

Trecode: a FAIR eco-system for the analysis and archiving of omics data in a combined diagnostic and research setting

Hindrik HD Kerstens^{1,*}, Jayne Y Hehir-Kwa^{1,*}, Ellen van de Geer¹, Chris van Run¹, Eugène TP Verwiël¹, Douwe van der Leest¹, Bastiaan BJ Tops¹, Patrick Kemmeren^{1,#}

¹ Princess Máxima Center for Pediatric Oncology, Utrecht, The Netherlands

* contributed equally

correspondence: p.kemmeren@prinsesmaximacentrum.nl

Keywords: Next-Generation Sequencing, FAIR, Data management platform, Data analyses, Data provenance, Automation

Abstract

Motivation

The increase in speed, reliability and cost-effectiveness of high-throughput sequencing has led to the widespread clinical application of genome (WGS), exome (WXS) and transcriptome analysis. WXS and RNA sequencing is now being implemented as standard of care for patients and for patients included in clinical studies. To keep track of sample relationships and analyses, a platform is needed that can unify metadata for diverse sequencing strategies with sample metadata whilst supporting automated and reproducible analyses. In essence ensuring that analysis is conducted consistently, and data is Findable, Accessible, Interoperable and Reusable (FAIR).

Results

We present “Trecode”, a framework that records both clinical and research sample (meta) data and manages computational genome analysis workflows executed for both settings. Thereby achieving tight integration between analyses results and sample metadata. With complete, consistent and FAIR (meta) data management in a single platform, stacked bioinformatic analyses are performed automatically and tracked by the database ensuring data provenance, reproducibility and reusability which is key in worldwide collaborative translational research.

Availability and implementation

The Trecode data model, codebooks, NGS workflows and client programs are currently being cleared from local compute infrastructure dependencies and will become publicly available in spring 2021.

Contact

p.kemmeren@prinsesmaximacentrum.nl

Introduction

The analysis of next-generation sequencing (NGS) data poses significant challenges for bioinformaticians. First, as NGS technologies continue to become cheaper, faster and more reliable, they are increasingly used both in research and diagnostics (1, 2). Current diagnostic utilization of NGS includes a wide range of applications, ranging from DNA variant discovery, to gene fusion detection by transcriptome assessment (RNA-Seq), to DNA methylation profiling. In addition, NGS technologies are also used within cancer research, for detailed genetic characterization of patient samples to study the molecular mechanisms that drive malignancies and to develop personalized treatment (3–5). These developments are only possible if bioinformatic and data services can accurately and efficiently process, store and distribute highly diverse NGS data (6). Automated data handling is required to match the increasing scale in which data is generated and to effectively use the expanding variety of compute- and data storage infrastructures. Furthermore, automation is essential to enable a relatively small number of bioinformaticians to process the data with minimal chance of human errors whilst using compute resources efficiently. If automated sufficiently, bioinformaticians can focus on method development and result interpretation rather than performing the analyses also with increasing workloads.

Secondly, (meta) data management is typically spread across separate clinical and/or research database(s), each using different data models. These systems are often loosely coupled to the genome analysis platform which tend to be monolithic workflows designed to support either clinical or research analysis. Scalable genomics analysis requires portable analysis workflows which can be implemented across a variety of software and hardware environments, including high performance computing clusters and cloud instances. Domain specific workflow languages (7, 8) and executors (9, 10) are being increasingly used to simplify analyses across a variety of execution environments. This progress facilitates a paradigm shift of moving analysis workflows to the data rather than the data to the analysis workflow. The latter is becoming increasingly difficult given the increasing volume of data. Furthermore, it fuels the development of standards for describing computational workflows with the aim of making genome analysis reproducible and transferrable to other labs (11).

The research of rare diseases and development of precision medicine requires combining data between institutes for large-scale analyses. Useful and sufficient metadata must be provided to enable findability, reusability and correct interpretation of the distributed data. Yet genomics is a dynamic field and capturing metadata for emerging analysis techniques is a moving target, placing demands on the flexibility of data models used in the genomics platform. Furthermore, the genomics platform should be accessible and interoperable to allow programmatic exchange of metadata thereby avoiding human error and increasing scalability.

The *Princess Máxima Center for Pediatric Oncology*, is the Dutch national children's oncology center and started sequencing research samples in 2016 and sequencing based diagnostics in 2018. For analyzing and tracking all sequencing data produced in the center, we have developed a platform called Trecode which is presented here. As of September 2020, this platform has archived the meta data and primary analysis of almost 1,000 sequenced transcriptomes and more than 500 whole exome and genome sequenced samples. Subsequent somatic variant analyses have been completed on 240 tumor-normal sample pairs for clinical (WXS) and 241 for research (WGS). With approximately 600 children per year being affected by cancer in the Netherlands this data collection will potentially grow by this number of sample pairs yearly. Trecode is a generic data management and analysis platform which allows a small team of 4 to 6 bioinformaticians to support all sequencing analysis related activities in both routine clinical diagnostics and research biobanking. Trecode integrates sequencing experiment description and computational analyses into a single data model, facilitating sample relation tracking and automated genome analyses. Furthermore, show that this integration promotes data exchange between users from different disciplines. Molgenis scientific data platform (12) is used to make this data model interoperable which aids in removing barriers in Findability, Accessibility, Interoperability and Reusability (FAIR) of genomics data.

The metadata standards enforced by the Trecode platform facilitates automated submissions to the European Genome-phenome Archive (EGA) (13) public repository. The computational analyses are fully integrated and designed with the aim of maximizing code re-use, reproducibility and transferability. The Trecode platform assists operators and bioinformaticians in performing reproducible data analyses and the generic, flexible and scalable design allows developers to adjust the platform with little effort.

System and methods

Platform design

We designed Trecode to enable the systematic collection, annotation and genomic characterization of samples for both tumor and healthy tissue of all patients entering a research hospital. RNA sequencing and WXS is performed systematically on all diagnostic samples. When informed consent is given, these samples additionally become part of the biobank and WGS is performed. Trecode captures all required information, models relations between samples, their derivatives and NGS results in a single data model and performs genomics analyses semi-automatically (Figure 1). Metadata and NGS analyses results remain findable and are kept accessible via the platform's standardized interfaces. NGS result files are stored using community-developed standard formats in Trecode and are annotated using well-established ontologies ensuring optimal reusability according to the FAIR principles (14). Based on pre-existing open software components we aim for maximum reuse of current developments.

Samples collected from an individual are identified as (primary) biosources from which new biosources can be derived by means of cell or organoid culturing. Extracts isolated from a biosource such as DNA or RNA, are termed biomaterials. These are stored as a separate entity and can be traced back to the originating biosource and individual using the data model. Within Trecode, biomaterials are the primary inputs for generating sequencing libraries. A sequencing experiment, such as whole genome (WGS), whole exome (WXS) or transcriptome (RNA-Seq), requires a context specific sequencing library and results in a sequencing run with associated sequencing run data (Figure 1, panel A).

An operator initiates the data analyses by querying the Trecode instance for the existence of sequencing data based on i) the sequencing strategy (WGS/WXS/RNA-Seq), ii) for a (set of) biomaterial(s) and iii) the type of analysis to be performed (Figure 1, panel C). As part of the query results, a predefined bioinformatics workflow for genome sequencing analysis is initiated using a set of sequencing strategy specific parameters. Parameters include required genomics and compute resources as well as workflow computational parameters that together form a complete list of workflow inputs. The workflow is then executed using an execution service which translates generic workflow task definitions into specific compute backend calls. By using an executor which supports multiple compute backend-services, analyses workflows become transferrable and can be executed on a variety of compute infrastructures. Workflows not only

include tasks for data analysis but also tasks for registering and linking performed analyses to biomaterial(s) in the database and archiving analysis results in a file or object store.

Implementation

Trecode is implemented using Molgenis, an open source scientific data management platform (12), and Cromwell Workflow Definition Language (WDL) executor (10). This creates a platform with a flexible and extensible data model which is tightly connected to an analysis workflow execution service using a REpresentational State Transfer (REST) protocol (15). We based our data model on SRA/EGA (13, 16) and extended with attributes from the Investigation/Study/Assay (ISA) open source metadata tracking framework (17). For describing the automated computational workflows, we created additional entities and attributes. The database is interoperable through Molgenis' web based graphical user interface (web GUI) and its REST based application programming interface (API) which offers accessibility to a broad range of end users.

Data model

In Trecode, samples and their relationships are recorded concisely together with the NGS sequencing experiments performed (Figure 2). Moreover, the genomics analyses created through workflows are interlinked with the primary 'raw' data and analysis result files. The core sequencing data tables are: 'study', 'individual', 'diagnosis', 'biosource', 'biomaterial', 'library', 'experiment', 'run' and 'analysis'. A 'study' can be nested, using 'sub-studies', and is related to 'individual' by the 'individual study' cross table (Supplemental figure 1). The data model enforces that both experiments and computational analyses are linked to studies.

An 'individual' can have one or more malignancies that each are recorded in Trecode as 'diagnosis' and is referenced from 'biosources' that represent samples from individuals, such as tumor and normal tissue/blood samples. An essential task of the 'diagnosis' entity is to not only describe the clinical diagnosis but also to link related biosources, such as tumor and normal samples, relapse tumor samples, or refractory cancer events from a single malignancy. Biomaterial is the biosource extract that is analyzed in NGS and is referenced from 'library' or 'assay' representing the sequencing library that is created from the extracted DNA/RNA or the assay that is performed e.g. for methylation analysis (Supplemental figure 2). The wet lab

sequencing or array-based analyses strategies in ‘library’/‘assay’ is further referenced from ‘experiment’ which holds details of the experimental setup. Entity ‘experiment’ is referenced from ‘run’, which holds links to the raw sequencing/methylation data, and from ‘analysis’ which describes the computational processing of the sequencing/methylation results using ‘Library’ or ‘Assay’ specific workflow configurations.

In addition, Trecode stores analyses workflows, their requirements and dependencies as well as analyses results created by each workflow execution. The series of computational processing steps performed by a specific bioinformatic workflow is described in the table ‘workflow’. Attribute values for ‘analysis_type_id’ and ‘library_strategy_id’ in table ‘workflowinstance’ determine for which experimental setups a generic bioinformatic workflow is used. The initiation of a generic bioinformatic workflow is configured in supporting workflow tables which refer to ‘workflowinstance’, namely ‘workflowsoftware’, ‘workflowparameter’, ‘workflowresource’ and ‘workflowcompute’. These tables select software (versions), the workflow and task level parameter values, genomics resources (references, genome annotations, variant database versions) to be used and the compute resources to be requested. As new versions of workflows are released, updates such as parameter changes remain traceable by tagging these new settings with a release version that is defined in the table ‘release_version’. This concept is implemented across all supporting workflow tables. Additionally, the specific code for a workflow is described in the table ‘workflowfile’. Compute environment differences are captured by parameters in Trecode tables ‘environment’ and ‘workflowexternal’. To promote consistent and uniform naming of workflow and task level inputs and support reuse, all input tags used in workflows and tasks are described in Trecode.

The computational analyses for (a set of) sequencing experiment(s) using a ‘workflowinstance’ is logged in the table ‘workflowrun’. This table links the workflow execution products (analysis results) which are described in the linked table ‘analysis’. All resulting files from a workflow run are registered in the table ‘data_file_store’. Computational analyses can be performed on biomaterials’ primary data via the linked experiments but can also be a secondary analysis that uses primary analyses results as input. In our primary analysis, the sequencing reads of an experiment are mapped to the reference genome and germline variants are detected. In a

secondary analysis, a pair of primary analysis mapping results are used as input and compared as in, for example, somatic copy number variant analysis.

Within Trecode, panels of normals are created using dedicated workflows and stored in the table ‘panelofnormals’ with a reference to the generic workflow analysis results table ‘analysis’. As a result, exactly how, and which samples are included in the panel of normal remains traceable. Panels of normals can subsequently be queried to be used for artifact reduction (18) purposes in e.g. copy number variant (CNV) and somatic single nucleotide variant (SNV) analyses.

Once the analysis is completed for a series of experiments, the data can be submitted to a sequencing repository to make it publicly accessible. The Trecode platform provides tools to perform automatic submissions to EGA by collecting raw sequencing data, analysis results and required metadata, given a number of experiment IDs (Supplemental figure 3A). The data model enforces researchers to collect sample metadata from the beginning of the project, is designed to perform submissions in a highly automated fashion and tracks what has been submitted (Supplemental figure 3B).

Accessibility to users from multiple disciplines

The Trecode platform links experimental and computational metadata as well as workflow execution code in a single data model. Consequently, the platform is useful to a broad range of users, in particular when users require exchange of information across research fields. Data analysis operators are provided with sufficient experiment (meta) data to perform computational analyses. Wet-lab scientists will find the computational analyses results linked to their experiments. Developers can exploit the data model flexibility and programmatic interfacing to efficiently add or improve functionality. By providing both a web GUI and REST API the Trecode platform keeps data and analyses accessible for all users in a broad variety of roles.

Operators

The platform provides genomics data, metadata, data analyses pipelines (workflows) and workflow management functionality to operators who routinely perform genome analyses. Operators typically insert metadata for new samples and subsequently execute workflows to analyze these samples. Newly inserted sample metadata is checked for consistency and vocabulary by the Trecode platform and implicitly linked to default analysis workflow instances.

Computational analysis queries are checked for sensibility by the platform command line client to avoid obvious human errors. For example, during somatic variant analysis, it is checked if the specified biomaterials originate from a single individual. In general, tumor versus normal comparisons are made in somatic analyses. Deviating queries like tumor versus (relapse) tumor result in a warning where the operator may explicitly indicate that this check should be ignored. Furthermore, queries are checked whether library strategy and sequencing platform correspond between the queried sequencing experiments and any panel of normals being used. Once the input is accepted, the workflow will be started and the platform provides the operator information about the workflow progress, consistent error reporting within a few mouse clicks and ensures that created analysis products become findable and linked to the sample metadata. Linked analysis products such as quality control (QC) metrics can be directly viewed, whereas other products can become subject of secondary analyses or can be used in reference panels, facilitating analysis plans of the operator.

Wet lab scientists

Trecode's web GUI with multiple search functions and wet lab intuitive data model allows lab technicians to easily find information about samples registered and processed, as well as associated metrics. Sample level QC metrics are a key resource for lab technicians to review and benchmark laboratory protocols. Furthermore, the platform offers browse and lookup functionality from individual to experiment level and the progress of computational analyses can be followed in real-time. More complex sample relations can also be traced via the web GUI and Molgenis' plugin architecture supports the creation of additional views on the data as desired.

Bioinformaticians

Bioinformaticians and developers interact with the Trecode platform when updating existing or implementing new workflows and registering genomics resources. This user group typically uses the REST API of the platform to make workflow and resource related changes programmatically. The REST API is also used by bioinformaticians who query experiment and analyses results tables to perform aggregate analyses based on a custom subset of primary analyses results.

Data security

Data is protected in Trecode using a password-based login, and role-based security is used to limit access or restricting operations to authorized users (see Table 1). For example: operators need write permissions for experiment and analysis metadata tables in order to upload new sample metadata or register their performed analyses. Other tables such as vocabularies, workflow and workflow compute tables only require read permissions. System administration entities in Trecode not relevant for data analyses, may even be completely invisible to operators. Wet-lab researchers typically use the platform to view sample relations and computational analyses results and are therefore granted read access to those tables. Bioinformaticians responsible for implementing analysis parameter optimizations and bugfixes have write permissions to all genomics data analyses tables. Making changes in the table structures, data model, user/role permission settings and GUI configuration requires access to the system tables which is restricted to administrators.

Row level security in Molgenis is under active development and is being implemented in Trecode for shared tables that contain sensitive data. The user's profile (roles and group memberships) is determinant for getting access to a particular row in these tables. Having controlled access in place allows to open up the platform instance to a wider audience and encourages (meta) data sharing without sacrificing data privacy and security.

Data governance

Enforcing data structure and terminology standards is crucial for data reusability and interoperability, as described in the FAIR guiding principles (14). In addition, adhering to these guidelines opens possibilities to increase automation of data analysis, making the system more scalable. We structure the data by unambiguously describing entity relationships using a common data model similar to SRA/EGA (13, 16). Entities and attributes in this model are annotated using well-established and maintained ontology terms which provide computers with the meaning of the data objects. Currently Trecode has an ontology reference for 480 of the total 498 attributes, and 58 of the 64 entities.

Trecode provides three routes of metadata entry; i) Operators can interactively insert rows in the database tables, ii) upload metadata in comma separated or excel files via the web GUI, or iii) upload metadata programmatically via the REST interface. In all cases the metadata is checked

for consistency and adherence to the implemented data standards, ensuring compatibility with the SRA/EGA (13, 14) data model.

Data integrity and reusability of outsourced data storage (19) is enforced through the md5 checksum and use of standardized metadata encapsulated output formats such as bam, cram, hdf5 and g(vcf). During data analysis, workflows check sample and file metadata for file integrity and to detect sample-file header inconsistencies. Furthermore, files in Trecode are required to have a Uniform Resource Locator (URL) using an open standard protocol. This ensures that data in the platform is findable and accessible regardless if data is stored on a webserver, object store and/or cloud bucket supporting the transfer of computational analyses from on-premise to cloud compute environments.

Records that are created in sample, NGS experiment or analyses result related tables have unique human readable and context traceable identifiers describing the source and type. These pre-configured Trecode identifiers have a human recognizable pattern and show similarity to the accessions used in the Encode project (20). The identifiers are in the format [A-Z]{3}[ID|DN|BS|BM|LB|RX|RR|RZ|AS|IC|IS|PR|ES|PO|DS|TR|PH|PN|WI|WE][0-9]{3}[A-Z]{3} where [ID|DN|BS|BM|LB|RX|RR|RZ|AS|IC|IS|PR|ES|PO|DS|TR|PH|PN|WI|WE] represents the metadata type (Supplemental table 1). The first three letters of the identifier are used to define the institute or data source of origin and the tailing 6 characters is the free namespace resulting in >17 million accessions per data source or entity type.

All workflows include tasks responsible for systematically storing analysis results and linking them to the sample and experiment metadata. The goal is to build a track-and-trace system for (bio)materials and their (analysis) products ensuring data provenance. The origin of analysis result entities is made transparent on the basis of consistent and descriptive label and titles that include Trecode identifiers involved and the experimental context. In addition, analysis result file names are automatically and systematically composed by concatenating identifiers from Trecode entities used and extended by a unique Trecode analysis id and experimental context. This results not only in human readable file names, but also sufficient context provided to the user/operator indicating what information can be expected in the file.

Data analysis

Bioinformatic analyses for variant detection in WGS and WXS data is a multistep and multi-software process. Analyses are often performed in a stepwise manner, with the output of one analysis being used as the input for the subsequent step. These analyses workflows are often designed to run automatically without human intervention. Bioinformatic workflows for variant discovery in NGS data typically includes steps that perform: 1) sequence read quality control, 2) read alignment to reference, 3) single-nucleotide polymorphism (SNP) and small insertion or deletion (indel) detection, 4) variant filtering, and 5) variant annotation (Figure 3). Due to the high similarity in variant discovery between WGS and WXS, the analysis workflows can be shared. However, errors in the sequencing experiment setup and platform biases differ per sequencing context and might require specific QC steps to monitor and detect deviations. In addition, individual steps in a workflow may have specific requirements for genomics resources. The computational resources for running a WXS or WGS workflow might also differ by an order of magnitude due to differences in data volume. Nonetheless, the similarity of WXS and WGS variant discovery is considerable, allowing for a common workflow definition as long as it has a modular design and is highly configurable. This concept can be extended to the task level, to further exploit similarity in omics data analyses by creating generic and configurable task definitions which can be reused across multiple analyses contexts.

Integrated data analysis workflows

Currently the GATK best practices (21) sequencing analyses based variant detection workflows have been implemented in Trecode. These include germline SNV (including QC), somatic-SNV and somatic-CNV. In addition, we have implemented transcriptome analysis for SNVs (GATK4), quantification and gene fusion detection (STAR-Fusion (22)). All of the aforementioned workflows generate QC metrics that are merged and uniformly parsed using MultiQC (23). For performing sequencing provider and platform specific noise reduction in somatic variant detection workflows, Trecode supports queries for creating, storing and (re)using panel of normals (18). Workflows that are currently under active development, and already supported in Trecode, are structural variant detection as well as DNA methylation profiling. For data visualization and publishing purposes the Trecode platform includes a workflow that outputs the results of the somatic variant detection analyses in a format that is compatible with

the visualization platform cBioPortal (24) and a workflow for automated data submission to EGA (13) using the REST protocol.

The workflows are all designed to be modular. In most cases, the workflows task definitions are imported from so called task containers that are listed as workflow dependencies and registered in the workflowfile table (Figure 3). Task definitions are generic and configured per workflow instance using key value pairs in the tables ‘workflowparameters’, ‘workflowcompute’ and ‘workflowresources’. For example, WGS and WXS germline analyses refer to the same workflow definition code at workflow instance level. The values for workflow parameters genomic targets (targetIntervalBedfile) and qc config (multiqc_config) determine if the step HSmetrics (hybridization metrics), that is defined in task container “Picard tasks”, is performed and what QC metrics are collected and how they are presented. Likewise, tasks have configurable runtime compute resource requests that are defined per workflow instance and can be optimized per analysis context. For example, the workflow step to identify SNV’s in the task container GATK4 is used for both WGS and WXS but the memory, cpu hours and tmp disk space differs between sequencing strategies. In addition to modulated use of tasks, complete workflows can also be called as a subworkflow. For example, variants identified in a SNV workflow are annotated with VEP (25) using a generic Vep workflow that is called as a sub workflow and controlled by a number of analysis context specific parameters. This same annotation workflow can be used for both germline SNVs, somatic SNVs as well as SNVs identified in RNA data. This exemplifies that in Trecode the reuse of analysis workflow code is not restricted to task level but can also be used in workflow nesting which offers high flexibility and low code redundancy in creating new workflows.

Analyses reproducibility by automation and tracking system

A fundamental issue with inter-project comparisons and long-term studies is the reproducibility of pipeline results. To achieve this, we have automated the data analyses and implemented a computational analysis tracking system in Trecode (Figure 4). The analyses metadata provides detailed information on how data was created, including which tool versions, parameters and genomics resources were used during data analyses. In addition, Trecode implements workflow instance versioning across workflow and task definition code as well as logging per workflow instance which software versions, parameters and genomics resources are used. This provides a

complete audit trail of developer and operator activity across all analyses (Figure 4). In some cases, default parameters may need to be overridden, a functionality that is provided by Trecode's client interfaces. These overrides, being part of the workflow execution information, are also included in the audit trail. For long term capturing of workflow execution parameters and statistics we have equipped Trecode with a persistent implementation of the Cromwell workflow executor (10).

Cromwell workflow execution

Trecode uses workflow definition language (WDL) to define workflows which are executed by the Cromwell executor (10). The main considerations for using this workflow description and execution software are human readability of the domain specific code, compute backend agnosticism, task result caching, error reporting, and the availability of a REST API. Advantages of using a workflow executor are that explicit calls to the compute backend can be omitted, keeping the code base small and maintainable while maintaining the portability of the workflows from e.g. "classical" HPC to cloud-based compute environments. The Cromwell workflow execution service will translate the implicit request for compute resources to specific calls on the underlying compute infrastructure. When a workflow is resumed Cromwell's task level caching makes analyses in Trecode computational and time-wise efficient as only failed or unfinished workflow tasks will be restarted. Workflows can be submitted (Figure 4) and subsequently monitored in detail for status and progress in Cromwell using REST API calls which allows for easy integration in data analyses platforms, client programs and user frontends.

Workflow Manager

The execution and monitoring of NGS analysis workflows is complex due to the many steps, the execution time of workflows and interaction between multiple samples. To maintain scalability and assist operators, Trecode provides the operator with a well-arranged interface, named 'Workflow Manager' which communicates via REST API with Molgenis and Cromwell (Figure 5). Within the Workflow Manager, routine analyses can be easily started and monitored for progress. In addition, via this web interface, links to completed analysis files can be retrieved and failed analyses cleaned up. The Workflow Manager is aimed at streamlining routine analysis as an alternative to the more versatile, but less intuitive, Trecode command line clients.

Discussion

By combining the open source initiatives Molgenis (12) and Cromwell (10) we have created Trecode, a (meta) data warehouse and analysis infrastructure that can be applied in both NGS molecular diagnostics as well as research activities. As part of this shared infrastructure we developed a data model in which diverse genomics experiment designs, analyses and their inter-relationships can be described. The tight integration of sample metadata with computational analyses assists operators in performing sensible analyses and provides an audit trail of parameters used and analysis results generated. Unique to the platform is the flexibility of the metadata model provided by Molgenis (12) together with an agnostic compute backend provided by Cromwell (10) and not seen in other NGS analyses platforms (see Table 2). The flexibility of the data model allows further integration of experiments such as (single cell) transcriptome sequencing and array-based methylation analysis with minimal effort.

In line with current international efforts of standardizing workflow descriptions (11), analysis workflows in Trecode are written using WDL (7) and are executed by the Cromwell workflow executor (11). When generating workflow code, our emphasis is on reuse, which has resulted in a compact non-redundant and well documented code base which is easy to maintain, extend and reuse. In order to address the challenge of sharing the bioinformatics tools and produce standardized and completely reproducible analyses (11) we are migrating from LMOD (26) software modules to container-based software stacks (27, 28) resulting in a code base free of local compute infrastructure dependencies. The final aim is making our workflows lab/site agnostic and cloud compute ready. This will additionally provide options to scale-out to the cloud and move workflows to data sources instead of the current approach of moving data to compute workflows.

Metadata are essential to understanding, interpreting and evaluation of genomic assays including the reuse of analyses results. By committing to FAIR data (14) principles we aim for complete data reusability. This will further be achieved by metadata extensions increasing the use of defined terms from ontologies such as SNOMED CT (clinical) (29), NCIT (translational research) (29), OBI (biomedical investigations) (30), EFO (experiments) (32), and EDAM

(bioinformatics)(33). Using curated definitions that are logically connected facilitates other researchers to computationally explore and understand our data and finding purposes for re-use.

Collecting rich metadata is a challenge in itself, as it often needs to be retrieved from multiple disciplines and sources and includes items that researchers do not deem relevant for their own research questions. We think that a platform such as Trecode which 1) enforces metadata collection at sample registration, 2) facilitates metadata handling through structure and integration, and 3) facilitates programmatic access to metadata shows its added value. Investing in the collection of high-quality metadata is indispensable for driving the use of in house or public data portals and results in improved data citation and credit in the scientific field.

A number of NGS data management and analysis platforms are available, each with its specific features and capabilities (see Table 2). Most platforms support per workflow reproducibility of standard data analyses but lack the data provenance features of Trecode for analyses that involve two or more subsequent workflows. Analyses parameter tuning on standard analyses is well supported by the Trecode, HTS-Flow (34), Closha (35) and Terra (36) platforms. Analyses parameters can be modified, but these changes are only recorded in the Trecode and Terra platforms by their workflow executor. Having a complete audit trail for future reference is a key feature in reproducible bioinformatics. Trecode is the only platform that offers the flexibility of allowing any parameter adjustment at the moment of workflows execution using a parameter override function in the command line client and workflow manager GUI while still providing an audit trail. Moreover, all workflow resources and parameters are stored and annotated in Trecode which makes it easier to find their meaning and encourages re-use across different workflows. Operators and lab staff welcome the integrated experiment QC and computational error reports provided exclusively by Trecode. However, analysis result visualization such as provided in OTP (37) and QuickNGS (38) is limited in Trecode. From the point of view of good software design, we separated the presentation layers by delegating result visualization to dedicated platforms such as cBioportal (24) and R2 (39) in pediatric oncology.

Trecode and OTP are the only platforms with a (meta)data model that shows similarity to SRA/EGA. Data model similarity is required for automated data submission and ingestion without the need for manual completion, and formatting of the data and metadata. With some

effort such a data model can be built in Terra, but functionality via a REST API and web-based navigation is lacking. Flexibility in adapting and extending the data model is superior in Trecode because it is built on Molgenis (12), which provides a framework to generate interoperable platforms based on a custom data model.

The Trecode platform demonstrates how FAIR data principles can be implemented in the context of NGS sequencing, harmonize (meta)data capture and representation and facilitate large-scale data (re)use. The platform can be used in a combined clinical and research setting to maximize translation of knowledge between research and healthcare, while addressing the requirements of all stakeholders, including patients/participants whilst addressing ethical, legal and social aspects. Several similar endeavors are being undertaken to promote NGS data and knowledge transfer. Currently guidelines are being drafted by multidisciplinary delegates from academic medical and research centers to facilitate large-scale (re)use of all human genomic data in the Netherlands. By announcing Trecode we aim to narrow the gap between the proposed far-reaching recommendations and its practical implementation in a clinical and research setting.

References

1. Höllein,A., Twardziok,S.O., Walter,W., Hutter,S., Baer,C., Hernandez-Sanchez,J.M., Meggendorfer,M., Haferlach,T., Kern,W. and Haferlach,C. (2020) The combination of WGS and RNA-Seq is superior to conventional diagnostic tests in multiple myeloma: Ready for prime time? *Cancer Genet.*, 10.1016/j.cancergen.2020.01.001.
2. Smadbeck,J., Peterson,J.F., Pearce,K.E., Pitel,B.A., Figueroa,A.L., Timm,M., Jevremovic,D., Shi,M., Stewart,A.K., Braggio,E., *et al.* (2019) Mate pair sequencing outperforms fluorescence in situ hybridization in the genomic characterization of multiple myeloma. *Blood Cancer J.*, **9**.
3. Kyrochristos,I.D., Ziogas,D.E., Goussia,A., Glantzounis,G.K. and Roukos,D.H. (2019) Bulk and Single-Cell Next-Generation Sequencing: Individualizing Treatment for Colorectal Cancer. *Cancers*, **11**, 1809.
4. Kumar-Sinha,C. and Chinnaiyan,A.M. (2018) Precision oncology in the age of integrative genomics. *Nat. Biotechnol.*, **36**, 46–60.
5. Seibel,N.L., Janeway,K., Allen,C.E., Chi,S.N., Cho,Y.-J., Glade Bender,J.L., Kim,A., Laetsch,T.W., Irwin,M.S., Takebe,N., *et al.* (2017) Pediatric oncology enters an era of precision medicine. *Curr. Probl. Cancer*, **41**, 194–200.
6. Kulkarni,P. and Frommolt,P. (2017) Challenges in the Setup of Large-scale Next-Generation Sequencing Analysis Workflows. *Comput. Struct. Biotechnol. J.*, **15**, 471–477.
7. Frazer,S. (2014) Workflow description language. <https://software.broadinstitute.org/wdl/>.
8. Amstutz,P. (2016) Common Workflow Language <https://github.com/common-workflow-language/common-workflow-language>.
9. Vivian,J., Rao,A.A., Nothaft,F.A., Ketchum,C., Armstrong,J., Novak,A., Pfeil,J., Narkizian,J., Deran,A.D., Musselman-Brown,A., *et al.* (2017) Toil enables reproducible, open source, big biomedical data analyses. *Nat. Biotechnol.*, **35**, 314–316.
10. Cromwell homepage. <https://cromwell.readthedocs.io/en/stable/>.
11. Fjukstad,B. and Bongo,L.A. (2017) A Review of Scalable Bioinformatics Pipelines. *Data Sci. Eng.*, **2**, 245–251.
12. van der Velde,K.J., Imhann,F., Charbon,B., Pang,C., van Enckevort,D., Slofstra,M., Barbieri,R., Alberts,R., Hendriksen,D., Kelpin,F., *et al.* (2019) MOLGENIS research: advanced bioinformatics data software for non-bioinformaticians. *Bioinformatics*, **35**, 1076–1078.
13. Leinonen,R., Akhtar,R., Birney,E., Bower,L., Cerdeno-Tarraga,A., Cheng,Y., Cleland,I., Faruque,N., Goodgame,N., Gibson,R., *et al.* (2011) The European Nucleotide Archive. *Nucleic Acids Res.*, **39**, D28–D31.

14. Wilkinson,M.D., Dumontier,M., Aalbersberg,I.J., Appleton,G., Axton,M., Baak,A., Blomberg,N., Boiten,J.-W., da Silva Santos,L.B., Bourne,P.E., *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**.
15. Fielding,R.T. (2000) Architectural Styles and the Design of Network-based Software Architectures.
16. Shumway,M., Cochrane,G. and Sugawara,H. (2010) Archiving next generation sequencing data. *Nucleic Acids Res.*, **38**, D870–D871.
17. González-Beltrán,A., Maguire,E., Sansone,S.-A. and Rocca-Serra,P. (2014) linkedISA: semantic representation of ISA-Tab experimental metadata. *BMC Bioinformatics*, **15**.
18. Cibulskis,K., Lawrence,M.S., Carter,S.L., Sivachenko,A., Jaffe,D., Sougnez,C., Gabriel,S., Meyerson,M., Lander,E.S. and Getz,G. (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, **31**, 213–219.
19. Imran,M., Hlavacs,H., Haq,I.U., Jan,B., Khan,F.A. and Ahmad,A. (2017) Provenance based data integrity checking and verification in cloud environments. *PLOS ONE*, **12**, e0177576.
20. Hong,E.L., Sloan,C.A., Chan,E.T., Davidson,J.M., Malladi,V.S., Strattan,J.S., Hitz,B.C., Gabdank,I., Narayanan,A.K., Ho,M., *et al.* (2016) Principles of metadata organization at the ENCODE data coordination center. *Database*, **2016**, baw001.
21. GATK Best Practices. <https://software.broadinstitute.org/gatk/best-practices/about>.
22. Haas,B.J., Dobin,A., Stransky,N., Li,B., Yang,X., Tickle,T., Bankapur,A., Ganote,C., Doak,T.G., Pochet,N., *et al.* (2017) STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq Bioinformatics.
23. Ewels,P., Magnusson,M., Lundin,S. and Käller,M. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048.
24. Cerami,E., Gao,J., Dogrusoz,U., Gross,B.E., Sumer,S.O., Aksoy,B.A., Jacobsen,A., Byrne,C.J., Heuer,M.L., Larsson,E., *et al.* (2012) The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov.*, **2**, 401–404.
25. McLaren,W., Gil,L., Hunt,S.E., Riat,H.S., Ritchie,G.R.S., Thormann,A., Flicek,P. and Cunningham,F. (2016) The Ensembl Variant Effect Predictor. *Genome Biol.*, **17**, 122.
26. Lmod: Environmental Modules System. <https://www.tacc.utexas.edu/research-development/tacc-projects/lmod>.
27. Merkel,D. Docker: Lightweight Linux Containers for Consistent Development and Deployment. *Linux J.*, **2014**.

28. Kurtzer,G.M., Sochat,V. and Bauer,M.W. (2017) Singularity: Scientific containers for mobility of compute. *PloS One*, **12**, e0177459.
29. El-Sappagh,S., Franda,F., Ali,F. and Kwak,K.-S. (2018) SNOMED CT standard ontology based on the ontology for general medical science. *BMC Med. Inform. Decis. Mak.*, **18**, 76.
30. Golbeck,J., Fragoso,G., Hartel,F., Hendler,J., Oberthaler,J. and Parsia,B. (2003) The National Cancer Institute's Thesaurus and Ontology. *SSRN Electron. J.*, 10.2139/ssrn.3199007.
31. Bandrowski,A., Brinkman,R., Brochhausen,M., Brush,M.H., Bug,B., Chibucos,M.C., Clancy,K., Courtot,M., Derom,D., Dumontier,M., *et al.* (2016) The Ontology for Biomedical Investigations. *PLOS ONE*, **11**, e0154556.
32. Malone,J., Holloway,E., Adamusiak,T., Kapushesky,M., Zheng,J., Kolesnikov,N., Zhukova,A., Brazma,A. and Parkinson,H. (2010) Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, **26**, 1112–1118.
33. Ison,J., Kalas,M., Jonassen,I., Bolser,D., Uludag,M., McWilliam,H., Malone,J., Lopez,R., Pettifer,S. and Rice,P. (2013) EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, **29**, 1325–1332.
34. Bianchi,V., Ceol,A., Ogier,A.G.E., de Pretis,S., Galeota,E., Kishore,K., Bora,P., Croci,O., Campaner,S., Amati,B., *et al.* (2016) Integrated Systems for NGS Data Management and Analysis: Open Issues and Available Solutions. *Front. Genet.*, **7**.
35. Ko,G., Kim,P.-G., Yoon,J., Han,G., Park,S.-J., Song,W. and Lee,B. (2018) Closha: bioinformatics workflow system for the analysis of massive sequencing data. *BMC Bioinformatics*, **19**, 43.
36. Terra cloud-native platform for biomedical researchers. <https://app.terra.bio/>.
37. Reisinger,E., Genthner,L., Kerssemakers,J., Kensche,P., Borufka,S., Jugold,A., Kling,A., Prinz,M., Scholz,I., Zipprich,G., *et al.* (2017) OTP: An automatized system for managing and processing NGS data. *J. Biotechnol.*, **261**, 53–62.
38. Wagle,P., Nikolić,M. and Frommolt,P. (2015) QuickNGS elevates Next-Generation Sequencing data analysis to a new level of automation. *BMC Genomics*, **16**, 487.
39. R2 bioinformatics platform. <http://r2.amc.nl>.

Funding

This work was financially supported by grants from KiKa and Adessium Foundation.

Figures

Figure 1: Overview of the Trecode platform.

Panel A depicts a graphical representation of the biobanking activities. Panel C depicts the architecture for computational analyses. The Trecode data model, of which the core entities are shown in panel B, describes the wet-lab experiment as well as the computational setup in detail and interconnects the two. Trecode makes experimental data programmatically accessible to operators, records computational analyses performed and takes care that analyses results are consistently linked back to the experimental data. For clarification, some tables are linked by dotted orange lines to the biobank or computational entity that they represent.

Figure 2: Overview of the core database model without vocabulary/ontology codebooks and Molgenis system tables.

The top row of tables describes the samples and their relationships in the wet-lab sequencing experiments. The tables at the bottom half of the diagram describe the computational analysis workflows including the genomics resources (references, variant databases) and computational resources required for workflow execution. Data integrating tables that describe sets of wet-lab experiment data or the results of computational analysis are in the middle (shaded).

Figure 3: Registration and re-use of genomics resources as well as workflow and task scripts in Trecode.

In the center a representation of a germline variant detection workflow in which the processing steps are indicated in red. The processing tasks are non-redundantly grouped by software or function, described and stored in table ‘Workflow_file’ (left side) and where applicable reused in other workflows. Required reference genome, the genomic targets and population variant databases are described in table ‘Genomics_resources’ (right side). The workflow is used to analyze both whole exome and whole genome data. Workflow parameters (top left) are analysis context specific and e.g. determine which genomic targets will be analyzed and if target specific metrics will be collected. Context specific computational parameters are stored in table ‘workflow_compute’.

Figure 4: Highly automated genome analyses in Trecode.

- 1) Genome analysis is started by providing the Workflow Manager or Trecode’s python command line client with a biomaterial, a library strategy (WGS/WXS/RNA-Seq), an analysis type and a genome version.
- 2) The client program queries Trecode using the Molgenis REST interface for run data and a computational analysis workflow being the core workflow ‘workflowSource’, an input template ‘workflowInputs’ and generic reusable workflow tasks ‘workflowDependencies’.
- 3) The inputs template is filled in by the command line client using queried experiment metadata and together with workflow core and dependencies submitted via a REST protocol to the Cromwell workflow executor.
- 4) Workflow execution is registered in Trecode as a ‘workflowrun’ record and successful workflow runs will have analyses records attached.
- 5) Analyses records are created by results archiving workflow tasks and include references to analyses inputs as well as result files ensuring data provenance.
- 6) All workflows include (shared) tasks for systematic archival of analysis results.

Figure 5: Interfaces for operators, bioinformaticians and researchers.

Workflowmanager (top left) aids operators in performing their daily duties. Client scripts help bioinformaticians in performing sensible analyses (top right). Molgenis Web-GUI (bottom) is an easy to use graphical Trecode querying tool for researchers involved in biobanking and clinical diagnostics.

Tables

Role	Vocabularies Ontologies	Workflows Parameters	Samples Metadata	Genomics Analysis	System	Trecode Users
Admin	R+W	R+W	R+W	R+W	R+W	admins
Manager	R+W	R+W	R+W	R+W	R	bioinformaticians
Editor	R	R	R+W	R+W	-	operators, bioinformaticians
Viewer	R	R	R	R	-	(wet-lab) researchers

Table 1: Roles and their permissions in Trecode. R=Read, W=Write, -=invisible

See file: TrecodeTable2.xlsx

Table 2: Feature comparison of currently available genomics (meta) data management and processing platforms.

Supplemental Figures and Tables

Supplemental figure 1: Data model for enrolling patients to studies.

Patients can be included in multiple studies and an informed consent is required for each study. In the crosstab ‘individual_study’ in the Trecode data model, it is recorded per patient - study combination which informed consent applies.

Supplemental figure 2: Description of DNA-methylation experiments in Trecode data model.

Array based DNA-methylation experiments are described similar to sequencing experiments. The preparation of the DNA sample (biomaterial) is described for array experiments in table ‘assay’. As for sequencing experiments, the results (raw data) are linked to run. Database modeling and integration of the computational analyses is currently under active development.

Supplemental figure 3: Automated data submission to EGA.

A) Ahead of sample metadata submission, queried sample data is EGA-pgp encrypted using Trecode’s data encryption workflow (1). Checksums of the encrypted files, which are required for metadata submission, are automatically recorded in Trecode and encrypted files are transferred to EGA (2). A similar query is performed by the operator for metadata submission in which all required metadata for the queried samples and experiments is gathered by the Trecode platform and submitted to EGA’s REST interface (3). The returned receipt is processed, and relevant information is passed to the operator and stored in the Trecode database (4).

B) The Trecode platform keeps track on EGA submissions using the tables ‘filehashes’ for storing encrypted file checksums and ‘submissions’ for keeping track on which sample batches have been submitted and when. Table ‘mappings’ records the relation between the assigned identifiers by EGA and Trecode identifiers. Datasets for publications, which are stored in ‘dataset’, are composed from submitted runs and analyses in the Trecode web GUI. EGA datasets require a reference to a data policy (‘policy’) and a data access committee (‘dac’).

Identifier (two letter code)	Metadata type
ID	Individual
DN	Diagnosis
BS	Biosource
BM	Biomaterial
LB	Library
RX	Experiment
RR	Run
RZ	Analysis
AS	Assay
IC	Informed consent
PR	Project / Study
ES	EGA submission
PO	EGA policy
DA	EGA data access committee
DS	EGA dataset

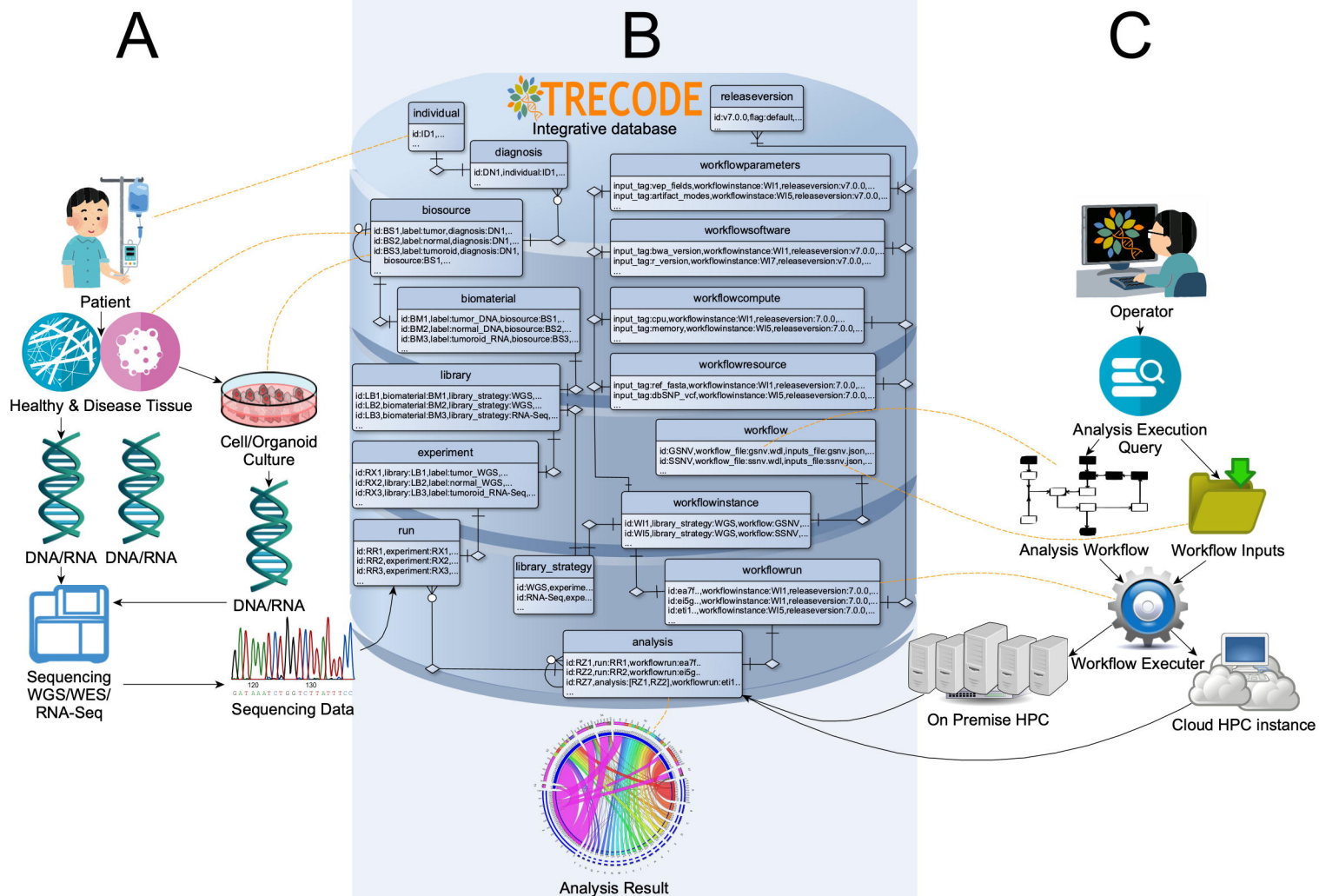
TR	Treatment
PH	PDX host
PN	PanelOfNormals
BC	Barcode
WI	Workflow instance
WE	Workflow environment

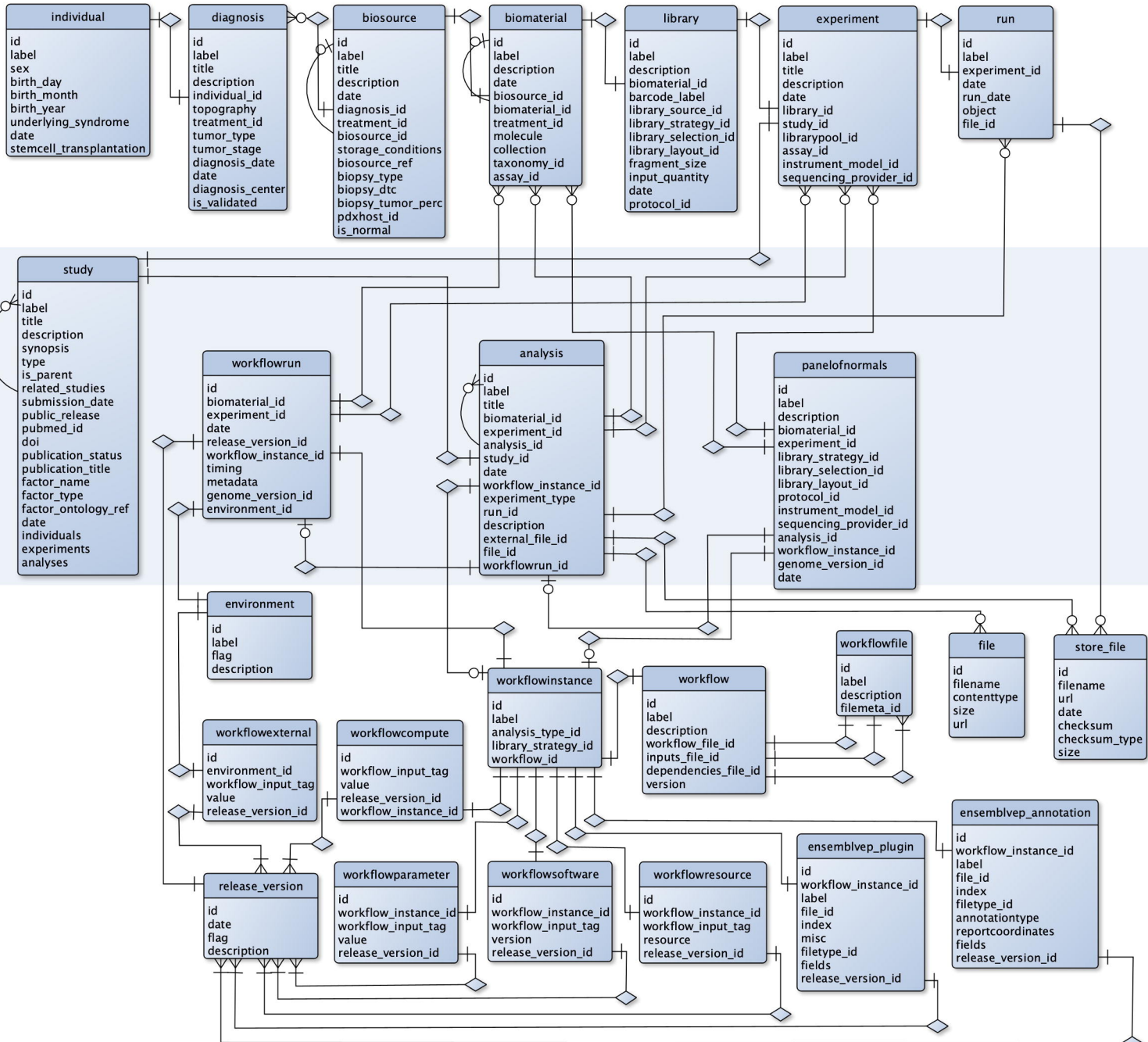
Supplemental table 1: Trecode metadata type by two letter code identifiers

A

B

C



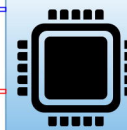


Workflow parameters

"multiqc_config_file": "/path/to/germline_snv_WXS.yaml"
"targetIntervalBedFile": "/path/to/hg38_diagnostic_somatic_2.0.bed"

"multiqc_config_file": "/path/to/germline_snv_WGS.yaml"
"targetIntervalBedFile": "null"

Workflow compute



CPU



Local disk
space



Memory

Start

Localize data

Stored sequencing
reads

Direct-access
sequencing reads

Align DNA

Reference
genome

Aligned reads

Sort

Sorted
aligned reads

Target Metrics

Metrics

SNV caller

Genomic
targets

QC metrics

Called Variants

QC variants

Filter

Filtered
called variants

Unify metrics

Annotate

Population
variant
database

Unified QC
metrics+report

Annotated filtered
called variants

Archive results

Archive results

Stored analysis
workflow results

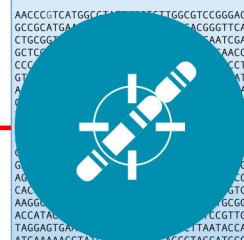
TRECODE
TRECODE
database

End

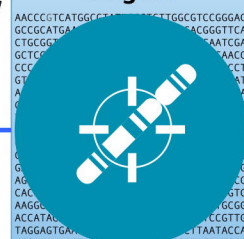
Genomics resources



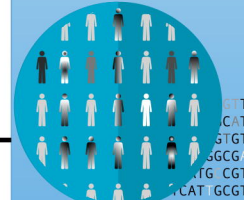
Reference genome



WGS genomic
targets



WXS genomic
targets



Population variant
database

WGS exclusive

WXS exclusive

Common

Workflow file



Sort



Filter



SNV caller QC variants

GATK4 tasks



Collect target
metrics



Collect
metrics

Picard tasks



Localize
data



Archive
results

Data transaction tasks



Unify metrics

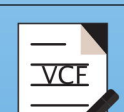


FastQC

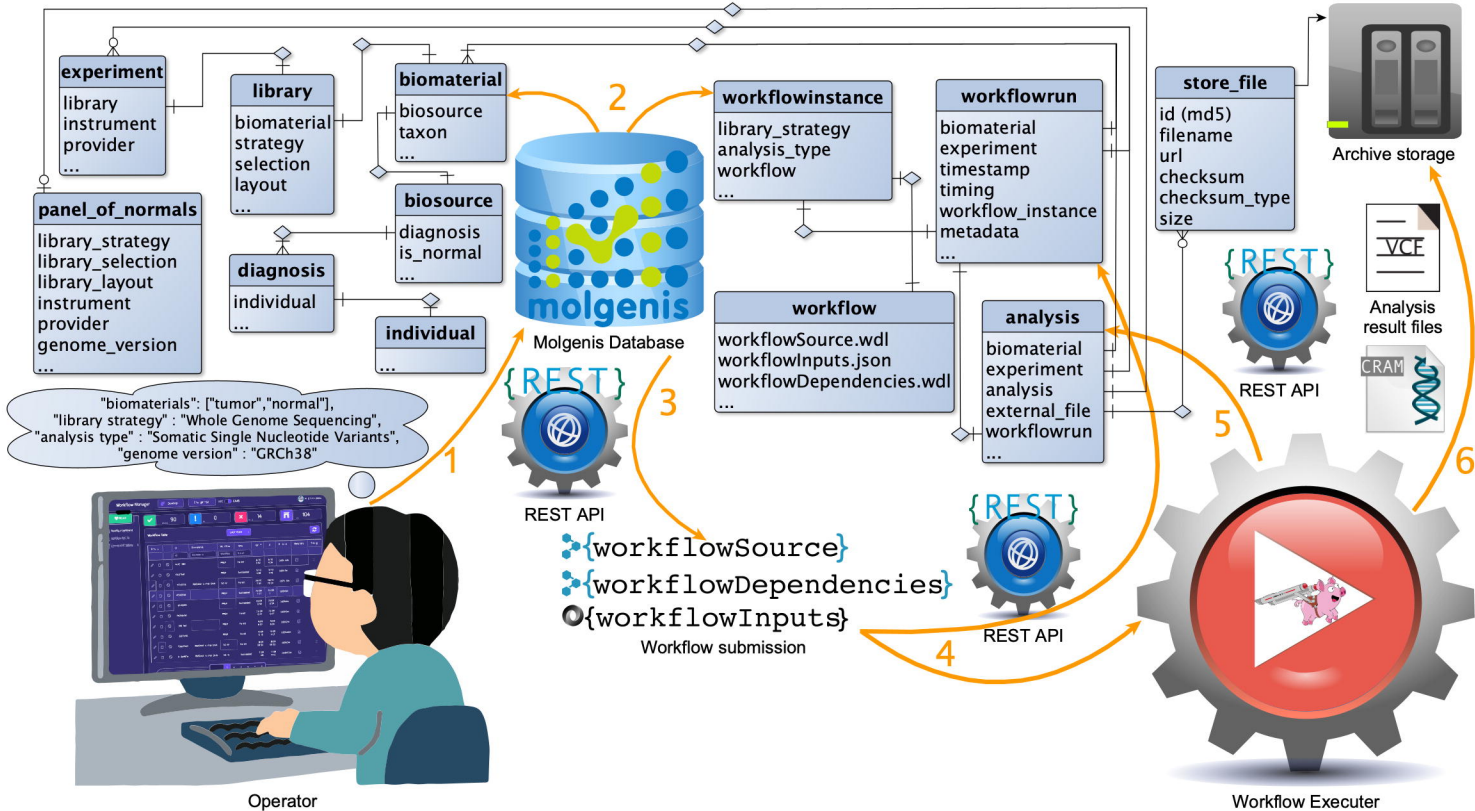
Quality Control tasks



Align DNA
Alignment
tasks



Annotate
Annotation
tasks



Workflow Manager

Production

Change User

HPC

CAAS

Eugene Verweil

Workflow Dashboard

Workflow Results

Command Templates

SUCCESS

64

!

RUNNING

0

✖

FAILED

0

🏠

TOTAL

64

Workflow Table

LAST 3 WEEKS

Hide preprocessing workflows

Actions	ID	Biomaterial	Workflow	Status	Start	End	Runtime	Metadata	Timing
	ID	Biomaterial	Workflow	Status					
	35dcf32b...	PMLBM000BFE	SSNV	Succeeded	29/10 08:26	29/10 09:38	0d1h12m		
	0968c9e4...	PMLBM000BGP	SSNV	Succeeded	28/10 10:22	28/10 15:40	0d5h17m		
	880d3a8d...	PMLBM000BFQ	SSNV	Succeeded	28/10 10:22	28/10 15:16	0d4h53m		
	aa00ac9e...	PMLBM000BEG	SSNV	Succeeded	28/10 10:22	28/10 15:37	0d5h15m		
	3be0c3b0...	PMLBM000BFE	SCNV	Succeeded	27/10 13:59	27/10 15:17	0d1h17m		
	6afcc745...	PMLBM000BGH	SCNV	Succeeded	27/10 11:06	27/10 12:46	0d1h40m		
	6c2f5d4c...	PMLBM000BGP	SCNV	Succeeded	27/10 11:06	27/10 13:07	0d2h1m		

usage: TreCode Pipeline Launcher [-h] -u USER -p PWD [--trecodeURL TRECODEURL] [--userPost USERPOST] [--pwdPost PWDPOST] [--trecodePostURL TRECODEPOSTURL] [-c CROMWELLURL] [-cu USERCROMWELL] [-cp PWD CROMWELL] [-cws CROMWELLWORKFLOWOPTIONS] [-r RUNIDS [RUNIDS ...]] [-a {GSNV,GSV,SCNV,SSNV,SSV,CNV PON,SNV PON,GeneFusionDetection,QcSnp sAdditional,Quantification} [{GSNV,GSV,SCNV,SSNV,SSV,CNV PON,SNV PON,GeneFusionDetection,QcSnp sAdditional,Quantification} ...]] [-d] [-b BIOMATERIALIDS [BIOMATERIALIDS ...]] [-o OVERWRITES] [-wd WORKINGDIR] [--ignoreTumorNormal] [-gv GENOMEVERSION] TreCode Pipeline Launcher: error: the following arguments are required: -u/--user, -p/--pwd

usage: EGASubmit [-h] -u USER -p PWD [--trecodeURL TRECODEURL] [--userPost USERPOST] [--pwdPost PWDPOST] [--trecodePostURL TRECODEPOSTURL] [-c CROMWELLURL] [-cu USERCROMWELL] [-cp PWD CROMWELL] [-cop CROMWELLOVERWRITESPARAMETERS] [-cof CROMWELLOVERWRITESFILE] -s STUDY [-x EXPERIMENTSUBSET [EXPERIMENTSUBSET ...]] -l {WGS,WXS,RNA-Seq} [-t TITLE] [-d DESCRIPTION] [-a ACCOUNT] [-A ACCOUNTPASSWD] [-n CENTERNAME] [-b DROPBOXURL] [-f {data,metadata}] [-d DATASETID] [-E ENTITYSUBSET [ENTITYSUBSET ...]] [-V VALIDATEONLY] [-R PROTECT] [-X DEBUG] [-o OVERWRITES] [-fakeStudy FAKESTUDY] EGASubmit: error: the following arguments are required: -u/--user, -p/--pwd, -s/--study, -l/--libraryStrategy

analysis

Entity describing the bioinformatic analysis of data

analysis

Delete

Data

Aggregates

Charts

Annotators

Identifier	Label	Title	Involved Biomaterials	Involved Experiments	Analysis entry date	Involved Analysis
PMCRZ564BZF	PMLBM000AZL_PMLBM000BJA_WXS_VCF	Somatic variant calling results	PMLBM000AZL,PMLBM000BJA	PMCRX000GKO,PMCRX000GKR	Nov 10, 2020 3:18 PM	PMLBM000AZL_WXS_GVCF,PMLBM000AZL_WXS_VCF,PMLBM000BJA_WXS_GVCF,PMLBM000BJA_WXS_VCF
PMCRZ083TYQ	PMLBM000BJL_PMLBM000BJB_WXS_VCF	Somatic variant calling results	PMLBM000BJL,PMLBM000BJB	PMCRX000GKQ,PMCRX000GLA	Nov 10, 2020 3:15 PM	PMLBM000BJL_WXS_VCF,PMLBM000BJB_WXS_VCF,PMLBM000BJL_WXS_CRAM,PMLBM000BJB_WXS_CRAM
PMCRZ645PWJ	PMABM000CEV_RNA-Seq_Quantification	Quantification levels of (meta) features.	PMABM000CEV	PMCRX000COO	Nov 10, 2020 3:11 PM	PMABM000CEV
PMCRZ209FKV	PMLBM000BIT_PMLBM000BIP_WXS_QCI_VCF	QCI Somatic variant calling results	PMLBM000BIT,PMLBM000BIP	PMCRX000GKN,PMCRX000GKZ	Nov 10, 2020 2:47 PM	PMLBM000BIT_WXS_CRAM,PMLBM000BIP_WXS_GVCF,PMLBM000BIT_WXS_QCI_VCF,PMLBM000BIP_WXS_QCI_VCF
PMCRZ212NTI	PMABM000CEV_RNA-Seq_CRAM	Mapping results	PMABM000CEV	PMCRX000COO	Nov 10, 2020 2:42 PM	PMABM000CEV
PMCRZ151KVL	PMABM000BBX_RNA-Seq_Quantification	Quantification levels of (meta) features.	PMABM000BBX	PMCRX000CTV	Nov 10, 2020 2:33 PM	PMABM000BBX
PMCRZ324IKD	PMLBM000BJP_PMLBM000BJJ_WXS_QCI_VCF	QCI Somatic variant calling results	PMLBM000BJP,PMLBM000BJJ	PMCRX000GKW,PMCRX000GLC	Nov 10, 2020 2:31 PM	PMLBM000BJJ_WXS_CRAM,PMLBM000BJP_WXS_VCF,PMLBM000BJJ_WXS_QCI_VCF,PMLBM000BJP_WXS_QCI_VCF
PMCRZ474OSK	PMABM000BBX_RNA-Seq_CRAM	Mapping results	PMABM000BBX	PMCRX000CTV	Nov 10, 2020 2:23 PM	PMABM000BBX
PMCRZ221BMC	PMABM000DQI_RNA-Seq_MultiQC	MultiQC report	PMABM000DQI	PMCRX000DIG	Nov 10, 2020 2:19 PM	PMABM000DQI
PMCRZ049PVI	PMLBM000BIV_PMLBM000BIQ_WXS_MultiQC	MultiQC report	PMLBM000BIV,PMLBM000BIQ	PMCRX000GKX,PMCRX000GLD	Nov 10, 2020 2:10 PM	PMLBM000BIQ_WXS_GVCF,PMLBM000BIQ_WXS_VCF,PMLBM000BIQ_WXS_QCI_VCF,PMLBM000BIQ_WXS_QCI_VCF