sondka, Z. *et al.* COSMIC: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Res* **52**, D1210–D1217 (2024).

12. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* **18**, 696–705 (2018).

13. Katayama, R. *et al.* Mechanisms of Acquired Crizotinib Resistance in ALK-Rearranged Lung Cancers. *Sci Transl Med* **4**, (2012).

14. Poei, D., Ali, S., Ye, S. & Hsu, R. ALK inhibitors in cancer: mechanisms of resistance and therapeutic management strategies. *Cancer Drug Resistance* (2024) doi:10.20517/cdr.2024.25.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Lea Zittlau, Dr. Saskia Ständler
Company: Chrestos Concept GmbH & Co. KG
Address: Girardetstraße 1-5, Essen, Germany, 45131
Email: lea.zittlau@chrestos.de, saskia.staendler@chrestos.de
Website: https://chrestos.de/