A Scalable Architecture for Smart Genomic Data Analysis in Medical Laboratories

9

Thomas Krause, Elena Jolkver, Michael Kramer, Paul McKevitt und Matthias L. Hemmje

Inhaltsverzeichnis

T. Krause (✉) · E. Jolkver · M. L. Hemmje
Faculty for Mathematics and Computer Science, University of Hagen, Hagen, Deutschland
E-mail: thomas.krause@fernuni-hagen.de

E. Jolkver
E-mail: elena.jolkver@studium.fernuni-hagen.de

M. L. Hemmje
E-mail: matthias.hemmje@fernuni-hagen.de

M. Kramer
ImmBioMed Business Consultants GmbH, Pfungstadt, Deutschland
E-mail: m.kramer@immbiomed.de

P. McKevitt
School of Arts & Humanities, Ulster University, Derry, UK
E-mail: p.mckevitt@ulster.ac.uk

© Der/die Autor(en), exklusiv lizenziert an Springer Fachmedien Wiesbaden GmbH, ein Teil von Springer Nature 2023

L. B Blum (Hrsg.), *Angewundte Data Science*,
https://doi.org/10.1007/978-3-658-39625-1_9

Abstract

Genomic data is an important building block for the era of personalized medicine. However, processing this data efficiently in diagnostic laboratories faces several challenges in distinct areas such as big data, artificial intelligence, regulatory environment, medical/diagnostic standards (evolving guidelines), and software requirements engineering.

Analysis of the state of the art in these areas shows promising approaches and suitable reference models but no direct solutions. Existing technical solutions for genomic data analysis tend to be specialized for research projects and do not take into account the requirements for routine medical diagnostics including the regulatory constraints in this area. This chapter introduces a technical architecture for the GenDAI (Genomic applications for laboratory Diagnostics supported by Artificial Intelligence) project that aims to create a platform for genomic data analysis that is specifically tailored to the needs and requirements of laboratory diagnostics. This includes the automation of processes using data analysis pipelines and artificial intelligence.

9.1 Introduction

Personalized medicine promises to improve the medical standard of care through diagnoses and treatments tailored to the individual patient (Goetz and Schork 2018). The basis for this is the evaluation of biomedical data. Genomic data plays a particular role here (Suwinski et al. 2019).

9.1.1 Genomic Data in Personalized Medicine

The pathological condition of a patient and his/her adequate treatment can be significantly influenced by his/her genetic make-up (Suwinski et al. 2019). Hence, individual sequencing of the genetic material or identification of single relevant alleles is an important basis for laboratory diagnoses. Evidently, this applies to hereditary diseases. However, apart from hereditary diseases, genomic applications may also provide expedient information for the diagnosis and treatment of acquired pathological conditions (Gebrayel et al. 2022). Examples here are, e.g., gene expression analysis and metagenomics.

The goal of gene expression analysis is to determine the activity of specific genes that are, e.g., linked to pathological conditions. A gene—i.e., a stretch of genomic (nuclear) DNA—is called active when it is transcribed particularly frequently by cell-internal mechanisms, i.e., if it is transferred into an mRNA (messenger RiboNucleic Acid) sequence. In a second step, mRNA is then translated into proteins, which then transform genetic information into metabolic activity in living organisms. Thus, one method for determining gene activity is the determination of the number of mRNA copies of given genes by the so-called RT-qPCR (Reverse Transcription quantitative Polymerase Chain Reaction) (Adams 2020).

Since direct measurement of the number of mRNA copies is not possible or only possible with great effort, the RT-qPCR method uses enzymes that are able to select specific mRNA segments, amplify them and make them visible by fluorescent substances. Figure 9.1 shows this process. First, mRNA consisting of a single strand is transcribed into cDNA (complementary DNA), which consists of two strands but contains the same information. The double strands enable amplification by separating the two strands and then complementing them to form a second identical double strand. Various methods allow fluorescent substances to be used in this duplication, which leads to an increase in fluorescence as the number of copies increases. Running through enough duplication cycles, a "chain" of polymerase reactions, produces a clear fluorescent signal that can be clearly distinguished from background noise, indicating the presence of the target sequence. If the number of amplification cycles that are needed to achieve a clearly measurable fluorescence is quantified, conclusions can be drawn as to the original number of copies present in a given sample. Assuming that with each cycle the amount of target sequences is approximately doubled, a sample with 1000 initial mRNA

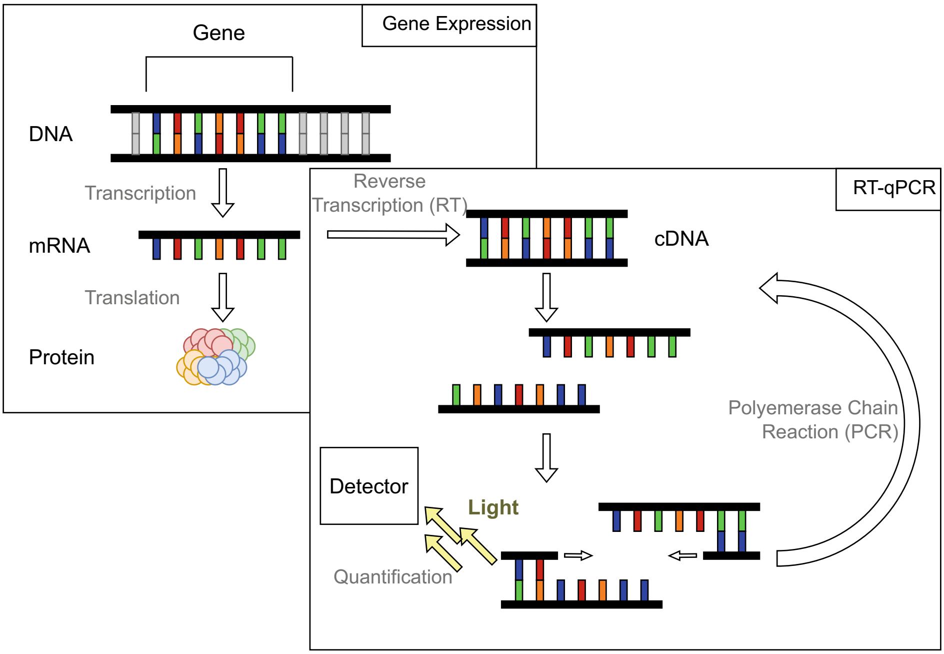


Fig. 9.1 Gene expression and RT-qPCR method for gene expression quantification

copies would need about three cycles more to achieve the same fluorescence as a sample with 8000 copies. This cycle number is called C_q or sometimes C_t. On the basis of the C_q values, further calculations can be made in the next step to normalize them and, using limit values or more complicated formulas, to provide diagnostic indications. Thus the raw data in the RT-qPCR method consist of the fluorescence signals of the individual samples over the duration of the cycles.

DNA microarrays, also called DNA-chip technology, are an alternative method to measure gene expression. Molecular probes are firmly attached to a carrier material. These probes specifically react with mRNA or cDNA sequences and bind tightly to them. Again, quantification can subsequently be performed by measuring fluorescence. In contrast to the qPCR method, the measurement range is significantly reduced and the quantification less precise. The fluorescence signal is only evaluated once at the end ("end-point method"). In other words, the raw data in this method consists of fluorescence values for the individual samples without a time course. In turn, up to several million probes can be attached to a single chip and tested in parallel. Hence, microarrays are well suited for obtaining an overall profile of the gene activity of specific cells, while qPCR is more suitable for the precise investigation of the activity of individual genes in multiple samples.

Finally, gene expression can also be determined by sequencing (RNA-seq). Instead of working with solid probes that recognize specific sequences, sequencing is able to read the existing mRNA/cDNA segments directly without the need to know them in advance. Here, too, different methods have been developed (Hong et al. 2020). Usually, similar to PCR, the natural amplification mechanism of DNA is used, in which a single strand is completed to form a complementary double strand. However, instead of allowing this process to proceed in an uncontrolled manner, it is observed using fluorescence signals to detect each base added to the strand. The advantage of this method is that any sequence can be detected, eliminating the need to develop or purchase special microarrays. Also, the measuring range and the accuracy of quantification are increased with RNA-seq compared to microarrays. A disadvantage can be seen in the significantly larger volume of data that is generated by sequencing which must be stored and processed. Sequencing has also traditionally been more expensive than chip technology, making the latter more common. However, with costs in the field decreasing rapidly, RNA-Seq is increasingly replacing DNA chip technology. The raw data for the sequencing method consists of an unordered collection of "reads" containing the individual bases of cDNA segments. By grouping these reads by sequence similarity (called "binning"), quantification can be performed.

Another example of genomic applications in diagnostics is metagenomics. The importance of microorganisms such as bacteria in and on the human body has been increasingly recognized in recent years. These microorganisms can be found, for example, in blood, saliva, or the gut. The assemblage of microorganisms in these defined environments are called "Microbiota" (Marchesi and Ravel 2015). Metagenomics is the study of these microbiota by characterizing the genomes and genes (the "metagenome") of its members.

Sequencing, the determination of the genome sequences, can be used to identify and categorize microorganisms or relevant genes. Based on the composition of microbiota, pathological conditions can be detected in medicine or general conclusions can be drawn about a patient's condition. Microbiota have been shown to influence metabolism (Fan and Pedersen 2021), mental health (Berding et al. 2021), diseases (Chiu and Miller 2019), disease treatment response (Zhang et al. 2019), and many other physiological and pathological conditions (Gebrayel et al. 2022; Daniel et al. 2021).

Microorganisms can be categorized on the basis of their evolutionary relationships and placed taxonomically in a tree of descent (the phylogenetic tree). In such a tree, evolutionarily closely related organisms whose genome sequences differ little are arranged close to each other, while distantly related organisms with greater differences in the genome sequence have a greater distance in the tree. At the top level of this tree the three domains of life: bacterias, archaea, and eukaryota are found. At the lowest levels, individual species or individual strains are found. In the context of a taxonomy, relevant branches of such a phylogenetic tree are assigned names, which can then be used for diagnostics and the generation of findings. The extent of diagnostics or sequencing determines how detailed, i.e. on which level of the phylogenetic tree the composition can be represented.

9.1.2 Current Challenges

As the aforementioned use cases show, medical data in genomics has high heterogeneity. Hence, the nature and order of the necessary processing steps differ considerably. Moreover, particularly in the field of genomic and metagenomic sequence analysis, large quantities of data in the range of many hundreds of gigabytes are being generated (Liu et al. 2021).

Due to the increasing importance of genomic applications in medicine, the speed at which new data are available and need to be processed is also increasing (Stephens et al. 2015). Figure 9.2 shows the exponential increase of sequence data including whole-genome sequencing records in the NCBI GenBank reference database (National Center for Biotechnology Information 2022) from December 1982 to February 2022, as the number of megabases on a logarithmic scale.

Big Data applications are commonly defined by the three criteria Variety, Volume, and Velocity (Abawajy 2015). All three criteria are present in at least some of the genomic applications as described, and thus data processing in genomics or sub-genomics can be considered a Big Data problem. This presents challenges for systems that seek to process this data. One of these challenges is the automated extraction of information from this data, as manual processing of all data is often not possible. A possible solution to this is the use of Artificial Intelligence (AI) and, in particular, Machine Learning (ML) to extract information automatically. However, this solution approach comes with its own challenges. First, for example, a suitable problem description must be found that can be solved using ML.

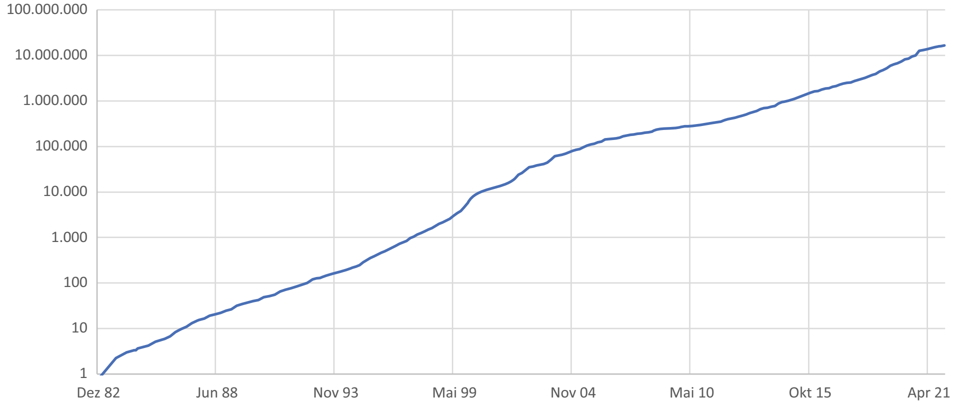


Fig. 9.2 Volume of sequencing data in NCBI GenBank database in megabases over time (National Center for Biotechnology Information 2022)

Then, this problem description must be transformed into suitable models and the established hypotheses must be carefully tested (Mc Kevitt and Partridge 1991).

Laboratory diagnostics are also subject to a great many regulatory challenges. In the European Union, for example, the In Vitro Diagnostic Regulation (IVDR) (The European Parliament and the Council of the European Union 2017) regulates the use of in vitro diagnostics. Similar regulations exist in other countries. In addition, laboratories must meet the requirements of various international and national standards.

A characteristic of laboratory diagnostics is the constant change resulting from advances in science and technology, including changes in the regulatory landscape. Laboratories must respond to this change and constantly adapt their procedures to keep up with the current state of the art. This not only ensures their continued competitiveness but is also increasingly an obligation by law. For example, this principle is reflected in the IVDR (The European Parliament and the Council of the European Union 2017) in the obligation of in vitro diagnostics providers to comply with pre-analytical, analytical and post-analytical requirements. Moreover, comprehensive market intelligence is required under the headlines, “Post-Marketing Surveillance” (PMS) and “Post-Marketing Performance Follow-Up” (PMPF). The many challenges from very different disciplines such as medicine, law, and computer science also make it difficult to gain comprehensive oversight over requirements. There is no established procedure for systematically determining these requirements which carries the risk that key elements may be overlooked.

9.1.3 Methodology

The analysis of genomic data in medical laboratories is a multi-faceted problem area. This raises the question of how information technology solutions must be designed to enable medically sound, efficient, automated, legally secure, and intelligent processing of genomic data within medical laboratories. To answer this question, the current state of research must be investigated first. Subsequently, based on these findings, a technical architecture can be designed that addresses the identified challenges, taking into account the state of the art in the individual sub-aspects, and integrating them.

Hence, the remainder of this chapter is structured as follows: Sect. 9.2 discusses the sub-areas of the state of the art, Sect. 9.3 proposes a technical architecture, and Sect. 9.4 discusses prospects for the next steps toward a complete system.

9.2 State of the Art

A thorough investigation of the State of the Art requires an analysis of relevant reference models, the possibilities of AI in genomics, architectural patterns for suitable system architectures, the regulatory framework under which laboratories operate, the current state in laboratories, techniques for requirements engineering, and systems for orchestrating and automating data analyses.

9.2.1 Reference Models

In the field of data analysis, various conceptual models attempt to standardize and formalize the process. For example, the CRISP-DM (CRoss-Industry Standard Process for Data Mining) (Chapman et al. 2000) model is well-known. It divides the process into the phases, “Business Understanding”, “Data Understanding”, “Data Preparation”, “Modeling”, “Evaluation”, and “Deployment”, whereby these phases are not strictly linear. Other well-known standard models include KDD (Knowledge Discovery in Databases) and SEMMA (Sample, Explore, Modify, Model, and Assess) (Shafique and Qaiser 2014). These standard models have been adapted and modified for different applications to represent better specific requirements in the domains and in turn define standards for processes. Examples are the CRISP4BigData model for Big Data processes (Berwind et al. 2016) and the AI2VIS4BigData model of Reis et al. (2021) for the combination of information visualization, Big Data, and AI.

Specifically developed for the requirements of laboratory diagnostics is the GenDAI (Genomic applications for laboratory Diagnostics supported by Artificial Intelligence) model (Krause et al. 2021a) (Fig. 9.3). The model was developed iteratively from the

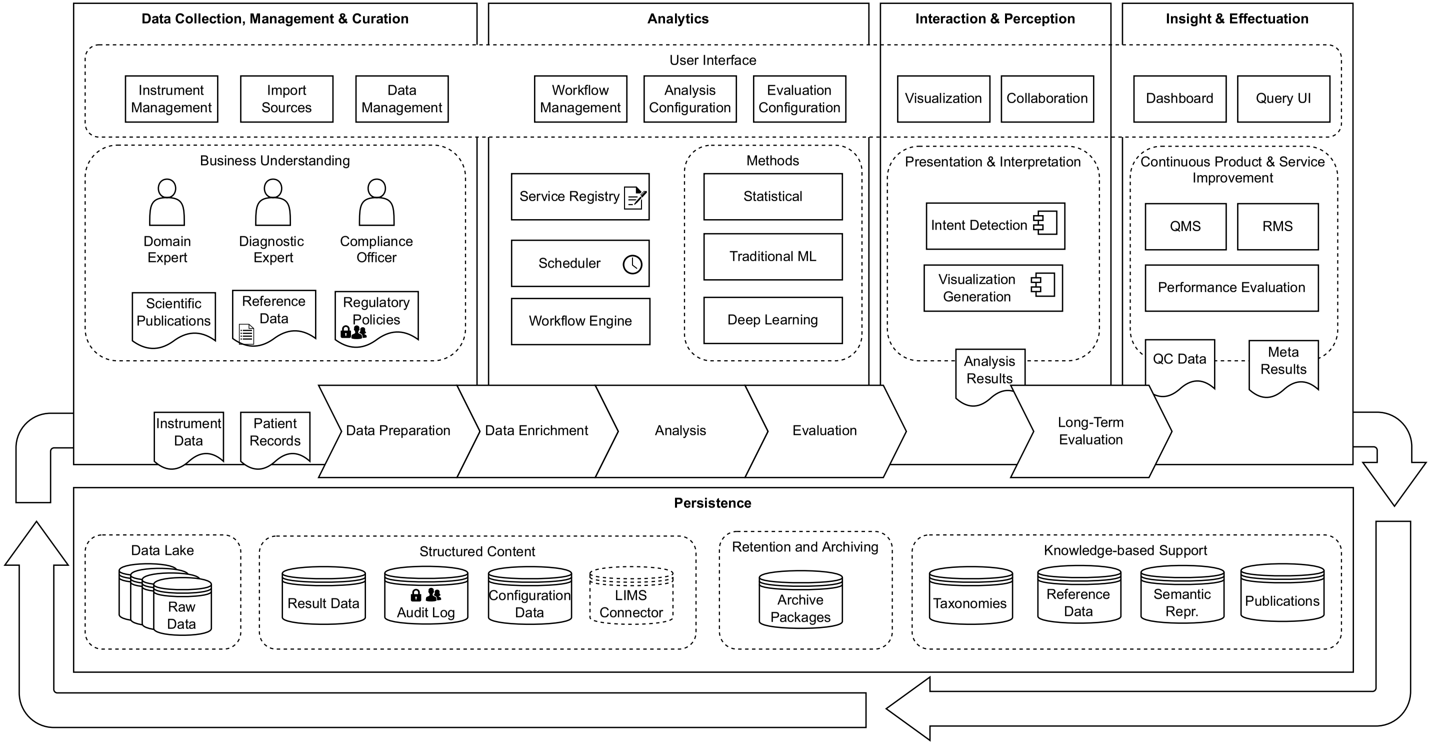


Fig. 9.3 GenDAI conceptual model. (©2021 IEEE. Reprinted, with permission, from Krause et al. (2021a))

AI2VIS4BigData reference model in which the requirements from genetics and the regulatory area were particularly taken into account.

9.2.2 Machine Learning

ML is a valuable tool for the analysis and classification of genomic data. By comparing the composition of microorganisms or their genes in samples from healthy and diseased humans, biomarkers can be developed in metagenomics, which in turn can be used for diagnosis. For example, Armour et al. (2019), used a random forest model in a meta-analysis to show how certain gene families of microorganisms correlate with the development of type 2 diabetes and other diseases. Figure 9.4 shows an example of the composition of samples from human gut microbiota in a sunburst diagram created with Krona (Ondov et al. 2011). The data from Qin et al. (2010) are samples from 124 individuals whose gut microbiota was examined for commonalities. Sunburst diagrams are used to show both abundance ratios and the

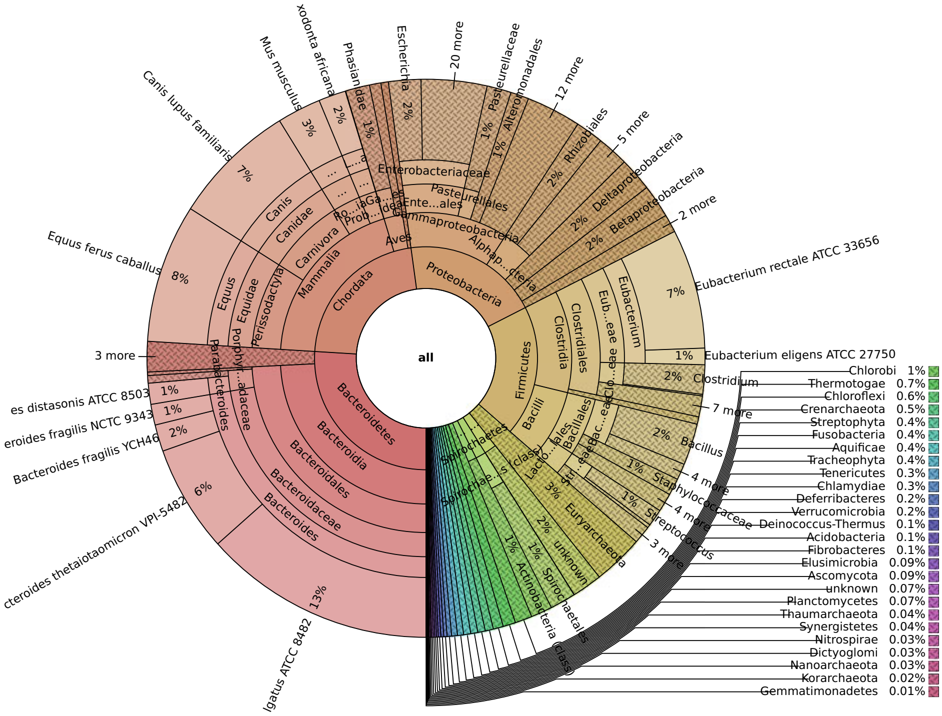


Fig. 9.4 Composition diagram created with Krona (Ondov et al. 2011) with data from (Qin et al. 2010)

phylogenetic structure of the data. In the innermost ring, the roots of the phylogenetic tree are represented. In the further rings, branchings within the phylogenetic tree are represented.

However, the consideration of evolutionary relationships is not only relevant for the representation, but also for the evaluation and classification of the data with ML. If a classifier is trained only on the level of the recognized species of a microbiome sample without taking into account how the individual species are evolutionarily related, important information will be lost. For example, it may happen that sequences of a sample cannot be assigned to a known species or that a recognized species was not part of the training material. However, at higher levels of the tree, these sequences would be clearly assignable to specific families, which may be sufficient for diagnostics. This consideration of evolutionary relationships is thus an active area of research using techniques such as Convolutional Neural Networks (CNN) (Fioravanti et al. 2018), Support Vector Machines (SVM), or Random Forest (RF) (Wassan et al. 2019).

Gene expression data are also suitable for classification using ML (Kuo et al. 2004). However, the focus here is more on data obtained from microarrays or RNA-seq, since in the RT-qPCR method usually only a few genes are considered and these can be evaluated with simple threshold methods or formulas. ML can also be used to guide clinical genomic testing (Dias and Torkamani 2019). Examples of this include the use of facial image recognition for rare genetic disorders (Gurovich et al. 2019) or predicting somatic mutations of cancer cells in histopathological images (Coudray et al. 2018). Thus, the findings generated by ML models based on phenotypic features can be used as recommendations for genetic testing.

Deep learning has proven particularly useful in recent years for recognizing patterns in large data sets (LeCun et al. 2015; Goodfellow et al. 2016). Deep Learning promises to do this without significant preprocessing of data. Feature selection, which is necessary for classical ML methods, also becomes less important in Deep Learning. Thus, Deep Learning algorithms are able to extract information from many features with high noise. This gain is usually bought by the necessity of large data sets and high computing capacities. Deep learning is also finding its way into genomic applications. Perhaps the most prominent example is the neural network, *AlphaFold2,* developed by Google (Jumper et al. 2021), which is able to predict the structure of proteins with high accuracy.

Many diseases arise from a complex interplay of genetic factors, environmental influences, and behavior (Dias and Torkamani 2019). Thus, genetic factors represent only part of the total individual risk and algorithms that determine a risk assessment based only on genomic data are limited in their performance and predictive power. Combining genomic data with other health data can significantly increase performance. In the future, Deep Learning could help to combine heterogeneous data from various sources such as electronic patient records and health trackers with genomic data to generate better diagnoses.

Despite these advantages, simple ML models (Soueidan and Nikolski 2015) currently dominate most genomic applications. Possible reasons for this are, (i) lack of expertise in model selection, (ii) the need for explainable and reproducible AI, (iii) Big Data challenges, and, (iv) lack of extensive training samples (Krause et al. 2021b). The selection of suitable models for a specific problem should be completed in a structured way, by first setting up hypotheses, which are then tested with preselected models (Mc Kevitt and Partridge 1991).

To employ ML models in the regular operation of a laboratory, these must be made available as IT services in order to be usable. The prerequisite for this is a suitable infrastructure.

9.2.3 Architectural Patterns

For the development of software technical architectures, there are various design patterns for distributed systems that have proven effective. The service-oriented architecture (SOA) approach (Gilbert 2021) creates an abstraction layer between the Domain Model, consisting of entities, and the applications which work with these entities. This abstraction layer models thereby no longer the entities themselves but the logic around certain executable processes or actions. Such a process can also include several entities. Related actions are grouped into services. SOA also offers a certain degree of scaling, since individual services can be deployed and scaled on different nodes.

Microservices can be regarded as a further development of a SOA. While the services in a SOA typically access a common backend, e.g. in the sense of an Enterprise Service Bus (ESB) (Gilbert 2021), microservices are more strongly separated from one another and form a bounded context. Hence, a microservice usually manages its own data exclusively and does not rely on other services or databases for its function (Gilbert 2021). This increases resilience.

Another design pattern that can be combined well with micro-services is the event-driven architecture, in which systems can communicate with each other separately via events. This pattern is particularly suitable if several systems coexist on an equal footing and entities are not managed centrally in one place, but are distributed across several systems. Changes to an entity in a system can be passed on to other systems via event-carried state transfer (Fowler 2017).

These architectures result in many independent services and components which can increase the complexity of deployment. Where in the past, perhaps a monolithic Web server and a connected database were used, now several services with their own databases as well as infrastructure components such as Event Queues are required. If individual components have to be scalable, the complexity is further increased by the deployment of additional nodes.

A solution for this is offered on the one hand by cloud services through the simple creation of new resources, features for the simple and automatic scaling of services, or approaches such as Infrastructure-as-Code (Morris 2020), in which required resources can be created programmatically and managed and versioned as part of the source code. This also simplifies the complete replication of an environment.

A further development is Kubernetes, which has established itself as the de facto standard for the deployment, scaling, and management of distributed resources (Bernstein 2014). It is supported by all common cloud providers but also enables deployment within an organization (on-premises) or mixed environments (hybrid cloud).

9.2.4 Law and Regulation

Laboratory medicine is highly regulated. This applies both to the laboratory infrastructure but also to devices, instruments, and consumables. ISO 13485 (ISO International Organization for Standardization 2016) sets extensive requirements for a quality management system for the manufacture of medical devices. ISO 15189 (ISO International Organization for Standardization 2012) concerns quality management in medical laboratories themselves. ISO 14971 (ISO International Organization for Standardization 2019) discusses terminologies, principles, and processes for risk management of medical devices. IEC 62366 (IEC International Electrotechnical Commission 2015) discusses requirements for the usability of medical devices. IEC 62304 (IEC International Electrotechnical Commission 2006) specifically addresses the development of medical software and software in medical devices. These standards are also taken up by regulatory authorities and in some cases extended. In the European Union, the IVDR (The European Parliament and the Council of the European Union 2017) regulates the development and use of devices and tools used for medical diagnosis. Software used for medical diagnoses is also covered by this regulation.

Concrete requirements for software include the use of quality and risk management systems, the use of the latest technological standards, the use of a software lifecycle process, the consideration of usability and security aspects, and verification and validation (The European Parliament and the Council of the European Union 2017; Grömminger 2018).

The lifecycle of software does not end with delivery but extends beyond. Post-Marketing Surveillance (PMS) is designed to ensure that manufacturers proactively collect experiences about their products that affect quality, performance, or risk. This collection must take place systematically so that Corrective and Preventive Actions (CAPA) can be initiated and implemented (The European Parliament and the Council of the European Union 2017).

In principle, the IVDR stipulates that only those diagnostics may be used in laboratories that are also approved for this purpose and fulfill the conditions of the IVDR (conformity declaration). Within limits, the IVDR allows laboratories to use their “own” test procedures, devices, and software which are not provided by commercial providers. The responsibility for the use of such Lab-Developed Tests (LDTs; synonym "In-house Tests", "Home-brew tests") (Spitzenberger et al. 2021) resides with the laboratory, which must demonstrate and document conformity with the essential requirements according to Annex I of the IVDR.

So-called research-use only products are often used as LDTs. These are products that are basically suitable for medical purposes, but for which the manufacturer has not undergone the necessary certifications and hence only offers them for research purposes. They have not undergone the rigorous validation process to demonstrate compliance with the basic requirements of scientific validity, technical performance, and clinical evidence required for in-vitro diagnostic products. Moreover, the requirements for products can differ significantly depending on whether they are to be used for research or for regular use. In research, for example, it may make sense to gear the user interface of a software tool to individual research projects. On the other hand, in regular deployment, such a project-related view could be rather a hindrance, and automation, integration and interoperability with other systems are more important.

There are also increasing efforts to formally regulate the use of ML. In 2020, the Joint Research Center (JRC) of the European Commission published a technical report describing how the trustworthiness and security of AI models can be increased (Hamon et al. 2020). The core criteria for this are transparency, reliability, and data protection. Specifically, the need for explainable AI is discussed. Since the risk of erroneous predictions by AI in the diagnosis of patients can be particularly serious, particularly strict requirements must be applied to the use of such techniques.

9.2.5 Analysis of an Example Laboratory

The actual situation in medical laboratories is recorded here as an example in a preliminary study. This preliminary study includes the documentation of use cases and processes, as well as the determination of requirements for future technical solutions. For this purpose, the use of gene expression analysis in a small laboratory in Heidelberg Biotechnology Park, Germany (operated by ImmBioMed GmbH & Co. KG) was investigated. The detailed methodology of the preliminary study was discussed in (Krause et al. 2022b). It is based on the framework of Nunamaker et al. (1990) and includes transcribed interviews, guided visits, use-case modeling, market analyses, and cognitive walkthroughs to validate the results.

The primary use cases in the laboratory are related to the development of new tests and test procedures, the implementation of tests, and post-market surveillance for ongoing review of tests offered (see Fig. 9.5). In the further course of this preliminary study, it focused on the running of gene expression tests and in particular on the measurement of so-called cytokine-dependent genes, which are important for the diagnosis of inflammatory or antiviral reactions (Barrat et al. 2019). For this purpose, the run of a sample was followed from initial

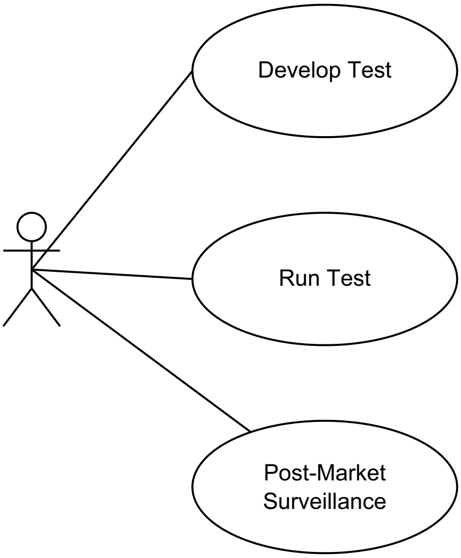


Fig. 9.5 Key laboratory use cases

registration to conversion into a lab report. Figure 9.6 shows the breakdown of this primary use case into more detailed use cases. These were assigned to four user stereotypes, “Lab Biologist”, “Data Analyst”, “QM & Compliance Officer”, and “Clinical Pathologist”.

Based on the observed processes and the results of the interviews, these use cases were then evaluated to determine whether they could be optimized or automated from an IT perspective. As expected, this evaluation differed depending on the use case. For example, it was observed that data has to be transferred between different systems at several points in the laboratory process and that this transfer sometimes must be performed manually due to different data formats or missing import/export interfaces. This, in turn, leads to additional effort to avoid or detect errors during the transfer (e.g. 4-eyes principle). Such time-consuming processes at system boundaries are examples of promising targets for optimization and automation some of which could be implemented with ML. An overview of the use cases with the evaluation of their automation potential (low/medium/high) is given in Table 9.1.

The preliminary study also confirmed the finding that medical laboratories are in a process of ongoing optimization of their test offerings. This optimization is partly driven by the market and partly by more stringent regulatory requirements. These advances are also changing the requirements for software. However, a market analysis conducted as part of the preliminary study found that these requirements are inadequately reflected by existing software, and even commercial systems have received few updates in recent years. In addition, the systems do not meet the regulatory requirements and hence can only be used for

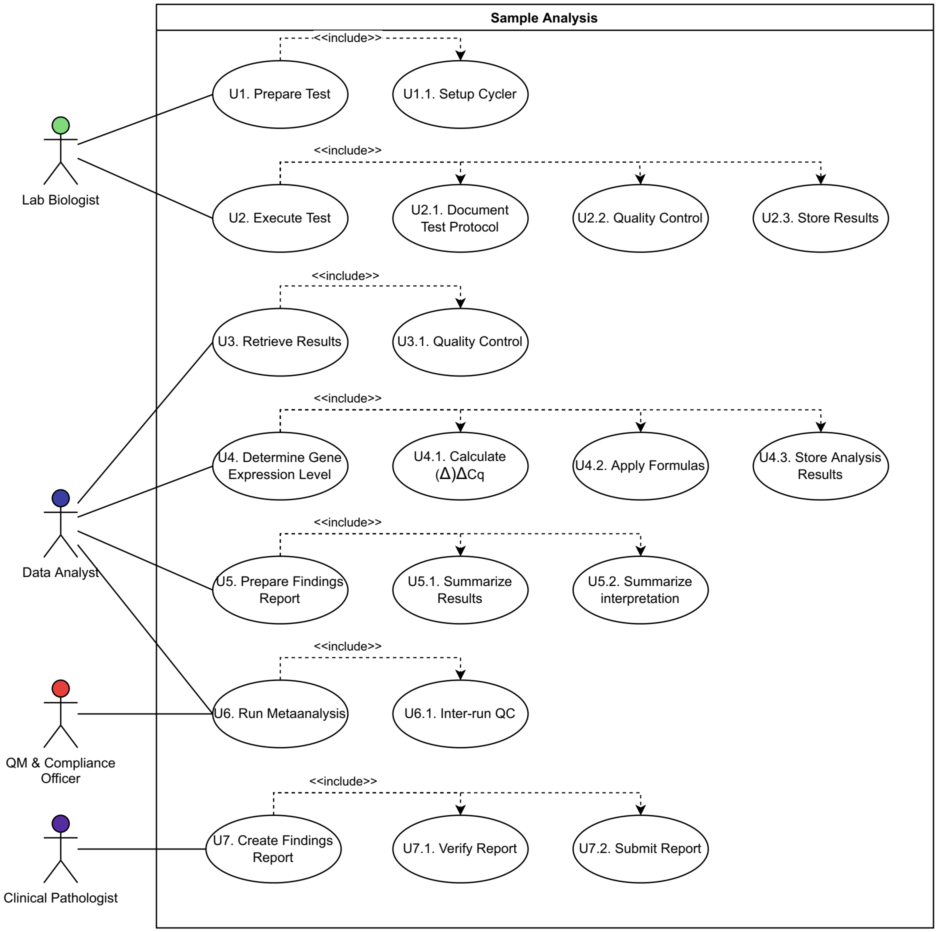


Fig. 9.6 “Run test” use case in laboratory—breakdown into detailed use cases

diagnostics with restrictions. Table 9.2 shows an overview of different software tools that were evaluated in the market study. The percent feature coverage in Table 9.2 is based on 13 essential features for gene expression analysis. A more detailed version of Table 9.2, listing each feature is given in (Krause et al. 2022b). The analysis is based on and extends the analysis done by Pabinger et al. (2014).

In order to evaluate possible analyses and user interfaces for future developments, a prototype called “PlateFlow” was also developed. This demonstrated the loading of raw data, the execution of elementary analyses, and the output of the analysis results in a report (see Fig. 9.7). PlateFlow was then evaluated with a cognitive walkthrough.

Table 9.1 Estimated Potential for Automatization of Use Cases. Previously published in Krause et al. (2022a)

Use Case	Potential			Limitations
	Low	Med.	High	
U1. Prepare Test		**x**		
U1.1. Program Cycler		x		Cycler Capabilities
U2. Execute Test		**x**	**x**	
U2.1. Document Test Protocol		x		User Input
U2.2. Quality Control			x	
U2.3. Store Results			x	
U3. Retrieve Results			**x**	
U3.1. Quality Control			x	
U4. Determine Gene Expression Level			**x**	
U4.1. Calculate (Δ)ΔCq			x	
U4.2. Apply Formulas			x	
U4.3. Store Analysis Results			x	
U5. Prepare Findings Report		**x**	**x**	
U5.1. Summarize Results			x	
U5.2. Summarize interpretation		x		Plausibility Checks
U6. Run Metaanalysis		**x**		
U6.1. Inter-run QC		x		Not Formalized
U7. Create Findings Report	**x**		**x**	
U7.1. Verify Report	x			Legal Responsibility
U7.2. Submit Report			x	

9.2.6 Automatization

There are many approaches to automating processing in bioinformatics with the aid of pipelines. A distinction must be made between approaches, (i) that attempt to cover as many use cases as possible with freely configurable pipelines, and (ii) those that map the most common use cases with the aid of fixed or less configurable pipelines. For example, the Galaxy Project (Afgan et al. 2018), a web-based platform for data analysis, falls into the first set of approaches. It offers several thousand tools from different scientific disciplines that can be used and connected in specially defined pipelines. Among them are many tools for genomic applications and for the use of ML models. The project supports the distribution of processes across many compute nodes. Deployment models range from local installation, to pay-as-you-go cloud services, to free public servers.

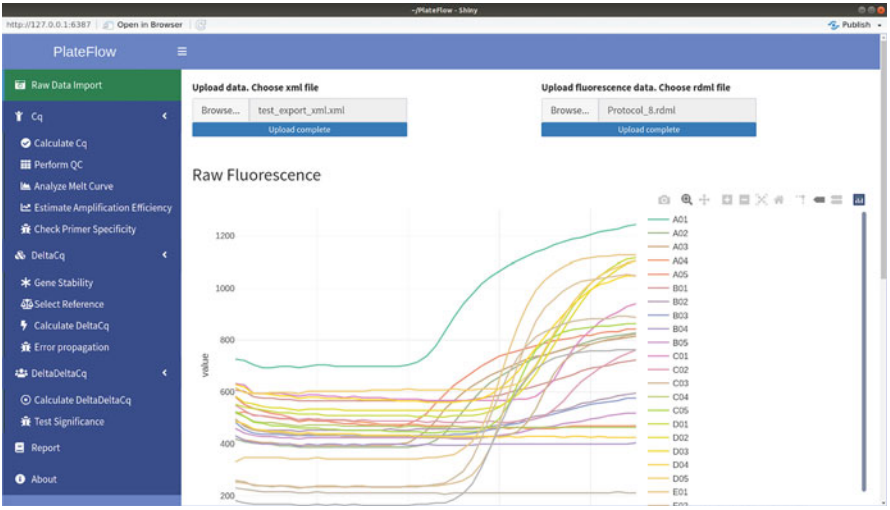
On the other hand, in principle, software with fixed or low configurable pipelines is application-specific. An example in the field of metagenomics is, e.g., MG-RAST (Meyer

Table 9.2 qPCR Software Evaluation. Summarized from Krause et al. (2022b); Pabinger et al. (2014)

Tool	Feature Coverage (%)	Last Update
CAmpER	38	2009
Cy0 Method	15	2010
DART-PCR	38	2002
Deconvolution	15	2010
ExpressionSuite Software	62	2019
Factor-qPCR	15	2020
GenEx	77	2019
geNorm	8	2018
LinRegPCR	46	2021
LRE Analysis	15	2012
LRE Analyzer	23	2014
MAKERGAUL	23	2013
PCR-Miner	23	2011
PIPE-T	54	2019
pyQPCR	54	2012
Q-Gene	31	2002
qBase	69	2007
qbase+	77	2017
qCalculator	38	2004
QPCR	69	2013
qPCR-DAMS	38	2006
RealTime StatMiner	69	2014
REST	46	2009
SARS	31	2011
SoFAR	31	2003

et al. 2008), which combines a database for metagenome sequences with automatic processing. The configuration options of the processing pipeline are limited to some parameters that can be set before starting the analysis. ML does not play an important role in this fixed pipeline.

The pipelines mentioned above provide a sound basis for biomedical analyses, but as stand-alone applications, they are more suitable for individual research projects rather than for ongoing diagnostics in medical laboratories. Reasons for this are the difficult integration into, and interoperability with, existing laboratory software, complex user interfaces, and legal requirements.



Upload data. Choose xml file

Browse... test_export_xml.xml

Upload complete

Upload fluorescence data. Choose rdml file

Browse... Protocol_8.rdml

Upload complete

Fig. 9.7 "PlateFlow" prototype user interface

9.2.7 Summary and Remaining Challenges

There are standardized process models for data analysis, which are also basically suitable for the requirements of laboratory diagnostics and particularly genomic applications. There are also generic technical architectures that support these approaches and enable scaling up. The use of sophisticated techniques such as deep learning is also possible in principle. However, in contrast, the analysis of the existing software landscape shows rather isolated solutions that only insufficiently cover the requirements of laboratories. Hence, the analysis of genomic data in laboratory diagnostics using modern methods such as AI remains a challenge.

9.3 GenDAI Technical Architecture

Since there is currently no solution on the market that satisfactorily meets the requirements of medical laboratories in genomic applications, we propose here a specialized technical architecture on which basis such a solution can be developed.

Conceptually, the architecture outlined in Fig. 9.8 is based on the GenDAI conceptual model discussed in Sect. 9.2. The technical basis is a modular frontend based on web technologies and a set of services.

The user interface is composed of pages and elements provided by individual modules. Thus the user interface can be easily extended or customized by incorporating new modules. On the other hand, modules can of course be hidden or removed if the functionalities they

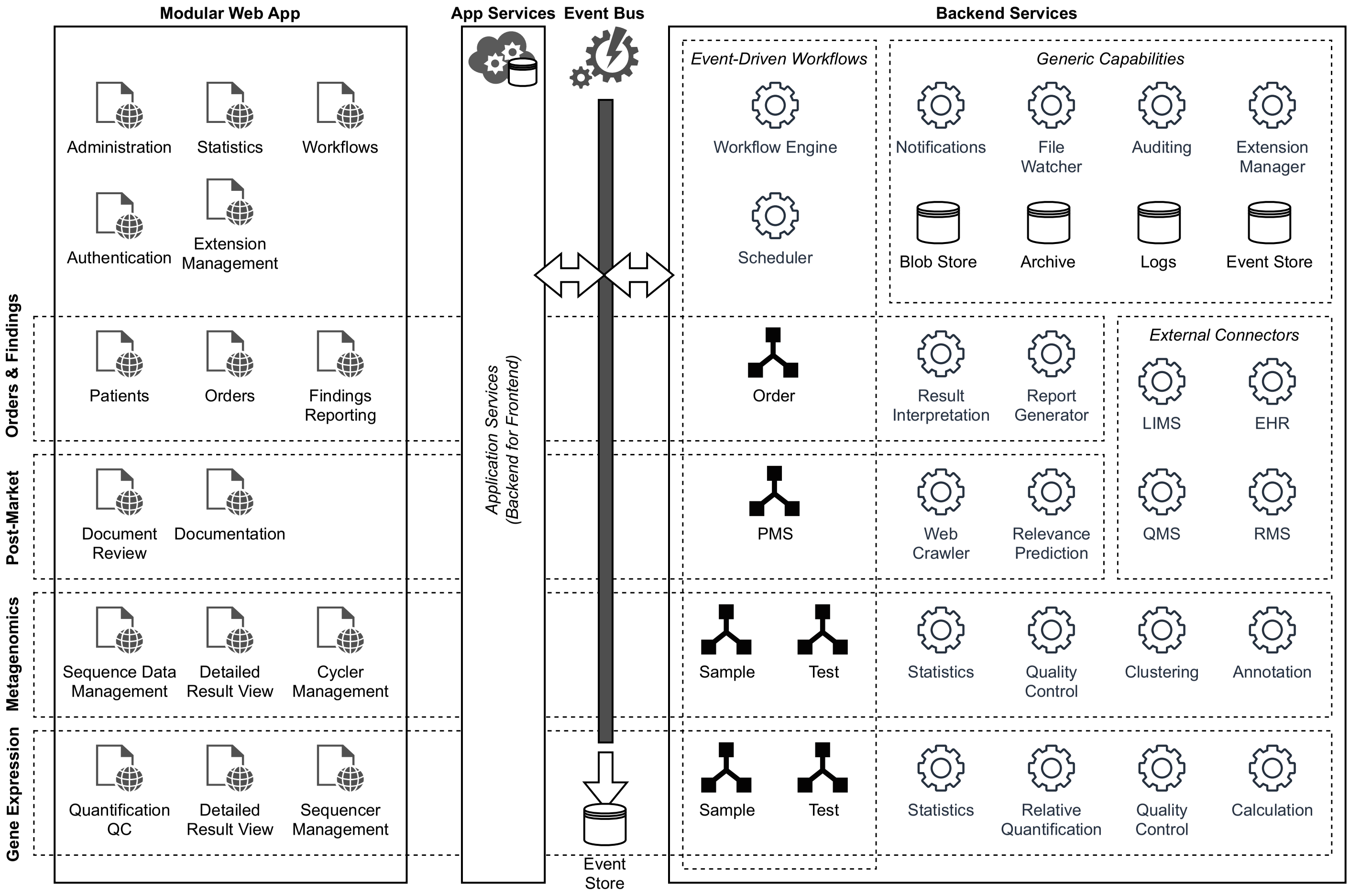


Fig. 9.8 GenDAI technical architecture

contain are provided by other applications and these cannot or should not be integrated into the GendDAI user interface. Basic modules allow, e.g., the registration of new orders, the management of tests, or the support of post-market surveillance.

The service layer consists of a micro-service architecture in which all relevant laboratory processes are mapped. The various services can be deployed independently of each other and are thus open to all technologies. Thus, the services can be either locally deployed or cloud-hosted. This also facilitates the scaling of individual services.

We distinguish between two types of services. So-called "application services" serve as backend-for-frontend (Gilbert 2021) for the support of user interfaces. For this purpose, they provide a synchronous REST-based API and a data model optimized for the frontend. As microservices, they are themselves responsible for managing the necessary data and making it available on request. To keep their data up to date, these services listen for relevant events on the event bus. At the same time, the services also generate new events based on user actions. "Backend services", on the other hand, operate largely asynchronously. They can be triggered by events on the event bus and often generate their own events. Examples of such services include work processes for analysis. Because these services are strongly oriented toward the associated user interface, they are not listed in detail in Fig. 9.8.

The concrete services presented in Fig. 9.8 are to be regarded as examples since depending on the circumstances in the laboratory, additional services may be added or certain services may not be needed. The examples were created using the event-storming method (Gilbert 2021). Thus, all relevant processes of a laboratory analysis were examined for their basic events (e.g., "Employee Authenticated", "Sample Registered", or "Results Computed"). Subsequently, these events were then assigned to subsystems and ultimately to specific services that are expected to produce or consume these events. The assignment is based on various factors such as the actors involved, the business capabilities, or the life cycle of the data.

9.3.1 Events and States

The use of events as the primary communication mechanism between services follows the principle of event-driven architecture with explicit modeling of state machine workflows. Events can cause state changes in entities that are mapped as state machine workflows. For example, an entity "Sample" can be in a state "Registered" after initial capture. When the sample is examined, an event is generated manually or automatically in the system that transfers the sample to the subsequent state "Processing".

This architecture has several advantages. For example, the state of an entity can be restored at any time by reproducing all relevant events (so-called event sourcing) (Gilbert 2021). Events can be recorded in an audit log and are thus available for audits, for documentation, or also for debugging errors.

Another advantage is the decoupling of components from each other. Individual components can be developed or exchanged separately and only have to be enabled to react to events. By the use of adapters, it is possible to connect existing components.

Figure 9.8 shows some examples of these workflows in the “Event-driven Workflows” area. The “Order” Workflow follows the state of a single order from a physician or other lab customer for a single patient starting from “Registration”. One order can consist of multiple samples for which different tests are required. The “Sample” Workflow follows a single sample as it moves through the lab. In a single test run in the laboratory, several samples are usually examined at the same time, so it makes sense to have a separate workflow for this. After a test has been performed, the results are assigned to the individual samples again, and finally, the combined results for a patient are assigned to the original order.

As a more thorough example we can look at the use case for PMS (see Sect. 9.2.4) as an important requirement for the use of in-vitro diagnostics. PMS requires the ongoing consideration of the current state of science, e.g. through regular keyword-related searches in the technical literature.

As this research should be structured and comprehensible, it is advisable to use information technology systems that support and document the PMS. The GenDAI architecture can represent these use cases through state-driven workflows. Such a workflow could look like that shown in Fig. 9.9.

First, a web crawler writes an event to the event bus when a new document with matching keywords or other features is found. This event leads to the creation of a PMS workflow for the document that tracks the current state. By assigning a reviewer, the document can be moved to the next “In Review” state. After the reviewer has assessed the relevance of the document, they can either discard the document or document the next course of action.

9.3.2 Application Specific Subsystems

The open, event-based architecture enables the use of different genomic applications in the same platform. The basis for this is the connection of new services to the event bus and, if necessary, the extension of the user interface with new modules. We can refer to the sum of all services and modules required for a new genomic application as a subsystem. In Fig. 9.8, the two subsystems “Metagenomics” and “Gene Expression” are shown as examples.

For example, for metagenomics, a user interface can be used to upload existing sequence data or transfer it from a partner laboratory. If sequencing is performed on-site, the sequencer must be programmed for this purpose. The actual processing in the backend is completed with several process steps, which are executed on one or more compute nodes, one after the other. These usually include “quality control”, “clustering”, and “annotation” (Krause et al. 2021b). In gene expression analysis, the processing is simpler. In addition to “Quality Control” and actual gene expression quantification from the raw data (“Relative Quantification”), the results can be summarized by calculation formulas (“Calculation”).

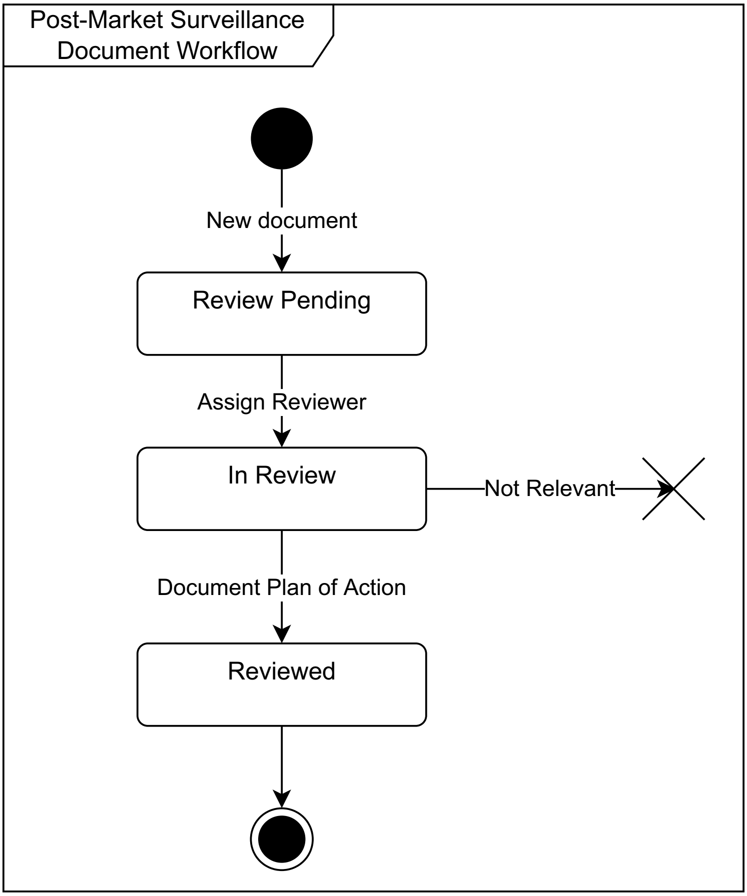


Fig. 9.9 Post-market surveillance document workflow

In summary, the GenDAI technical architecture is based on modern and proven methods for highly scalable solutions. Hence, we believe that GenDAI can also meet the technical requirements of genomics in laboratory diagnostics. Moreover, the improved capabilities of building and deploying such distributed architectures using Cloud, Infrastructure-as-a-Service, or Kubernetes in recent years leave these architectures easier to deploy than previously.

9.4 Conclusion & Future Work

Genomic applications in laboratory diagnostics have the potential to further advance the era of personalized medicine. AI, and particularly ML techniques such as deep learning, can provide support to efficiently analyze data and facilitate relevant results.

There are several technical and regulatory challenges in order to achieve this goal that must be addressed. The "GenDAI Technical Architecture" presented here is based on the "GenDAI Conceptual Model" and makes specific technical proposals to address the complex requirements of laboratory software for genomic applications.

Remaining challenges include the exact technical design of software modules, the implementation of a complete solution, evaluation, and the legally compliant use in practical applications. As a next step, we plan a prototypical implementation of the GenDAI technical architecture for a particular use case.

References

Abawajy, J. 2015. Comprehensive analysis of big data variety landscape. *International Journal of Parallel, Emergent and Distributed Systems* 30 (1): 5–14.

Adams, G. 2020. A beginner's guide to rt-pcr, qpcr and rt-qpcr. *The Biochemist* 42 (3): 48–53.

Afgan, E., D. Baker, B. Batut, M. van den Beek, D. Bouvier, M. Čech, J. Chilton, D. Clements, N. Coraor, B. A. Grüning, A. Guerler, J. Hillman-Jackson, S. Hiltemann, V. Jalili, H. Rasche, N. Soranzo, J. Goecks, J. Taylor, A. Nekrutenko, and D. Blankenberg. 2018. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research* 46 (W1): W537–W544.

Armour, C. R., S. Nayfach, K. S. Pollard, and T. J. Sharpton. 2019. A metagenomic meta-analysis reveals functional signatures of health and disease in the human gut microbiome. *mSystems* 4 (4): 1–15.

Barrat, F. J., M. K. Crow, and L. B. Ivashkiv. 2019. Interferon target-gene expression and epigenomic signatures in health and disease. *Nature Immunology* 20 (12): 1574–1583.

Berding, K., K. Vlckova, W. Marx, H. Schellekens, C. Stanton, G. Clarke, F. Jacka, T. G. Dinan, and J. F. Cryan. 2021. Diet and the microbiota-gut-brain axis: Sowing the seeds of good mental health. *Advances in Nutrition* (Bethesda, Md) 12 (4): 1239–1285, https://pubmed.ncbi.nlm.nih.gov/33693453/.

Bernstein, D. 2014. Containers and cloud: From lxc to docker to kubernetes. *IEEE Cloud Computing* 1 (3): 81–84.

Berwind, K., M. X. Bornschlegl, M. A. Kaufmann, and M. Hemmje. 2016. Towards a cross industry standard process to support big data applications in virtual research environments. In *Proceedings of the Collaborative European Research Conference (CERC) 2016*, ed. U. Bleimann, B. Humm, R. Loew, I. Stengel, and P. Walsh. https://www.cerc-conf.eu/wp-content/uploads/2018/06/CERC-2016-proceedings.pdf.

Chapman, P., J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. 2000. *Crisp-dm 1.0: Step-by-step data mining guide. SPSS Inc., U.S.A.*

Chiu, C. Y., and S. A. Miller. 2019. Clinical metagenomics. *Nature Reviews Genetics* 20 (6): 341–355, https://www.nature.com/articles/s41576-019-0113-7.

Coudray, N., P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, and A. Tsirigos. 2018. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine* 24 (10): 1559–1567, https://pubmed.ncbi.nlm.nih.gov/30224757/.

Daniel, N., E. Lécuyer, and B. Chassaing. 2021. Host/microbiota interactions in health and diseases-time for mucosal microbiology! *Mucosal Immunology* 14 (5): 1006–1016, https://www.nature.com/articles/s41385-021-00383-w.

Dias, R., and A. Torkamani. 2019. Artificial intelligence in clinical and genomic diagnostics. *Genome medicine* 11 (1): 70.

Fan, Y., and O. Pedersen. 2021. Gut microbiota in human metabolic health and disease. *Nature reviews Microbiology* 19 (1): 55–71, https://pubmed.ncbi.nlm.nih.gov/32887946/.

Fioravanti, D., Y. Giarratano, V. Maggio, C. Agostinelli, M. Chierici, G. Jurman, and C. Furlanello. 2018. Phylogenetic convolutional neural networks in metagenomics. *BMC bioinformatics* 19 (Suppl 2): 49, https://link.springer.com/article/10.1186/s12859-018-2033-5.

Fowler, M. 2017. *What do you mean by „event-driven"?*. https://martinfowler.com/articles/201701-event-driven.html, 2022-04-19.

Gebrayel, P., Nicco, C., S. Al Khodor, J. Bilinski, E. Caselli, E. M. Comelli, M. Egert, C. Giaroni, T. M. Karpinski, I. Loniewski, A. Mulak, J. Reygner, P. Samczuk, M. Serino, M. Sikora, A. Terranegra, M. Ufnal, R. Villeger, C. Pichon, P. Konturek, and M. Edeas. 2022. Microbiota medicine: Towards clinical revolution. *Journal of Translational Medicine* 20 (1): 111, https://translational-medicine.biomedcentral.com/articles/10.1186/s12967-022-03296-9.

Gilbert, J. 2021. *Software architecture patterns for serverless systems: Architecting for innovation with events, autonomous services, and micro frontends*, 1st ed. Birmingham: Packt Publishing Limited.

Goetz, L. H., and N. J. Schork. 2018. Personalized medicine: Motivation, challenges, and progress. *Fertility and Sterility* 109 (6): 952–963.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep learning*. Cambridge: MIT Press. http://www.deeplearningbook.org/.

Grömminger, S. 2018. *Ivdr – in-vitro-diagnostic device regulation*. https://www.johner-institute.com/articles/regulatory-affairs/ivd-regulation-ivdr/, 2021-08-29.

Gurovich, Y., Y. Hanani, O. Bar, G. Nadav, N. Fleischer, D. Gelbman, L. Basel-Salmon, P. M. Krawitz, S. B. Kamphausen, M. Zenker, L. M. Bird, and K. W. Gripp. 2019. Identifying facial phenotypes of genetic disorders using deep learning. *Nature Medicine* 25 (1): 60–64, https://www.nature.com/articles/s41591-018-0279-0.

Hamon, R., H. Junklewitz, and I. Sanchez. 2020. *Robustness and explainability of Artificial Intelligence: From technical to policy solutions, EUR*, vol. 30040. Luxembourg: Publications Office of the European Union.

Hong, M., S. Tao, L. Zhang, L. T. Diao, X. Huang, S. Huang, S. J. Xie, Z. D. Xiao, and H. Zhang. 2020. Rna sequencing: New technologies and applications in cancer research. *Journal of Hematology & Oncology* 13 (1): 166. https://jhoonline.biomedcentral.com/articles/10.1186/s13045-020-01005-x.

IEC International Electrotechnical Commission. 2006. Medical device software—software life cycle processes. *IEC* 62304: 2006.

IEC International Electrotechnical Commission. 2015. Medical devices—part 1: Application of usability engineering to medical devices. *IEC* 62366–1: 2015.

ISO International Organization for Standardization. 2012. Medical laboratories—requirements for quality and competence. *ISO* 15189: 2012.

ISO International Organization for Standardization. 2016. Medical devices—quality management systems—requirements for regulatory purposes. *ISO* 13485: 2016.

ISO International Organization for Standardization. 2019. Medical devices—application of risk management to medical devices. *ISO* 14971: 2019.

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.

W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. 2021. Highly accurate protein structure prediction with alphafold. *Nature* 596 (7873): 583–589.

Krause, T., E. Jolkver, S. Bruchhaus, M. Kramer, and M. Hemmje. 2021a. Gendai—AI-assisted laboratory diagnostics for genomic applications. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, ed. IEEE Computer Society.

Krause, T., J. T. Wassan, P. Mc Kevitt, H. Wang, H. Zheng, and M. Hemmje. 2021. Analyzing large microbiome datasets using machine learning and big data. *BioMedInformatics* 1 (3): 138–165.

Krause, T., E. Jolkver, S. Bruchhaus, P. Mc Kevitt, M. Kramer, and M. Hemmje. 2022. A preliminary evaluation of "gendai", an ai-assisted laboratory diagnostics solution for genomic applications. *BioMedInformatics* 2 (2): 332–344.

Krause, T., E. Jolkver, P. Mc Kevitt, M. Kramer, and M. Hemmje. 2022. A systematic approach to diagnostic laboratory software requirements analysis. *Bioengineering* 9 (4): 144.

Kuo, W. P., E. Y. Kim, J. Trimarchi, T. K. Jenssen, S. A. Vinterbo, and L. Ohno-Machado. 2004. A primer on gene expression and microarrays for machine learning researchers. *Journal of biomedical informatics* 37 (4): 293–303.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. *Nature* 521 (7553): 436–444, https://www.nature.com/articles/nature14539.pdf.

Liu, Y. X., Y. Qin, T. Chen, M. Lu, X. Qian, X. Guo, and Y. Bai. 2021. A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein & Cell* 12 (5): 315–330.

Marchesi, J. R., J. Ravel. 2015. The vocabulary of microbiome research: A proposal. *Microbiome* 3 (1): 31. https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-015-0094-5.

Mc Kevitt, P., and D. Partridge. 1991. Problem description and hypotheses testing in artificial intelligence. In *AI and cognitive science'90, Workshops in Computing*, ed. M. McTear and N. Creaney, 26–47. London: Springer.

Meyer, F., D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards. 2008. The metagenomics rast server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9 (1): 1–8.

Morris, K. 2020. *Infrastructure as code: Dynamic systems for the cloud age*, 2nd ed. Beijing: O'Reilly.

National Center for Biotechnology Information. 2022. *Genbank and wgs statistics*. https://www.ncbi.nlm.nih.gov/genbank/statistics/, 15.04.2022.

Nunamaker, J. F., M. Chen, and T. D. Purdin. 1990. Systems development in information systems research. *Journal of Management Information Systems* 7 (3): 89–106.

Ondov, B. D., N. H. Bergman, and A. M. Phillippy. 2011. Interactive metagenomic visualization in a web browser. *BMC bioinformatics* 12: 385.

Pabinger, S., S. Rödiger, A. Kriegner, K. Vierlinger, and A. Weinhäusel. 2014. A survey of tools for the analysis of quantitative pcr (qpcr) data. *Biomolecular Detection and Quantification* 1 (1): 23–33.

Qin, J., R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J. M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Dore, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, P. Bork, S. D. Ehrlich, and J. Wang. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464 (7285): 59–65.

Reis, T., M. X. Bornschlegl, and M. Hemmje. 2021. Ai2vis4bigdata: A reference model for ai-based big data analysis and visualization. In *Advanced visual interfaces*, ed. T. Reis, M. X. Bornschlegl, M. Angelini, and M. Hemmje. Lecture Notes in Computer Science, 1–18. Springer Nature, Switzerland.

Shafique, U., and H. Qaiser. 2014. A comparative study of data mining process models (kdd, crisp-dm and semma). *International Journal of Innovation and Scientific Research* 12 (1): 217–222.

Soueidan, H., and M. Nikolski. 2015. Machine learning for metagenomics: Methods and tools. https://arxiv.org/pdf/1510.06621.

Spitzenberger, F., J. Patel, I. Gebuhr, K. Kruttwig, A. Safi, and C. Meisel. 2021. Laboratory-developed tests: Design of a regulatory strategy in compliance with the international state-of-the-art and the regulation (eu) 2017/746 (eu ivdr in vitro diagnostic medical device regulation). *Therapeutic innovation & regulatory science.* 56 (2022): 47–64.

Stephens, Z. D., S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson. 2015. Big data: Astronomical or genomical? *PLoS biology* 13 (7): e1002195.

Suwinski, P., C. Ong, M. H. T. Ling, Y. M. Poh, A. M. Khan, and H. S. Ong. 2019. Advancing personalized medicine through the application of whole exome sequencing and big data analytics. *Frontiers in Genetics* 10: 49.

The European Parliament and the Council of the European Union. 2017. In vitro diagnostic regulation: Ivdr. http://data.europa.eu/eli/reg/2017/746/2017-05-05.

Wassan, J. T., H. Wang, F. Browne, and H. Zheng. 2019. Phy-pmrfi: Phylogeny-aware prediction of metagenomic functions using random forest feature importance. *IEEE transactions on nanobioscience* 18 (3): 273–282.

Zhang, X., L. Li, J. Butcher, A. Stintzi, and D. Figeys. 2019. Advancing functional and translational microbiome research using meta-omics approaches. *Microbiome* 7 (1): 154. https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-019-0767-6.

Thomas Krause (M.Sc.) is a researcher in the areas and intersections of artificial intelligence, cloud, Big Data, genomics, and laboratory diagnostics. He draws on more than 15 years of industry experience in highly scalable solutions developed as a consultant for well-known international clients. He holds multiple certifications from industry leaders, such as Microsoft, for machine learning, cloud architecture, and other areas. He is also pursuing a Ph.D. at the Fernuniversität Hagen, Germany since 2019.

Elena Jolkver studied biology (diploma, 2005) and biochemistry (Ph.D., 2009) at the University of Cologne, Germany. Switching the lab coat for a laptop, she proceeded to research as a data scientist at at BASF metanomics GmbH, identifying biomarkers for plant yield. She continued her education in machine learning and computer science (B.Sc., Fernuniversität Hagen, Germany, 2022) while passing on her skills and knowledge in modern data science approaches as a consultant at xValue GmbH and a Guest Lecturer at DHBW Karlsruhe, Germany. Her major focus is on supporting third-party biopharmaceutical companies in the area of biomarker research and software development.

Michael Kramer (M.D.) studied medicine at the University of Heidelberg, Germany. He is a board-certified specialist in laboratory medicine. From 1990 to 1999 he was head of the section for Immunopathology at the University Hospital Heidelberg. Since 1999 he is an adjunct professor of Applied Immunology at the University of Heidelberg. He runs the consulting company ImmBioMed GmbH & Co. KG with a focus on the translational development of laboratory diagnostic procedures and advice on the associated medical-scientific and regulatory issues. To support these activities, ImmBioMed GmbH & Co. KG operates a laboratory infrastructure in the Biotechnology Park of the University of Heidelberg.

Paul McKevitt completed his Ph.D. in Computer Science at the University of Exeter, England (1991). He also completed a Master's degree in Education (M.Ed.), at the University of Sheffield, England (1999). He has 40 years of international experience in Computer Science (Artificial Intelligence [AI], Intelligent MultiMedia/MultiModal Systems, Medical Informatics) focused on teaching, research & technology transfer (with software demonstrators and patents) as an academic deployed at university and R&D institutions in Germany, Northern Ireland, France, Denmark, England, Ireland & the USA. He is currently Professor Emeritus at Ulster University, Magee, Derry/LondonDerry, Northern Ireland and Visiting Professor of AI at FTK—Research Institute for Telecommunication and Cooperation e. V. in Pfungstadt, Germany and at the Academy for International Science & Research (AISR), also in Derry.

Matthias L. Hemmje received a Ph.D. degree from the Department of Computer Science of the Technical University of Darmstadt, Germany. After that, he managed a research division at Fraunhofer IPSI in Darmstadt, Germany. Since 2004 he is a Full Professor of Computer Science at the University of Hagen, Germany, where he holds the Chair of Multimedia and Internet Applications. His primary research interests include Information Systems, Knowledge management, Semantic Technologies, Big Data Analysis, Information Visualization, and Long Term Archival. Since 2009, Matthias Hemmje is director and chairman of the board of FTK—Research Institute for Telecommunication and Cooperation e. V. in Dortmund, Germany. Having worked in many international R&D projects with research and industrial partners, his R&D and innovation teams ensure the transfer of results into widely available prototypes, products, and services.