Deliverable D8.3

Project Title:	Building data bridges between biological and medical infrastructures in Europe
Project Acronym:	BioMedBridges
Grant agreement no.:	284209
	Research Infrastructures, FP7 Capacities Specific Programme; [INFRA-2011-2.3.2.] "Implementation of common solutions for a cluster of ESFRI infrastructures in the field of "Life sciences"
Deliverable title:	Demonstration of interoperability between different types of PM data
WP No.	8
Lead Beneficiary:	16: UH
WP Title	Use case: Personalized Medicine
Contractual delivery date:	31 December 2015
Actual delivery date:	15 December 2015
WP leader:	16: UH, Imre Vastrik
Contributing partner(s):	1: EMBL, 3: KI, 5: UDUS, 16: UH, 19:UMCG

Authors: Imre Vastrik, Alexander Thorarinsson, Helen Parkinson, Roxana Merino Martinez, Morris Swertz, Joeri van der Velde, Christian Ohmann

Contents

1 Executive Summary

The purpose of D8.3 is to describe the prototype IT systems for storing, structuring as well as providing controlled access and annotation/fitering to the data pertinent for the personalised medicine (referred as the "PM system" in subsequent text). As already defined in D8.1 and D8.2 we use personalised medicine of leukemia as the use case and in D8.3 we added personalized medicine of new-born babies with metabolic disease as second use case.

The report starts with quick recapitulation of the use cases and the challenges identified. It then describes the data access control mechanisms followed by a walk-through of the PM systems: demonstrates the main views and utility of the systems as well as how these match with the stakeholder needs. The report also contains some statistics of the current usage and lays out future work. The report concludes by providing instructions for obtaining an account on the demo installations for 1st hand experience using the systems.

2 Project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives:

No.	Objective	Yes	No
1	Develop a process for secure sharing of and access to PM data		x
2	Define types of PM data and mapping between them		x
3	Develop a PM informatics pipeline	x	

# 3 Detailed report on the deliverable

3.1 Background

Personalised medicine uses the genetic profile and other molecular characteristics from a diseased individual or his/her samples in order to guide treatment (Topol 2014). "Omic" technologies, such as "Next Generation Sequencing" (NGS), are widely used in research and increasingly also in the medical practice, especially when dealing with cancer. In WP8, we have based our work on two very concrete use cases:

- The hematologists at the Helsinki University Central Hospital treating leukemia patients
- The genome-first differential diagnostics of new-born babies with acute metabolic disease using whole-genome sequencing at the neo-natal intensive care unit of the UMCG Medical Center Groningen.

These are good test cases for personalised medicine since the samples can be taken relatively easily and even in a serial manner, patients are concentrated into a medical speciality, and complex drug treatment regimes are required.

The leukemia sample aliquots are biobanked at the storage facility affiliated with the Finnish BBMRI node. The samples are subjected to molecular profiling – identification of somatic mutations, estimates of gene copy number changes and gene expression levels – at the Finnish EATRIS infrastructure and analysed for drug sensitivity and resistance at the Finnish EU-OPENSCREEN facility. Equipped with the results of these tests, the hematologists can then make more educated decisions on how to treat a particular patient, for example prediction of response to a particular drug treatment regime.
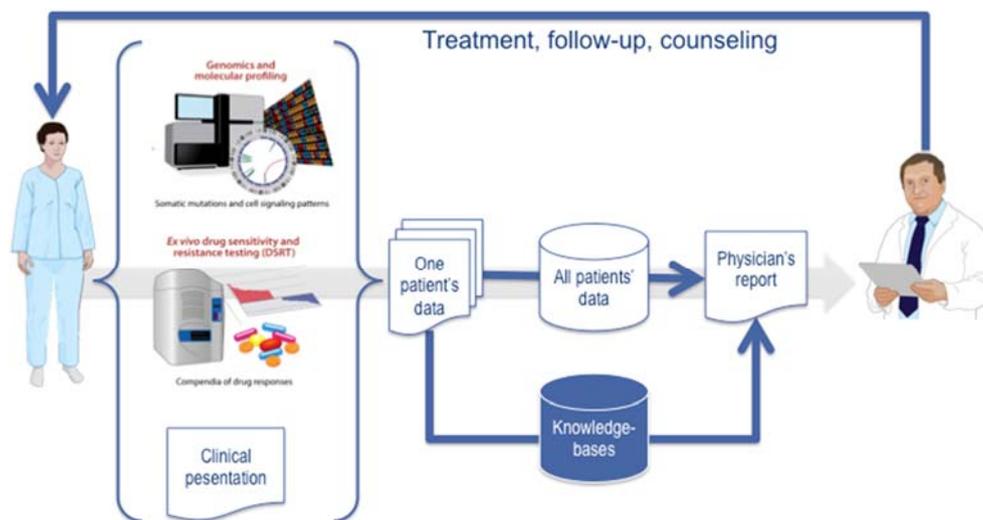
The DNA samples and phenotype profile (HPO terms) of new-born babies with acute metabolic disease are collected at the neo-natal care unit of the University Medical Center Groningen. These samples are immediately sequenced on Illumina 2500 using whole genome sequence protocol. Then genetic variants are called and the results are discussed in a multi-disciplinary team of clinicians, lab technicians, geneticists and bioinformaticians. The use case addressed in BioMedBridges WP8 was how to rapidly integrate and filter these data to inform actionable differential diagnosis of these newborns to chose treatment, i.e. identify potentially pathogenic mutations and check if the genes involved are known to affect relevant phenotypes observed in the new-borns.
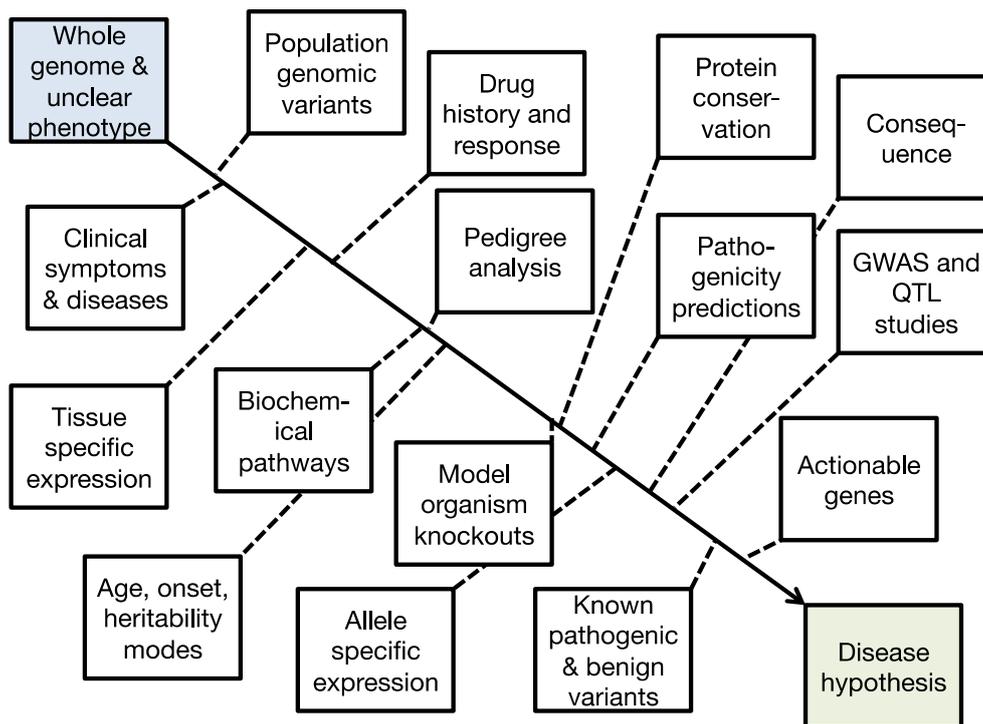
3.2 Personalised medicine: challenges and solutions

Whereas the approach is rather simple in principle the use case revealed several opportunities to improve the flow of information. The table below describes the challenges and unmet needs identified in D8.1 and D8.2 as well as the solutions to these as implemented in the prototype PM IT systems.

Challenge	Solution
Omics results in large data files, physicians (as well as majority of researchers) lack time and/or expertise to wade through these.	Patient and sample level summary reports that show only the most pertinent information while nevertheless enabling to delve/drill into entirety of the data.
Generation of the reports entails integration with the existing knowledge. This knowledge is not static but evolves and accumulates over time. As a result the same patient-derived data may provide additional or even different insights at different points in time. Also the recipes to interpret the data may change over time.	Create reports dynamically, i.e. a report is created at the time when a user requests to see it. For example, if the list of known cancer related genes is updated in the system, all the reports utilising such a list will immediately reflect the update.
A way of gaining insights on the patient from his/her data is by looking at the outcome of other	Structure the data of all the patients according to the same data model and store it in the same IT system,

patients with similar molecular characteristics. This necessitates that other patients' data be interoperable as well as accessible and "queryable".	i.e. use software as a service rather than on-premise software.
Whereas IT-literate users may be happy with downloading or querying large volumes of data which they can interrogate programmatically, the "ordinary" users - physicians, biologists, geneticists etc - appreciate the possibility to interact with the data via graphical user interface. Furthermore, even the computationally advanced users still face the challenge of having to understand how different types of data relate to each other.	Provide a browser based point-and-click user interface. For every data item provide an easy way to navigate to items referred by this item as well as items referring to this data item.
Different user roles need access to different "kinds" of data. Or put in other way – different kinds of data have different access restrictions. E.g. medical personnel treating the patient needs to access the personal details whereas other user roles typically don't and are even forbidden to see it.	Provide the access control mechanism that enables different user roles to have different access rights.
Prior to publishing reserachers typically want to limit the access to the data being collected. Such need is particularly understandable in commercially oriented research. Typically (in biomedical research) this is handled by creating different data (base) instances for each project/group of users. However, this means increased maintenance costs.	Provide the access control mechanism that enables different users to have different access rights so that different parties can share the same instance of software and data.
To enable data interpretation the patient data must be viewed within the context of relevant existing data.	Create an annotation tool to rapidly integrate patient data with existing data (e.g. known genetic variants) and allow for rapid filtering.

3.3 Data access control

The PM system uses software as a sevice (SaaS) software delivery model and is based on the multi-tenant architecture with all the data (and code) hosted centrally. Whereas this releaves the user from having to install and maintain software and hardware it requires the system to be able to partition the data so that every user can access only his/her data. In order to achieve this, every data item (referred as "document" in subsequent text) in the PM

system has an access control list (ACL) specifying the users which can read it.

In order to make the access rights management easier, users can also be "assembled" into user groups which can also be included in the ACLs. Thus a user can only see documents where the reading ACL either contains this user or a user group which has this user as a member.

Similarly, every document has also an ACL specifying who can edit it. Editing right also entails the right to grant with or revoke other users the right to view or edit this document. Adding a new user group to the readers ACL gives the members of this group the right to see this document. Similarly, removing a user from the list, revokes the access right of this user on the given document.

The PM system also aims to make the changing the access rights easy in practise. A user with the write-privileges can change the access rights of a single document, all documents returned by a query and even documents of specified data types related to, say, a person or sample (more on this below).

Our hope is that, perhaps counter-intuitively, such a controlled access mechanism actually facilitates data sharing and re-use. Rather than creating a downloadable data dump for a party interested in given data, the the data custodian can just add this party to the readers ACLs of relevant documents. Similarly, the requesting party is releaved from having to deal with raw data dumps as the PM system makes the data readily browsable and searchable.

As at some point we may need to be able to show who has accessed what data and when, the PM system logs all the interactions users have with it. As all the users have to be authenticated to access the data, the logs will hence not only have the information about the IP address of the user's device, but also the user's identity.

3.4 Walk-through 1 - case leukemia

In order to use the PM system the user needs to obtain an account for which the user needs to register with the PM system. As this is still work-in-progress we want to work with limited number of users and hence the registration is currently by invitation only.

Whereas the system allows to create system specific login credentials we reasoned that from the perspective of both the user and the developer, it would be easier to use some existing identity provider. This way the user has one less set of credentials to remember. Furthermore, the user wouldn't have to (re)create his/her own profile since this can be obtained from the identity

provider. Since the most widely used identity plaform among the immediate potential user base was Google, we opted to use Google Sign-in. However, this can be easily changed if the data owners or legislation requires it and we indeed already have the code to implement two-factor authentication from the user.

Provided that a user has a Google account and knows the registration token the user can enter the token and click the 'Register with Google account'. This will take the user to a Google sign-in page and ask user's permission for the PM system to get the e-mail address and public profile info, e.g. name. These will be used to identify the user in the PM system.

Provided that the user is authenticated by Google and the user grants permission for this application to access his/her information, the she/he is then taken to the PM system starting.

This page provides easy entry points to data categories most pertinent to the use case. Note that since the signed-in user belongs to a user group which has custom theme specified – in this case BioMedBridges' theme – the page is styled accordingly. For other users the theme may be different. The number in the right corner of each data category panel indicates the total number of entries in this category. For example, the number 449 in the 'Persons' panel indicates that there are this many Person entries in the PM system at the time of taking this screenshot. However, this does not mean that the given user can actually access all or any of them. In fact, a newly registered user can access only the data items which have been made visible to all users of the PM system. These data items are typically controlled

vocabulary terms for e.g. genes, diseases, chemicals, gender etc.

In order for the user to see the actual PM data the user needs to be given the reading rights to the relevant data entries. For the purpose of this demo we have created a user group named BioMedBridges to which we add individual users wanting to try this demo. This group has been given viewing rights to a limited set of data – samples and analysis results from three well-published patient cases [Pemovska 2013, Pemovska 2015].
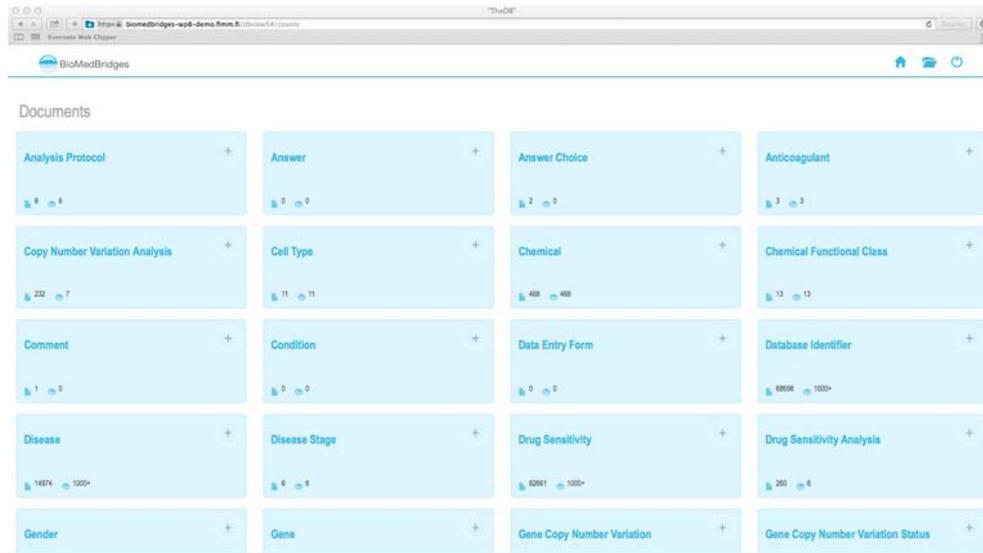
The numbers in the lower left corner of each box indicates the total number of documents and the number of documents visible to this user, respectively. For example, there are 232 Copy Number Variation Analysis documents in the PM system of which 7 are visible to this user. For the sake of speed the latter number is precise only if there are less than 1000 documents. Click the folder icon in the navigation bar to get to this page.

3.4.1 The Persons page

The main purpose of the Persons page is to provide medical doctors treating the patients with an easy way of finding the patient's data. The page lists all the patients whose data the user has the right to see together with samples donated by these patients and analysis performed.

The page enables to limit the data shown to those matching user provided query terms. For example, by typing the identifier of the relevant patient into the search box in the 'study id' column and pressing enter, the page will show only the data from patient(s) with matching study id.

The user can also change the columns displayed. By clicking the navigation bar menu icon > 'Columns', the user can change the order and remove existing columns as well as add new ones. The latter requires certain knowledge of the data model, but is facilitated by an auto-complete functionality.

For example, the user may choose to add columns for person's personal details and thus be able to find the relevant records by name or personal identification number. This is a functionality relevant for medical personnel treating the patient. Of course, this requires the user to have the right to see the personal details documents. Hence, whereas any user can add the columns to his/her view, only those with the permission to see the data will be shown it. That said, in order to be on the safe side, the demo installation of the PM system does not contain any real personal information.

The eye icons on the Persons page provide links into person/patient level summary reports. The summary report provides in reverse cronological order a succinct overview of the most pertinent information for this use case:

- It lists potential driver mutations found in the sample(s). These are known cancer genes as defined by the Cancer Gene Census [http://cancer.sanger.ac.uk/census/] and found mutated in the sample. When gene expression data is available for the sample, also the expression level of the mutated gene is indicated (column 'expression (RPM)'). As the expression value on its own is usually rather meaningless, the report also provides a percentile rank of this gene's expression among all the samples in the PM system that the given user has the right to see.
- Gives graphical overview of the copy number abberrations in the genome. Amplified areas are indicated in read and deleted one in blue with darker colors showing the more severe change.
- Lists the approved drugs and investigational chemicals most effective on this person's tumor sample in ex-vivo drug sensitivity and resistance testing (DSRT) assay [Pemovska 2013]. By default the list is ordered by differential drug sensitivity score (dDSS) values [Yadav 2014] which is a measure of drugs effect on given tumor cells compared to the effect on normal cells. To cater for the need of the physicians to see the effect of

given drug in other samples also the percentile rank is given. The grey bar in the background indicates the relative size of the value compared to minimal and maximal values seen in all the samples with the given drug. The user can also see the distribution of all the values of the drug accross all the samples by clicking on the bar chart icon.
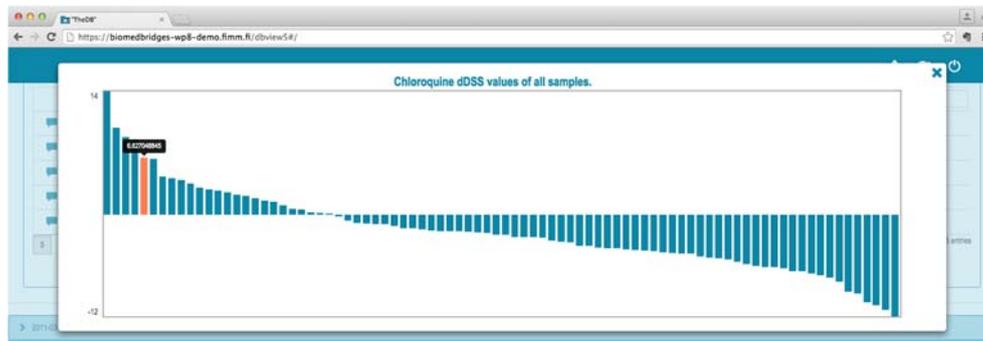
The user can filter further the values in mutations and drug effect tables. For example, by entering the word 'approved' and clicking enter in the search box of the 'development phase' column of the 'DSRT results' section, the table will list the results from approved drugs only.

The report is created anew every time it is requested. Hence a new sample or new results become availabel once added to the PM system. If a new gene gets annotated as a cancer gene, and it is mutated in this person's sample, it will show up in the list of potential drivers. Similarly, if an investigational chemical gets annotated as an approved drug, its classification in the drug testing results table is also changed.

3.4.2 The Samples page

The Samples page lists sample and donor identifiers, diagnosis and disease

stage associated with the sample, sampling time, tissue of the sample as well as analysis performed on the sample and location of the stored aliquots. It serves several purposes:

- Provides users with and entry point for all the information on a sample. If the user knows the sample id or donor id, he/she can quickly limit the content of the page to these samples only. As on the Persons page, click on the eye icon will take the user to the succinct summary of the results otained from this sample.
- Enables users to identify samples matching certain criteria, e.g. representing particular disease, disease stage and tissue, having aliquots in storage and specific analysis done.
- Provides user with the summary information on the set of samples matching user-provided criteriae. For example, the '# of distinct values' section on Figure Samples indicates that in this set of 14 samples originated from three donors, and contains three diagnoses, three disease stages and two tissue types.

3.4.3 The DSRT results, Somatic mutations, Copy number ideograms, Gene copy number alterations, Integrated gene reports pages

All of these pages provide a means of browsing, searching/filtering as well as viewing specific data types accross all the samples and are aimed at researchers.

For example, the screenshots above display the drug sensitivity analysis results for sunitinib accross all relapse stage AML samples and for all

approved kinase inhibitors across diagnosis stage AML samples. In both cases the results are in the descending order of dDSS and only those with dose response curve available are displayed. Once the user has identified a set of interesting samples, he/she can then investigate the mutation status and gene expression of these samples in order to, say, find a mechanistic explanation for the observed phenotype. For brevity, only the top 5 items are displayed.

3.4.5 The Sample storage page

The page is primarily aimed at lab personnel handling the tissue samples sent to the lab for analysis. Clicking the navigation bar menu icon > 'Edit' takes the user to the editing mode enabling to create, edit and delete the entries in spreadsheet-like user interface.

In fact, the editing capability is available on every data page described above. As with sharing, any user who can see the document can also attempt to edit it. However, the changes are stored in the database only if the user has write permission on the document.

The page allows to search for stored sample aliquots and provides a starting point for browsing around in various storage containers, e.g. freezers.
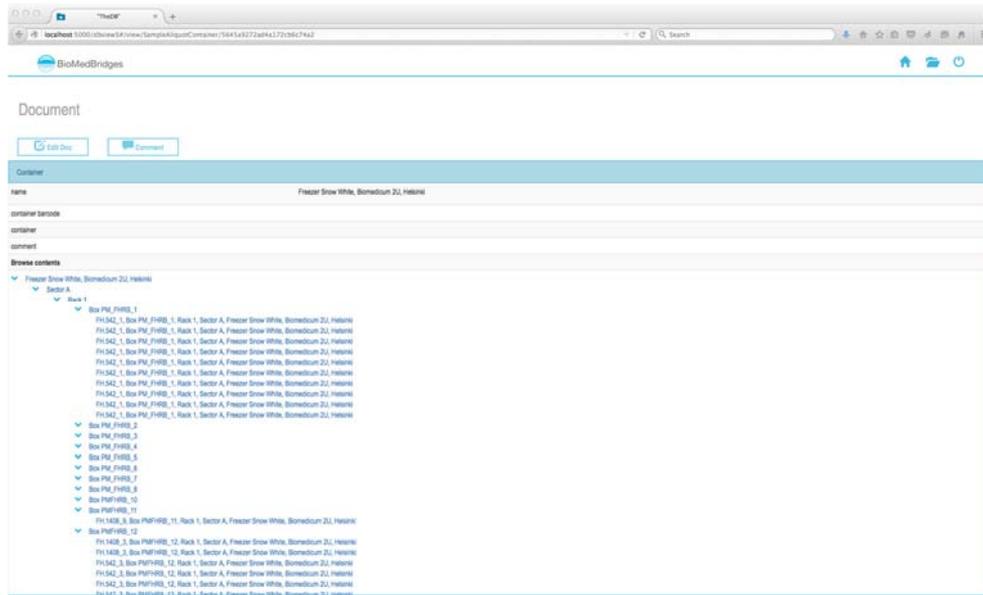
3.4.6 Data sharing

All the pages described above can also be used to change the access permissions of the data. Clicking the navigation bar menu icon > 'Share' brings up a window which allows to select the users and groups whose access rights the user wants to change. In the case of sharing the Person documents, the user can also specify, which other data derived from the person(s) he/she wants to share. Currently the options are samples (entailing the diagnosis, time and tissue information), DSRT results and sequencing results. Whereas any user can attempt to change the access rights, the operation will succeed only if the user has the write permissions on the documents he/she tries to share.
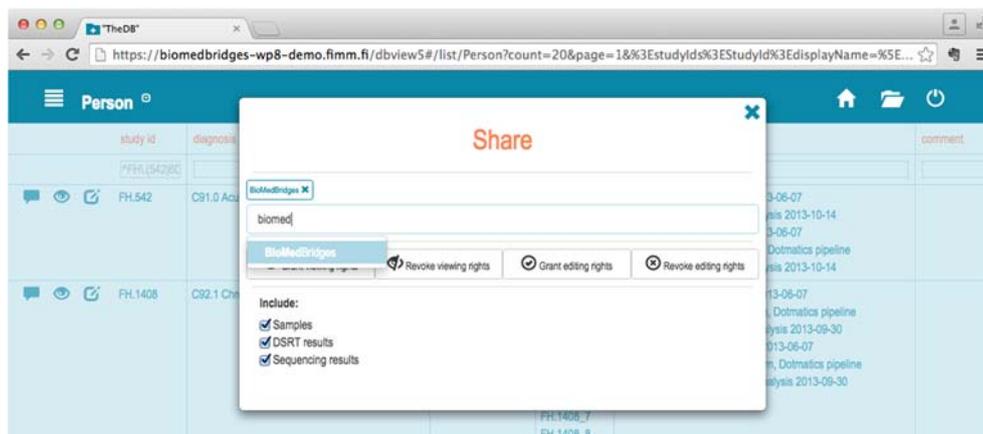
3.4.7 A platform for chatter

Sometimes the data provokes thoughts and opinions and there is need to comment it. At best this can be an important tacit knowledge which

unfortunately ends up in disparate places such as e-mails or private notes. In the PM system any user can comment on any document he/she can see. The user can also specify the visibility of his/her comment. By default, the comment will have the same visibility of the document being commented, but the user can over-ride it in case it is meant for a narrower audience.

3.5 Walk-through 2 – case acute metabolic disease in new-borns

The walkthrough has been part of the BioMedBridges workshop "Translational research infrastructure: submit, use, manage and combine imaging, clinical and molecular data" at the BioMedBridges symposium 2015 and has been implemented using MOLGENIS (http://molgenis.org). A public demo server is available at http://www.molgenis.org/demo. Contact Morris Swertz for access to the servers that where used as part of this walkthrough (m.a.swertz@rug.nl)

Aim of the workshop is to start with a next-generation sequencing genetic variation data file (VCF) with data from 3 patients, to wish to find out:

- which of these patients has a pathogenic mutation
- what disease/condition this patient might have.

For this we use the MOLGENIS annotation system to add data from various sources to answer these questions.

3.5.1 Introduction to the tools used

Cadd

CADD is a tool for scoring the deleteriousness of single nucleotide variants as well as insertion/deletions variants in the human genome.

While many variant annotation and scoring tools are around, most annotations tend to exploit a single information type (e.g. conservation) and/or are restricted in scope (e.g. to missense changes). Thus, a broadly applicable metric that objectively weights and integrates diverse information is needed. Combined Annotation Dependent Depletion (CADD) is a framework that integrates multiple annotations into one metric by contrasting variants that survived natural selection with simulated mutations.

CADD-scores strongly correlate with allelic diversity, pathogenicity of both coding and noncoding variants, and experimentally measured regulatory effects, and also highly rank causal variants within individual genome sequences. Finally, C-scores of complex trait-associated variants from genome-wide association studies (GWAS) are significantly higher than matched controls and correlate with study sample size, likely reflecting the increased accuracy of larger GWAS.

CADD can quantitatively prioritize functional, deleterious, and disease causal variants across a wide range of functional categories, effect sizes and genetic architectures and can be used prioritize causal variation in both research and clinical settings.

(Source: http://cadd.gs.washington.edu)

CGD

A key barrier to translating the power of genomic sequencing to the clinical setting involves the time and resources required for clinically-relevant analysis. To help address this barrier, we constructed the Clinical Genomic Database (CGD), a manually curated database of conditions with known genetic causes, focusing on medically significant genetic data with available interventions.

All conditions with identified genetic causes are included in the CGD. For each entry, the database includes the gene symbol, condition(s), allelic conditions, inheritance, age (pediatric or adult) in which interventions are indicated, clinical categorization, and a general description of interventions/rationale. The contents are not intended to serve as nor substitute for comprehensive clinical guidelines, but are rather intended to briefly describe the types of interventions that might be considered.

The database includes only single gene alterations (it does not include contiguous gene syndromes, although some conditions with, for example, digenic inheritance are included), and does not include genetic associations or susceptibility factors related to more complex diseases, such as identified through association-based studies. Somatic alterations, such as commonly occur in cancerous processes, are not included unless a germline change in the same gene results in disease.

(Source: http://research.nhgri.nih.gov/CGD/)

SnpEff

SnpEff is a variant annotation and effect prediction tool. It annotates and predicts the effects of variants on genes (such as amino acid changes). The inputs are predicted variants (SNPs, insertions, deletions and MNPs). The input file is usually obtained as a result of a sequencing experiment, and it is usually in variant call format (VCF). SnpEff analyzes the input variants. It annotates the variants and calculates the effects they produce on known genes (e.g. amino acid changes).
(Source: http://snpeff.sourceforge.net)

3.5.2 Sign in

Sign in to the workshop VM with username 'workshop' and password 'workshop' (without quotes). Alternatively use the latest version at http://molgenis.org/demo
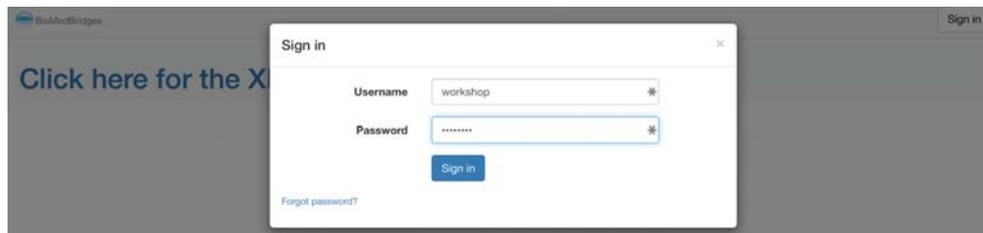
3.5.3 View data in data explorer

- Click "Data Explorer" in the main menu
- Select BMB_Workshop in the entity select dropdown at the right side of the screen
    - This data is directly imported from this VCF File
- Browse the data

There might be an error message in the genome browser, this occurs if reference tracks in the browser could not be loaded due to slow or unstable internet connections

3.5.4 Annotate using SnpEFF

We start with adding the information from SnpEff. This information can help us with the first step of determining the potentially dangerous mutations. For more information on SnpEff read the "background" section at the end of this document.

- Click the "Annotators" tab in the data explorer
    - You'll notice that not all sources are available since the required input for these sources is not present in our dataset.
- Tick the checkbox in front of SnpEff
- Click "Annotate"
    - This might take a minute.
- When finished, click "show result" in the green message bar

3.5.5 View and filter the data

- Select and open the snpEff folder by ticking the checkbox and arrow in front of the folder in the tree on the left side of the screen.
- Click the filter icon in front of "Putative_impact"
- Set filter on impact HIGH
- This gives a first indication of potentially dangerous mutations, however we'd like to filter the set further and for that we need some extra information added.

There might be an error message in the genome browser, this occurs if reference tracks in the browser could not be loaded due to slow or unstable internet connections

3.5.6 Annotate using CADD

We like to add some more data in order to answer our questions.
- Click on the "Annotators" tab
  - You'll notice that all sources are now available since the required input for the previously unavailable sources was added by annotating with SnpEff.
- Tick the box for
  - cadd -> to get the CADD scores, a method to predict pathogenicity

    For more information on these sources read the "background" section at the end of this document.
- Click "Annotate"
  - This will take a few minutes...
- When finished, click "show result" in the green message bar

3.5.7 View and filter the CADD results

By selecting the relevant columns, filtering and sorting we can now find at least 1 variant in our patients that is likely pathogenic, we also have a description of the disease our patients might suffer from.

BioMedBridges Deliverable D8.3

Note that there is a scrollbar at the bottom of the table (you might need to scroll down to show it), you might need to use this scrollbar depending on your screen size adn resolution.

- Select and open the snpEff folder by ticking the checkbox and arrow in front of the folder in the tree on the left side of the screen.
- Click the filter icon in front of "Putative_impact"
- Set filter on impact HIGH
- Add a filter for GT = 1 :
  - collapse the "Samples" folder in the tree
  - click the filter icon in front of GT.
  - type "1" in the textbox and hit "enter"
    - The GT (=genotype) column indicates if the mutation was found in none (0|0) one (1|0 or 0|1) or both (1|1) allele's. By filtering for "1" we exclude the lines where the variant was not seen in any of the patients.
- Select the relevant columns using the tree on the left side of the screen:
  - POS
  - REF
  - ALT
  - SAMPLES
  - cadd> CADDSCALED
  - SnpEff-> Annotation
- Sort on cadd/CADDSCALED score by clicking on the column header
- Collapse the samples by clicking the icon in front of "SAMPLES" in the table header.
- Based on the information on your screen, 15 rows remain.
  - CADDSCALED scores indicate higher pathogenicity for higher scores.
- If we take a look at the "samples -> GT" column, we notice that one of our patients has the variant in one of the allele's. To check if the variant is pathogenic in this situation we need more information, we'll add that in the next step.

3.5.6 Annotate using CGD

We like to add some more data in order to get an answer on the question if one (Heterozygous) or both (Homozygous) alleles should be affected for this variant to be pathogenic. And off course we'd like to know the effect of the variant on the patient. For this information we'll annotate with CGD.

- Tick the box for
  - CGD -> to get information about inheritance and associated conditions
  - Click "Annotate"
  - This will take a few minutes...

- When finished, click "show result" in the green message bar

3.5.9 View, sort and filter the data

By selecting the relevant columns, filtering and sorting we can now find at least 1 variant in our patients that is likely pathogenic, we also have a description of the disease our patients might suffer from.
Note that there is a scrollbar at the bottom of the table (you might need to scroll down to show it), you might need to use this scrollbar depending on your screen size adn resolution.

- Select and open the snpEff folder by ticking the checkbox and arrow in front of the folder in the tree on the left side of the screen.
- Click the filter icon in front of "Putative_impact"
- Set filter on impact HIGH
- Add a filter for GT = 1 :
  - collapse the "Samples" folder in the tree
  - click the filter icon in front of GT.
  - type "1" in the textbox and hit "enter"
    - The GT (=genotype) column indicates if the mutation was found in none (0|0) one (1|0 or 0|1) or both (1|1) allele's. By filtering for "1" we exclude the lines where the variant was not seen in any of the patients.
- Select the relevant columns using the tree on the left side of the screen:
  - POS
  - REF
  - ALT
  - SAMPLES
  - cadd> CADDSCALED
  - SnpEff-> Annotation
  - CGD-> CGDGIN, CGDCOND
- Sort on cadd/CADDSCALED score by clicking on the column header
- Collapse the samples by clicking the icon in front of "SAMPLES" in the table header.
- Based on the information on your screen, 15 rows, you can now easily pick out the patients with highly pathogenic mutations:
  - In case of one affected allele (0|1 or 1|0 in the GT field) only the DOMINANT information is relevant (shown in the CGDGIN (CGD generalized inheritance) column). If both alleles are affected Dominant as well as Recessive information is relevant.
  - CADDSCALED scores indicate higher pathogenicity for higher scores.
  - The CGDCOND (CGD Condition) column shows the condition associated with the mutation.

BioMedBridges Deliverable D8.3

Select the "ID" column for the line with the highest ClinVar and CADD scores, if all went well this line should contain a link to the clinvar page about this variant.

3.6 Current usage of the PM systems

The system described in walk-through 1 is currently hosted at UH/FIMM and is used to store and provide controlled access to the data produced from the samples sent to UH/FIMM for analysis as part of the pesonalised medicine research program. Current user base consists mainly of UH/FIMM researchers and hematologists from the Helsinki University Hospital. However, we are also working towards making this a platform for delivering the analysis results to users elsewhere, e.g. clinicians in other hospitals.

At the time of writing this report the PM system contains molecular profiling data from 323 samples originating from 234 people representing 12 diseases and 10 tissue types.

TissueSample

	sample id	donor	diagnosis	disease stage	tissue	analyses performed
		^(GMALL)				*
💬 👁 ✎	FH.4098_2	FH.4098	C91.1 Chronic lymphocytic leukaemia	Relapse	C42.0 Blood	FH.4098_2 Exo-CNV analysis 2015-01-20
💬 👁 ✎	FH.393_6	FH.393	C92.0 Acute myeloid leukaemia [AML]	Diagnosis	C42.1 Bone marrow	FH.393_6 Exo-CNV analysis 2015-01-27
💬 👁 ✎	FH.393_3	FH.393	C92.0 Acute myeloid leukaemia [AML]	Relapse	CD34+ cells	FH.393_3+ Exo-CNV analysis 2013-06-07
💬 👁 ✎	FH.393_4	FH.393	C92.0 Acute myeloid leukaemia [AML]	Relapse	CD34- cells	FH.393_4- DSRT, MCM medium, Dotmatics pipeline
💬 👁 ✎	FH.393_1	FH.393	C92.0 Acute myeloid leukaemia [AML]	Relapse	C42.1 Bone marrow	FH.393_1 DSRT, MCM medium, Dotmatics pipeline

# of distinct values

323	234	12	5	10	

The system described in walk-through 2 is currently hosted at UMCG and is used to help diagnose patients as part of the diagnostics practice. At the time of writing 20+ babies have been processed with help of the system and in 2 cases a diagnosis could be reached, which was not possible using the traditional tests.

3.7 Future Work

Whereas the BioMedBrdiges is ending, U/FIMM has secured funding to continue the development of the PM system 1 for another two years. This will entail:

- Inclusion of further data types such as clinical chemistry measurements, treatment and treatment outcome information.

- Broadening the use to other indication areas such as solid tumors and other use cases such as biobank sample and information management.
- Broadening the user base to include clinicians from other hospitals and researchers from other institutions.

The aim is to become self-sustainable with the owners of the data paying for the use of the system. Obviously such a system and in particular the data in it is also an important research infrastructure, which hopefully also atracts support from the sponsors of research infrastructures.

Also PM system 2 has secured funding for 2 more years as part of both BBMRI-NL as well as the foundation for heriditary research. The system will continue to be available as open source via http://github.com/molgenis/molgenis and with documentation at http://molgenis.github.io

3.8 How to access the PM systems

The demo version of PM system 1 is hosted at https://biomedbridges-wp8-demo.fimm.fi. In order to create an account you need to have a Google account and registration token. To request a registration token, please send an e-mail to imre.vastrik@fimm.fi, preferably from your institutional e-mail address, and specify a reason why you need the account.

The demo version of PM system 2 is hosted at http://molgenis.org/demo . If desired a private version can be provided for evaluation by emailing m.a.swertz@rug.nl. Documentation how to install locally is provided via http://molgenis.github.io

4 Background Information

This deliverable relates to WP8 Use case: Personalized Medicine. Background information on this WP as originally indicated in the description of work (DoW) is included below. **However, during the course of the projects, the list of work package particants was extended to include also UMCG (participant #19) with funding from the part of the budget originally allocated to UH.** Obviously this is not reflected in the original DoW.

WP8   Title: Use case: Personalized Medicine
      Lead: 16: UH
      Participants: EMBL, KI, UDUS, TUM-MED, UH

WP8 will integrate complex data sets to understand disease pathogenesis and improve biomarker and treatment selection

Work package number	WP8	Start date or starting event	month 1
Work package title	Personalised Medicine		
Activity Type	RTD		

Participant number	1: EMBL	3: KI	5: UDUS	7: TUM-MED	16: UH
Person-months per participant	16	8	5	8	32

Objectives

1. Definition of a process for secure sharing of and access to personalized medicine (PM) data.
2. Definition of existing PM data types and mappings between them.
3. Pilot the use of PM data to support the clinical decision making process.

Description of work and role of participants

Use case: Personalized Medicine - integrating complex data sets to understand disease pathogenesis and improve biomarker and treatment selection

Task 1. Develop a process for secure sharing of and access to PM data
Building on the work carried out in Secure access work package (WP5) we will develop a process by which a producer of the data can share and the user of the data can gain access to the PM data in a secure, legal yet easiest possible manner. FIMM will have the role of a prototype PM data provider as well as a user. TUM, as the leader of WP5 will provide expertise in privacy protection as well as secure sharing and access matters.

Task 2. Define types of PM data and mapping between them
Measurements made with different technologies may not be (and usually are not) directly comparable even though the underlying thing measured (e.g. certain mRNA level) may be exactly the same. This creates a situation where a user of the data may inadvertently be "comparing apples with oranges". To avoid that we will catalogue data types (as well as pertinent standards) relevant to PM and provide mapping between them if applicable. FIMM will provide PM domain expertise. KI, as the leader of the Standards work package (WP3) will provide know-how of existing standards.

Task 3. Develop a PM informatics pipeline
As a proof of concept that the tasks above facilitate the interoperability of different PM data types we will develop a prototype PM informatics pipeline to support the decision making process in PM. This prototype pipeline will utilise the data type specifications and standards established in Task 2 and be subject to constraints of access procedures established in Task 1. FIMM will be a prototype PM data producer and user. EMBL-EBI as the leader of Technical Integration work package (WP4) will provide expertise on general framework and architecture of the implementation.

Deliverables		
Number	Name	Due Month
D8.1	Process specification for secure sharing of and access to PM data	30
D8.2	Definition of PM data types (report)	30
D8.3	Demonstration of interoperability between different types of PM data	48

5 References

Topol EJ. Individualized medicine from prewomb to tomb. Cell. 2014 Mar 27;157(1):241-53. doi: 10.1016/j.cell.2014.02.012. Review. PubMed PMID: 24679539; PubMed Central PMCID: PMC3995127.

Pemovska T, Kontro M, Yadav B, Edgren H, Eldfors S, Szwajda A, Almusa H, Bespalov MM, Ellonen P, Elonen E, Gjertsen BT, Karjalainen R, Kulesskiy E, Lagström S, Lehto A, Lepistö M, Lundán T, Majumder MM, Marti JM, Mattila P, Murumägi A, Mustjoki S, Palva A, Parsons A, Pirttinen T, Rämet ME, Suvela M, Turunen L, Västrik I, Wolf M, Knowles J, Aittokallio T, Heckman CA, Porkka K, Kallioniemi O, Wennerberg K. Individualized systems medicine strategy to tailor treatments for patients with chemorefractory acute myeloid leukemia. Cancer Discov. 2013; 3:1416-1429.

Pemovska, T., Johnson, E., Kontro, M., Repasky, G.A., Chen, J., Wells, P., Cronin, C.N., McTugue, M., Kallioniemi, O., Porkka, K., Murray, B.W., Wennerberg, K.; Axitinib effectively inhibits BCRABL1(T315I) with a distinct binding conformation. 2015, Nature., 519:102-105.

Yadav B, Pemovska T, Szwajda A, Kulesskiy E, Kontro M, Karjalainen R, Majumder MM, Malani D, Murumägi A, Hintsanen P, Knowles J, Porkka K, Heckman C, Kallioniemi O, Wennerberg K, Aittokallio T. Quantitative scoring of differential drug sensitivity for individually optimized anticancer therapies. Scientific Reports 2014; 4, 5193; DOI:10.1038/srep05193