A Demonstration of 4D Digital Avatar Infrastructure for Access of Complete Patient Information

Project acronym: MyHealthAvatar

Deliverable No. 6.6
Final comprehensive datasets

Grant agreement no: 600929

Dissemination Level		
PU	Public	**X**
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

COVER AND CONTROL PAGE OF DOCUMENT	
Project Acronym:	MyHealthAvatar
Project Full Name:	A Demonstration of 4D Digital Avatar Infrastructure for Access of Complete Patient Information
Deliverable No.:	D6.6
Document name:	Final comprehensive datasets
Nature (R, P, D, O)[1]	R
Dissemination Level (PU, PP, RE, CO)[2]	PU
Version:	Final
Actual Submission Date:	14/11/2014
Editor: Institution: E-Mail:	Kostas Marias FORTH kmarias@ics.forth.gr

ABSTRACT:
This deliverable reports deliverable 6.6 – final comprehensive datasets for this project. This document, extents previous submitted deliverable D6.5 and describes the final comprehensive datasets that will be used for experimental purposes by the MyHealthAvatar platform. The final dataset are in line and will be used for the demonstration use cases and scenarios of MHA in order to demonstrate the platform's capabilities. These comprehensive datasets of MHA related to Task 6.7 are: Full-scale datasets of patients with cancer diseases, Multi-scale datasets of patients with osteoarthritis (OA) and/or Bone Marrow Edema (BME) and Datasets of next generation sequencing data. In the following sections, the above datasets are described in detail, focusing on the processes of data collection, storage and access.

[1] **R**=Report, **P**=Prototype, **D**=Demonstrator, **O**=Other

[2] **PU**=Public, **PP**=Restricted to other programme participants (including the Commission Services), **RE**=Restricted to a group specified by the consortium (including the Commission Services), **CO**=Confidential, only for members of the consortium (including the Commission Services)

KEYWORD LIST:
avatar, datasets, imaging, mri, dicom, pacs, genomic, sequence, genes

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 600929.

MODIFICATION CONTROL			
Version	**Date**	**Status**	**Author**
1.0	10/10/2014	Draft	E. Maniadi
1.1	04/11/2014	Draft	E. Maniadi, G. Manikis, G. Papagiannakis, A. Karantanas
1.2	12/11/2014	Draft	E. Maniadi, E. Spanakis
2.0	13/11/2014	Draft	K. Marias

List of contributors

- Evangelia Maniadi, ICS-FORTH
- George Manikis, ICS-FORTH
- Katantanas Apostolos, ICS-FORTH, University of Crete, Heraklion University Hospital
- George Papagiannakis, IMBB-FORTH
- Dimitris Kafetzopoulos, IMBB-FORTH
- Emmanouil Spanakis, ICS-FORTH
- Konstantinos Marias, ICS-FORTH

Contents

CONTENTS 4
1 INTRODUCTION 5
2 MR IMAGING DATASETS 6
2.1 COLLECTION 6
2.1.1 HR-MRI *6*
2.1.2 MRI-relaxometry *6*
2.1.3 DWI - MRI *7*
2.1.4 DCE - MRI *7*
2.2 STORAGE AND RETRIEVAL 8
2.2.1 DICOM standard *9*
2.2.2 DICOM Objects *10*
2.2.3 PACS server Implementation *14*
3 NEXT GENERATION SEQUENCING (NGS) DATA WORKFLOW 16
3.1 OVERVIEW 16
3.2 SEQUENCING STEPS 16
3.3 DATA ANALYSIS 17
3.4 STORAGE 17
3.5 ION TORRENT SEMICONDUCTOR SEQUENCING OVERVIEW 17
3.6 DATA FLOW 18
3.7 DATA FORMATS 19
3.8 SEQUENCE GENERATION 20
3.8.1 Sequence alignment *21*
3.8.2 Variant Calling (Variant Caller plugin) *21*
3.8.3 Annotation of variant calls (Ion Reporter) *22*
3.8.4 Filtering Variants in Ion Reporter *23*
3.8.5 Output files *24*
3.9 DIABETES MELLITUS 28
3.10 CONGESTIVE HEART FAILURE 31
3.11 OSTEOARTHRITIS 33
5 CONCLUSION 38
6 REFERENCES 39
APPENDIX 1 – ABBREVIATIONS AND ACRONYMS 40

1 Introduction

WP6 has as main objective to develop data collection utilities and create a collected data repository to store health related data of individual citizens. In order to align the activities of T6.7 with the main project scenarios, WP6 will focus in the collection of a small number of multiscale datasets. This will be handled by FORTH, though experienced clinicians and molecular biologists that are formal collaborators of the computational medicine laboratory. Although an exhaustive data collection is out of the scope of the project, a small but adequate for experimentation number of full scale and comprehensive datasets (images) will be collected focusing on:

- A multiscale medical imaging dataset of Osteoarthritis patients will be collected prospectively
- A small dataset of next generation sequencing data will be acquired for MHA

This follows the first year recommendations of the project to focus more on scenarios that can really show the full potential of the project and have the greater impact. In the following sections we report the preparation for this data collection first for the medical imaging (section 1) and then for the next generation sequencing (section 2). The data will be collected dynamically during the course of the project and in line with the corresponding needs and technological advancements.

This document, extents previous submitted deliverable D6.5 and describes the final comprehensive datasets that will be used for experimental purposes by the MyHealthAvatar platform. The final dataset are in line and will be used for the demonstration use cases and scenarios of MHA in order to demonstrate the platform's capabilities. These comprehensive datasets of MHA related to Task 6.7 are:

- Full-scale datasets of patients with cancer diseases
- Multi-scale datasets of patients with osteoarthritis (OA) and/or Bone Marrow Edema (BME)
- Datasets of next generation sequencing data

In the following sections, the above datasets are described in more detail, focusing on the processes of data collection, storage and access.

2 MR Imaging datasets

An adequate number of full scale and comprehensive imaging datasets are collected for facilitating R&D purposes of the MHA platform. These datasets cover a range of cancer diseases as well as osteoarthritis disease.

2.1 Collection

All MRI examinations will be performed on a 1.5T clinical MRI system form the University Hospital of Heraklion (Vision-Sonata, SIEMENS) enforced with a powerful gradient system (Gradient strength: 45 mT/m, Gradient slew rate: 200 mT/m/ms). Regarding the osteoarthritis dataset, a standard quadrature excitation/detection knee coil will be used for signal excitation and subsequent signal detection. In specific cases a special cylindrical loop small surface coil (2.5 cm diameter) will be used for signal detection. This coil will be attached and appropriately immobilized on patient's knee. Quantitative MRI will be focused on (a) High Resolution morphometric (HR-MRI), (b) relaxometric (MRI-relaxometry), (c) water molecular diffusion (DWI-MRI) and (d) Dynamic paramagnetic Contrast Enhanced (DCE-MRI) methodologies which will be applied on all patients.

2.1.1 HR-MRI

HR-MRI techniques will include:

a) T2w based morphometry using a 3D Constructive Interference at Steady State (CISS) multislice sequence with basic MR imaging parameters : (TR/TE/FA : 12ms/5ms/70o, 60 space filling slices, slice thickness : 0,6mm, 80 μm in plane spatial resolution and bandwidth > 100 Hz/pixel.

b) T1w based morphometry using a 3D Volume Interpolated Breath-Hold Examination (VIBE) multislice sequence with basic MR imaging parameters: (TR/TE/FA: 11ms/4ms/15o, 60 space filling slices, slice thickness: 0,6mm, 140 μm in plane spatial resolution and bandwidth > 100 Hz/pixel.

Both sequences provide superb in-plane spatial resolution for the anatomic area of the knee. Special post-processing algorithms (fourier analysis, fractal analysis etc) will be used for the quantitative estimation of trabecular thickness, geometry, architecture and 3D trabecular bone spatial distribution for the specific anatomic area.

2.1.2 MRI-relaxometry

MRI-relaxometry techniques will include:

a) T1 relaxomerty using a 3D variable flip angle multisclice GRE sequence, 3D FLASH (3D-FLASH) with basic MR imaging parameters : (TR/TE : 2,9ms/1,3 ms, 40 space filling slices, slice thickness : 3mm, bandwidth > 500 Hz/pixel and 7 variable flip angles : [5o, 10o, 15o, 20o, 25o, 30o and 60o]).

b) T2 relaxometry using a 2D multiecho spin echo multislice SE sequence, (MESE) with basic MR imaging parameters : (TR/FA: 2000ms/90o, 32 equidistant spin echoes with TE's : [6,7, 13,4, 20,1, 26,8, 33,5, 40,2, 46,9, 53,6, 60,3, 67, 73,7, 80,4, 87,1, 93,8, 100,5, 107,2 , 113,9, 120,6, 127,3, 134, 140,7 , 147,4, 154,1, 160,8, 167,5, 174,2, 180,9, 187,6, 194,3, 201, 207,7 and 214,4]msecs, 5 slices, slice thickness : 3mm and bandwidth > 500 Hz/pixel).

c) T2* relaxometry using a 2D multiecho multislice GRE sequence, (MEGRE) with basic MR imaging parameters : (TR/FA: 200ms/25o, 12 equidistant gradient echoes with TE's : [2,4, 4,8, 7,2, 9,6, 12, 14,4, 16,8, 19,2, 21,6, 24, 26,4 and 28,8)] msecs, 5 slices, slice thickness : 3mm and bandwidth > 500 Hz/pixel).

The spatial distribution of T1, T2 and T2* values will be estimated and visualized by utilizing the appropriate color T1, T2 and T2* parametric map. These parametric maps will be also used for the final estimation of the true spatial distribution of the main magnetic field inhomogeneity (ΔH0) introduced by the presence of trabecular bone structures, via the calculation of a final ΔH0 color parametric map. The final (ΔH0) map will be used in turn for the overall assessment of the skeletal status of the knee at the anatomical area of interest.

2.1.3 DWI - MRI

Diffusion Weighted MRI techniques will include:

a) Diffusion weighted MR imaging based on the utilization of a 2D multi-b multislice Spin Echo – Echo Planar Imaging sequence (SE-EPI) with basic MR imaging parameters : (TR/TE/FA : 1500ms/94ms/90o, 10 b values : [0, 50, 100, 150, 200, 300, 500, 700, 1000 και 1500]sec/mm2, with fat saturation, 5 slices, slice thickness : 3mm and bandwidth > 1000 Hz/pixel. The Apparent Diffusion Coefficient (ADC) will be calculated from a selected set of b values.

With more than two b-values available, the Intravoxel Incoherent Motion (IVIM) model will be also applied to the DW-MRI data. By altering the b-value, the IVIM relies on two distinct water motion phenomena; the fast and the slow motion of the water molecules. From the biological perspective, the fast motion reflects the microvasculature blood flow which is mostly apparent in the low b-value range (0-200 sec/mm^2) and is given by the micro-perfusion parameter D*. From the other hand, the slow motion (200-1500 sec/mm^2) is related to the true water diffusivity and is given from the IVIM model by parameter D.

b) Diffusion weighted MR imaging based on the utilization of a 3D multi-Gradient Moment, multislice Precession-Steady-Imaging-Fast (PSIF) GRE sequence (PSIF-Diffusion) with basic MR imaging parameters : (TR/TE/FA : 27ms/8ms/70o, 7 Gradient Moment values : [0, 25, 50, 100, 150, 200, and 250](mT/m)*msec, without fat saturation, 40 slices, slice thickness : 1mm and bandwidth > 500 Hz/pixel.

The spatial distribution of both the D and D* parameters will be estimated and visualized by pseudo-coloured parametric maps.

2.1.4 DCE - MRI

Dynamic paramagnetic Contrast Enhancement MRI techniques will include:

a) T1 relaxomerty using a 3D variable flip angle multisclice GRE sequence, 3D FLASH (3D-FLASH) with basic MR imaging parameters : (TR/TE : 2,9ms/1,3 ms, 40 space filling slices, slice thickness : 3mm, bandwidth > 500 Hz/pixel and 7 variable flip angles : [5o, 10o, 15o, 20o, 25o, 30o and 60o]).

b) T1w based perfusion dynamic paramagnetic contrast enhanced MR imaging based on the utilization of a 3D Volume Interpolated Breath-Hold Examination (VIBE) multislice dynamic sequence with basic MR imaging parameters : (TR/TE/FA : 5,1ms/2,3ms/10o, 20 space filling

slices, slice thickness : 2,5mm and bandwidth > 300 Hz/pixel. The temporal resolution (a time for a single set of slices) is 8.2 secs. The specific sequence will be repeated 30 times. The first 3 sets of slices will be taken as baseline data. The subsequent 27 sets of slices will be obtained post bolus contrast agent injection and they will be used as the basis for the actual dynamic perfusion analysis based upon the kinetics of the paramagnetic contrast agent. The total sequence time will be 4 mins.

T1 perfusion analysis will be performed utilizing the Tofts basic pharmacokinetic model. Quantitative (K_{trans}, K_{ep}, V_p) and semi quantitative (WIN, WOUT, AUC, TTPK etc) perfusion parameters will be estimated utilizing both sequences (a) and (b).

2.2 Storage and retrieval

Several technologies have to be developed in order to support reliable and safe data storage as well as data communication between the local repositories and the avatar system (imaging data storage/retrieval). Thus, the deployment of a PACS server will provide all the relevant functionalities becomes apparent. PACS are medical imaging archive systems (consisting of necessary hardware and software) built to run digital medical imaging. They comprise:

- Modalities: Digital image acquisition devices, which are typically, computed tomography (CT), ultrasound, nuclear medicine, positron emission tomography (PET), and magnetic resonance imaging (MRI).
- Digital image archives: Where the acquired images are stored.
- Workstations/Clients: Where experts view (“read”) the images.

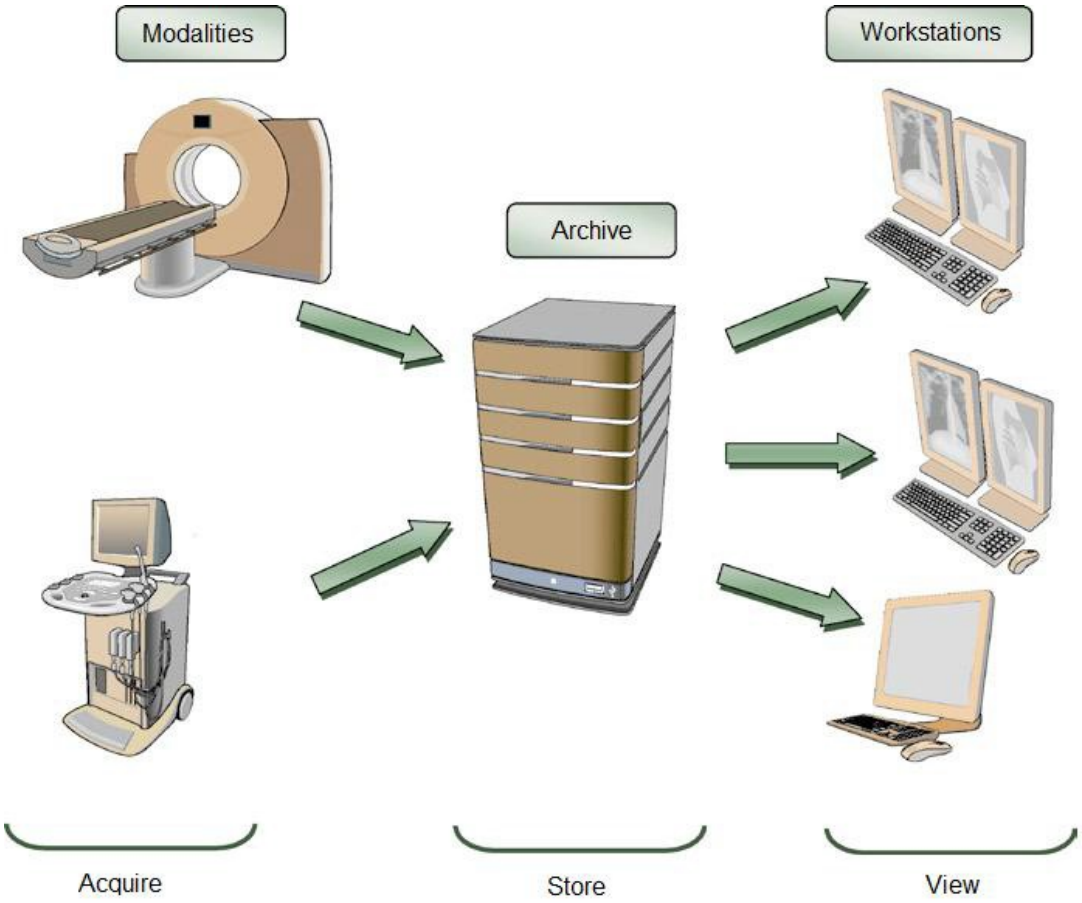


Figure 1: Major PACS components

PACS are directly related to DICOM, which is the reference standard in medical imaging. Their functionality is DICOM-driven, which guarantees their interoperability.

2.2.1 DICOM standard

DICOM stands for Digital Imaging and COmmunications in Medicine and represents years of effort to create the most universal and fundamental standard in digital medical imaging. As such, it provides all the necessary tools for accurate representation and processing of medical imaging data. Moreover, contrary to popular belief, DICOM is not just an image or file format. It is an all-encompassing data transfer, storage and display protocol built and designed to cover all functional aspects of contemporary medicine.

- **DICOM File Format** – All medical images are saved in DICOM format. Medical imaging equipment creates DICOM files. Doctors use DICOM Viewers, computer software applications that can display DICOM images, to diagnose the findings in the images. DICOM files contain more than just images. Every DICOM file holds ***patient information*** (name, ID, sex and birth date), ***important acquisition data*** (e.g., type of equipment used and its settings), *and* ***context of the imaging study*** that is used to link the image to the medical treatment it was part of.
- **DICOM Network Protocol** – All medical imaging applications that are connected to a (hospital) network use the DICOM protocol to exchange information, mainly DICOM images but also patient and procedure information. Moreover, a new imaging equipment in the network can immediately query the medical imaging archive (PACS), retrieve the images that

are created by other systems and display them. Also, the DICOM network protocol is used to search for imaging studies in the archive and restore imaging studies to the workstation in order to display it.

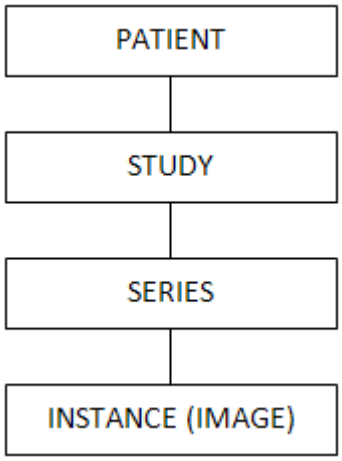
Without a doubt, DICOM truly governs practical digital medicine [1]. The ability of modern imaging equipment to seamlessly collaborate and integrate together in a multi-vendor environment is the most notable achievement of DICOM that led to a great advancement in medical imaging.

One can hardly imagine modern digital medicine without DICOM and PACS. The DICOM standard – conceived over 20 years ago – plays an integral role in the digital medicine evolution, ensuring the highest diagnostic standards and the best performance. DICOM truly shaped the landscape of contemporary medicine by providing [1]:

- *Universal standard of digital medicine*. All current digital image acquisition devices produce DICOM images and communicate through DICOM networks. Modern medical workflow is implicitly controlled by a multitude of DICOM rules.
- *Excellent image quality*. DICOM supports up to 65,536 (16 bits) shades of grey for monochrome image display, thus capturing the slightest nuances in medical imaging. In comparison, converting DICOM images into JPEGs or bitmaps (limited to 256 shades of grey) often renders the images unacceptable for diagnostic reading. DICOM takes advantage of the most current and advanced digital image representation techniques to provide the utmost diagnostic image quality.
- *Full support for numerous image-acquisition parameters and different data types*. Not only does DICOM store the images, it also records a legion of other image-related parameters such as patient 3D position, sizes and orientations, slice thickness, radiation doses and exposures, image processing filters, and so on. This data immensely enriches the informational content of DICOM images, and facilitates processing and interpreting of image data in various ways (for example, creating 3D images from several sequences of 2D CT slices).
- *Complete encoding of medical data*. DICOM files and messages use more than 2000 standardized attributes (defined in the DICOM Data Dictionary) to convey various medical data from patient name to image color depth to current patient diagnosis. Often essential for accurate diagnostics, the data captures all aspects of current radiology.
- *Clarity in describing digital imaging devices and their functionality – the backbone of any medical imaging project*. DICOM defines medical device functionality in very precise and device-independent terms. Working with medical devices through their DICOM interfaces becomes a straightforward, predictable process leaving little room for errors.

2.2.2 DICOM Objects

A DICOM object is comprised of *DICOM elements or DICOM attributes*. Every element is one piece of typed data with a pre-defined, well specified meaning. Every DICOM element has a *Tag* that uniquely defines the element and its properties. There are thousands of DICOM elements from the very basic attributes of patient name and birth date to the most esoteric uses of 3D surface vortices. The specification of DICOM objects are documented in chapter 3 of the DICOM standard [2] that defines the **DICOM data model**. In its most simplified form the DICOM Data Model looks like this:



The DICOM data model defines *Information Entities (IE)*; **Patient**, **Study**, **Series** and **Image**. There are more IE like *Visit, Equipment, Clinical Trial, Procedure* and many others and they are all defined in chapter 3 which is the longest chapter of the standard [2]. A *module* is a collection of elements from one information entity. The DICOM Data Model that is made of IE's is normalized. It is a perfect relational database definition.

All DICOM Objects must include modules from the four main IE: Patient, Study, Series and Image. The mandatory attributes per information entity is presented at the below table. "Type" equals to 1 means mandatory with actual value (not zero length), equals to 2 means mandatory that can be null (zero length) and equals to 3 means optional. Type 1 and 2 can also have a 'C' for conditional so 1C is mandatory if some condition is met and the same for 2C.

Attribute Name	Tag	Type	Attribute Description
PATIENT MODULE ATTRIBUTES			
Patient's Name	(0010,0010)	2	Patient's full name.
Patient ID	(0010,0020)	2	Primary hospital identification number or code for the patient.
Patient's Birth Date	(0010,0030)	2	Birth date of the patient.
Patient's Sex	(0010,0040)	2	Sex of the named patient. Enumerated Values: M = male F = female O = other
GENERAL STUDY MODULE ATTRIBUTES			
Attribute Name	**Tag**	**Type**	**Attribute Description**
Study Instance UID	(0020,000D)	1	Unique identifier for the Study.
Study Date	(0008,0020)	2	Date the Study started.

Study Time	(0008,0030)	2	Time the Study started.
Referring Physician's Name	(0008,0090)	2	Name of the patient's referring physician
Study ID	(0020,0010)	2	User or equipment generated Study identifier.
Accession Number	(0008,0050)	2	A RIS generated number that identifies the order for the Study.
GENERAL SERIES MODULE ATTRIBUTES			
Attribute Name	**Tag**	**Type**	**Attribute Description**
Modality	(0008,0060)	1	Type of equipment that originally acquired the data used to create the images in this Series. See C.7.3.1.1.1 for Defined Terms.
Series Instance UID	(0020,000E)	1	Unique identifier of the Series.
Series Number	(0020,0011)	2	A number that identifies this Series.
Laterality	(0020,0060)	2C	Laterality of (paired) body part examined. Required if the body part examined is a paired structure and Image Laterality (0020,0062) or Frame Laterality (0020,9072) are not sent. Enumerated Values: R = right L = left
SC EQUIPMENT MODULE ATTRIBUTES			
Attribute Name	**Tag**	**Type**	**Attribute Description**
Conversion Type	(0008,0064)	1	Describes the kind of image conversion. Defined Terms : DV = Digitized Video DI = Digital Interface DF = Digitized Film WSD = Workstation SD = Scanned Document SI = Scanned Image DRW = Drawing SYN = Synthetic Image
Modality	(0008,0060)	3	Source equipment for the image. This type definition shall override the definition in the General Series Module. See C.7.3.1.1.1 for Defined Terms.
GENERAL IMAGE MODULE ATTRIBUTES			

Attribute Name	Tag	Type	Attribute Description
Instance Number	(0020,0013)	2	A number that identifies this image. Note: This Attribute was named Image Number in earlier versions of this Standard.
Patient Orientation	(0020,0020)	2C	Patient direction of the rows and columns of the image. Required if image does not require Image Orientation (Patient) (0020,0037) and Image Position (Patient) (0020,0032). May be present otherwise. See C.7.6.1.1.1 for further explanation.
Content Date	(0008,0023)	2C	The date the image pixel data creation started. Required if image is part of a series in which the images are temporally related. Note: This Attribute was formerly known as Image Date.
Content Time	(0008,0033)	2C	The time the image pixel data creation started. Required if image is part of a series in which the images are temporally related.
IMAGE PIXEL MACRO ATTRIBUTES			
Attribute Name	**Tag**	**Type**	**Attribute Description**
Samples per Pixel	(0028,0002)	1	Number of samples (planes) in this image. See C.7.6.3.1.1 for further explanation.
Photometric Interpretation	(0028,0004)	1	Specifies the intended interpretation of the pixel data. See C.7.6.3.1.2 for further explanation.
Rows	(0028,0010)	1	Number of rows in the image.
Columns	(0028,0011)	1	Number of columns in the image
Bits Allocated	(0028,0100)	1	Number of bits allocated for each pixel sample. Each sample shall have the same number of bits allocated. See PS 3.5 for further explanation.
Bits Stored	(0028,0101)	1	Number of bits stored for each pixel sample. Each sample shall have the same number of bits stored. See PS 3.5 for further explanation.
High Bit	(0028,0102)	1	Most significant bit for pixel sample data. Each sample shall have the same high bit. See PS 3.5 for further explanation.
Pixel Representation	(0028,0103)	1	Data representation of the pixel samples. Each sample shall

			have the same pixel representation. Enumerated Values: 0000H = unsigned integer. 0001H = 2's complement
Pixel Data	(7FE0,0010)	1C	A data stream of the pixel samples that comprise the Image. See C.7.6.3.1.4 for further explanation. Required if Pixel Data Provider URL (0028,7FE0) is not present.
Planar Configuration	(0028,0006)	1C	Indicates whether the pixel data are sent color-by-plane or color-by-pixel. Required if Samples per Pixel (0028,0002) has a value greater than 1. See C.7.6.3.1.3 for further explanation.

2.2.3 PACS server Implementation

The PACS that was chosen for the implementation of the current service is the so called DCM4CHEE [3]. DCM4CHEE is a free and open-source DICOM archive and image manager, forming the server side of a PACS system. It is actively developed (in Java) and updated, with modules including HL7 and WADO, and is based on JEE, JMX, and the JBOSS Application Server.

DCM4CHEE offers a web-based interface for the administrators responsible for the administration of the DCM4CHEE archive application. It also uses an internal database to store information from the DICOM headers, index information for locating objects on the file system, and other pertinent system and clinical data. Six different databases are supported for deployment with the system:

- PostgreSQL
- MySQL
- Oracle
- SQL Server
- DB2
- Firebird
- HSQL

Table 1: Mandatory Information included in any DICOM file

For the interaction between the DCM4CHEE and the workstation (avatar system) several DICOM services have to be implemented on both sides. Fortunately, these services have been already implemented by DCM4CHEE for us. More specifically, DCM4CHEE is able to store, query and retrieve any type of DICOM objects by implementing the necessary DICOM services.

The Upload/Storage service allows the upload and storage of DICOM imaging data, identifying and rejecting any non-DICOM elements inside the uploaded file. Upon uploading to PACS, DCM4CHEE indexes and stores the DICOM elements automatically. Data are indexed based on patient, study ID and series number. The actual uploaded data should be provided using a ZIP archive format. This is a universally supported format and allows multiple files and directories to be combined together in a single file.

The Query/Retrieve service takes as an argument a DICOM object that represents a query. The server transforms the object that was sent in SQL, runs it and then transform every result record

back into a DICOM object and send it back to the client. The server sends one response for every result record. While still running, the status field of the response command is pending. The last response has a status success. It may of course fail and then the server will throw an exception with the failure reason and status. It may also succeed with no matches.

As explained above, the imaging data are indexed based on patient, modality, study ID and series number or in combination of these options. By choosing a patient, all the studies and modalities related to that patient are retrieved and subsequently all relevant series from each study can be retrieved. By selecting now a patient and a modality that is available for that patient, the avatar can retrieve a whole study or the relevant/needed DICOM image series from the selected modality.

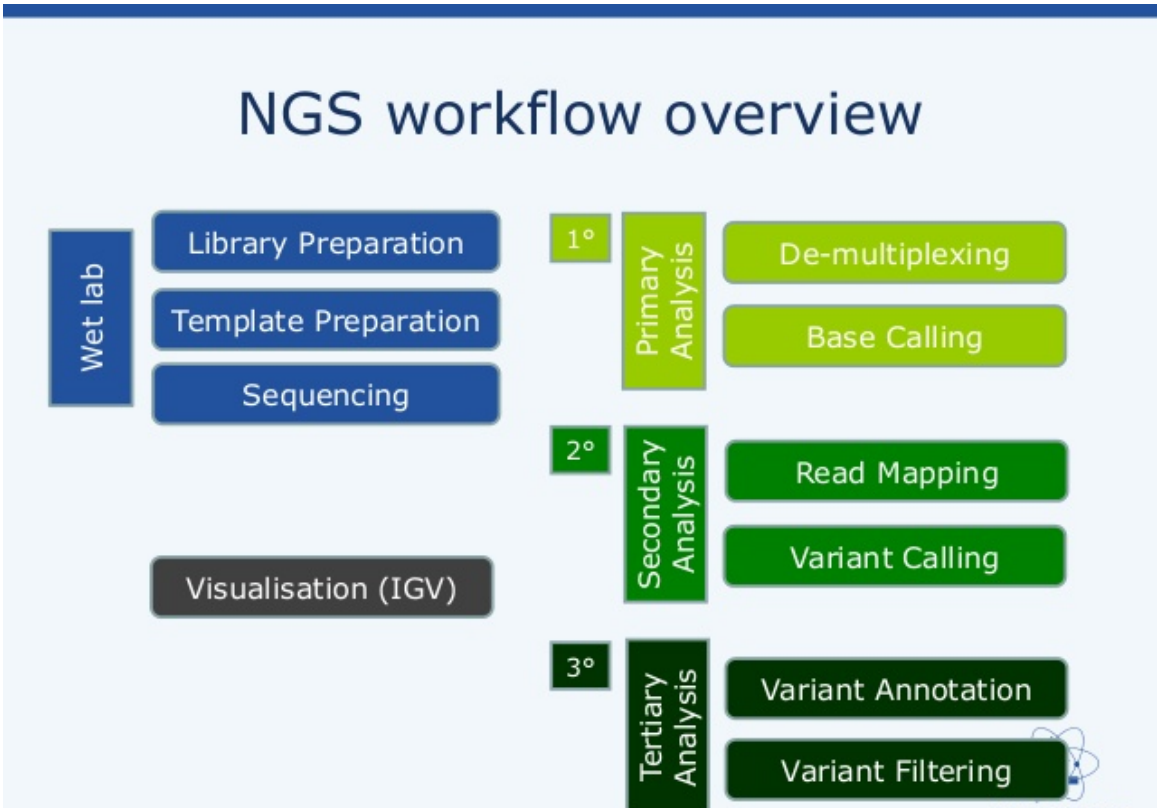
3 Next Generation Sequencing (NGS) Data Workflow

3.1 Overview

Sequencing techniques are used to determine the base composition of DNA/RNA molecules. Advances in sequencing technology have succeeded in parallelizing the sequencing process, a term referred to as “next-generation sequencing”. These new high-throughput techniques have resulted in massive amounts of biological data. NGS applications range from microbial to mammalian genome sequencing, exome sequencing, as well as targeted resequencing, mitochondrial, miRNA and chromatin-immunoprecipitation sequencing.

3.2 Sequencing steps

The usual steps performed during an NGS run are the wet lab preparation, the sequencing run itself, data analysis and filtering, and finally storage. The preparation consists of DNA or RNA sample purification, enrichment, and preparation of the to-be-sequenced library. During the sequencing run, sensors inside the instrument perform base-calling in the sequences they read, and the produced data is exported into one or more files, usually stored in the computer server that accompanies the instrument. NGS output formats can be either text or binary, with the most common text formats being FASTA and FASTQ. The most common binary formats are BAM and SFF. A quality control step is usually first performed on the data, which includes removal of sequences or nucleotides with a low quality threshold, as well as trimming of adapter sequences that have been read together with the targets.



3.3 Data analysis

Subsequent data analysis can be performed by specialized software, which can be either free/open-source or commercial. Some common free and/or open-source software are Galaxy (a suite which includes many different free programs), NovoAlign (software distributed by Illumina for use with Illumina data), and Ion Torrent Suite (distributed by Life Technologies for use with Ion Torrent data). Common commercial software which can be used with output data from most platforms are CLC Genomics Workbench, Partek Genomics Suite, and NextGENe.

3.4 Storage

Due to the volume of produced data and the cost associated with each sequencing run, storage of both raw and analyzed data is usually done in specialized servers with large storage capacity as well as data corruption/loss prevention mechanisms. Such an example is the EBI SRA: The partners of the International Nucleotide Sequencing Database Collaboration, which includes the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI), and the DNA Data Bank of Japan (DDBJ), have established the Sequence Read Archive (SRA) to provide the scientific community with an archival destination for next generation data sets. The SRA captures and presents information relating to experimental workflows that are based around nucleotide sequencing. Data arrive at SRA from a variety of sources. These include submissions of raw data, assembled sequences and annotation from small-scale sequencing efforts, data provision from the major European sequencing centres and routine and comprehensive exchange with partners in the International Nucleotide Sequence Database Collaboration (INSDC).

3.5 Ion Torrent Semiconductor sequencing overview

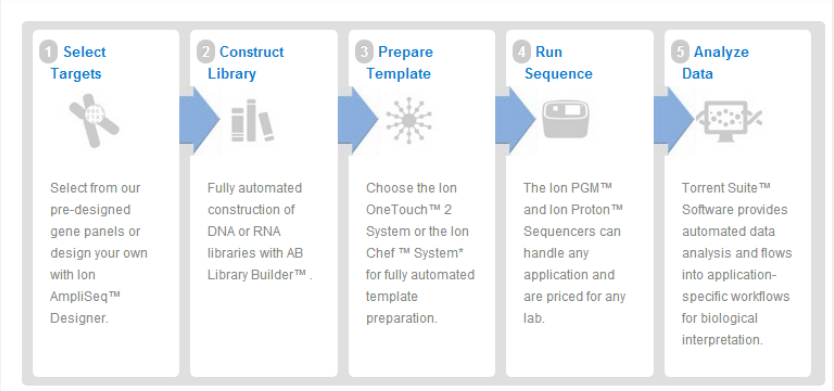
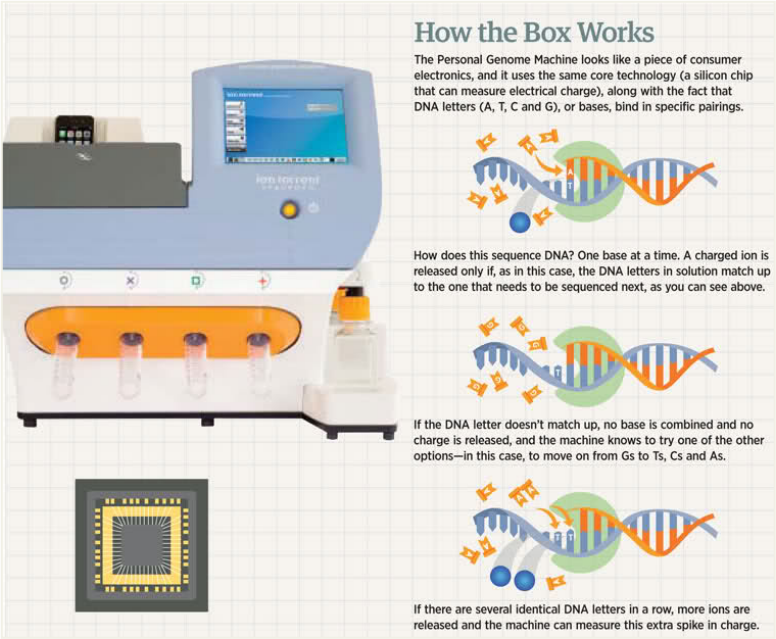


Figure 2 - Ion Torrent sequencing workflow (©Life Technologies)

Ion Semiconductor Sequencing is a method of DNA sequencing based on the detection of hydrogen ions that are released during the polymerization of DNA. This is a method of "sequencing by synthesis", during which a complementary strand is built based on the sequence of a template strand. A microwell containing a template DNA strand to be sequenced is flooded with a single species of deoxyribonucleotide triphosphate (dNTP). If the introduced dNTP is complementary to the leading template nucleotide, it is incorporated into the growing complementary strand. This causes the release of a hydrogen ion that triggers an ISFET ion sensor, which indicates that a reaction has occurred. If homopolymer repeats are present in the template sequence, multiple dNTP molecules will be incorporated in a single cycle. This leads to a corresponding number of released hydrogens and a proportionally higher electronic signal.

Ion semiconductor sequencing chips contain a high-density array of micro-machined wells over an ion-sensitive layer and a proprietary ion sensor. During sequencing, each well contains a different DNA template. The Ion Sequencer sequentially fills the chip with one nucleotide after another. When a nucleotide is incorporated into a strand of DNA, a hydrogen ion is released. The charge from that ion changes the pH of the solution and is detected by the ion sensor. The process transforms the chemical information to digital information in a very direct manner.

After release of the hydrogen ion during nucleotide incorporation, the chemical changes from nucleotide incorporation are now expressed in digital information as changes in voltage. The Torrent software later analyzes the signal data to generate the base sequence of the DNA sample and then to align the sequence to the reference genome.



3.6 Data flow

The Ion sequencer outputs raw sequencing data in the form of DAT files. The raw measurements in DAT files are the conversion of the raw pH value in a well to a digital representation of the voltage. These raw data files are transferred to the Torrent Server for analysis pipeline processing. On the Torrent Server, the Torrent Suite analysis pipeline converts the raw signal measurements into incorporation measurements and, ultimately, into base calls for each read.

The following diagram shows the steps in the Torrent pipeline from the point of view of the data files that are generated. A data file output by one step is input into the next step in the pipeline.

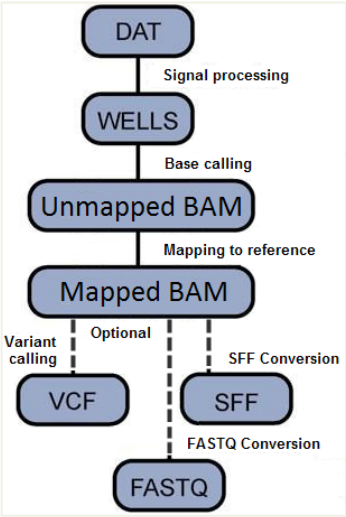


Figure 3 - Generation of data files

3.7 Data Formats

- **DAT** (Data Acquisition) – Contains raw voltage measurements from the chip.
- **WELLS** – Contains the nucleotide incorporation signals for each flow for each well.
- **BAM** (Binary Alignment Map) – Contains unmapped reads or mapped reads with their alignment to the reference genome. Ion Torrent BAM files include flow space information that contains nucleotide sequence and quality scores. The files may contain aligned (based on a reference) or unaligned reads. The compression of this format allows a lot of data to be stored in a file that takes up relatively little hard drive space. Furthermore, the file can also store metadata (e.g. flags for paired reads, QC pass/fail), offering a distinct advantage over FASTQ. http://samtools.sourceforge.net
- **VCF** (Variant Call Format) – Contains variants called on the input DNA sample. Each variant call details how the sequence of bases found in the DNA sample differs from the reference genome, for instance, by an insertion of bases, deletion of bases, or change in a base or in sequential bases. is a text file format (most likely stored in a compressed manner) for storing the most prevalent types of sequence variation, including SNPs, indels and larger structural variants, together with rich annotations. It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome. http://vcftools.sourceforge.net
- **SFF** (Standard Flowgram File) – Contains the nucleotide calls (sequence calls), associated quality values, signal values in flow space, and a mapping between sequence and flow spaces.
- **FASTQ** – Contains the nucleotide calls (sequence calls), and associated quality values. This file contains a sequence header followed by a line of sequence then by a quality header and another line of quality metrics. The sequence header line always starts with a "@" directly

followed by a unique identifier and may contain additional information. The quality header always starts with "+" and may contain additional information. Like the FASTA, there is typically one line for each header, sequence, and quality metrics, but more than one line may appear for sequence and quality metrics. The line(s) containing quality metrics may appear different from each sequencer. Typical quality metrics are character encoded and appear quite complex.

- **BAI** - This is the index of an associated BAM file. Indexing a BAM file allows for quick access of reads, regions (chromosomes) or flagged data without having to read through an entire file. Sorting and indexing a BAM file can significantly reduce downstream computation time for certain processes.

VCF, SFF, and FASTQ files are optionally generated by plugins, not by the main analysis pipeline. The variantCaller plugin (described below) generates VCF output. The FileExporter plugin can generate SFF, FASTQ, or zipped output, or also rename output files.

```
##fileformat=VCFv4.1
##fileDate=20140116
##source=Torrent Unified Variant Caller (Extension of freeBayes)
##reference=/results/referenceLibrary/tmap-f3/hg19/hg19.fasta
##reference=file:///results/referenceLibrary/tmap-f3/hg19/hg19.fasta
##contig=<ID=chr1,length=249250621,assembly=hg19>
##LeftAlignVariants="analysis_type=LeftAlignVariants bypassFlowAlign=true kmer_len=19 min_var_count=5 short_suffix_match=5
min_indel_size=4 max_hp_length=8 min_var_freq=0.15 min_var_score=10.0 relative_strand_bias=0.8 output_mnv=0 sse_hp_size=0
sse_report_file= target_size=1.0 pref_kmer_max=3 pref_kmer_min=0 pref_delta_max=2 pref_delta_min=0 suff_kmer_max=3 suff_kmer_min=0
suff_delta_max=2 suff_delta_min=0 motif_min_ppv=0.2 generate_flow_position=0 input_file=[] read_buffer_size=null phone_home=STANDARD
gatk_key=null read_filter=[] intervals=null excludeIntervals=null interval_set_rule=UNION interval_merging=ALL
reference_sequence=/results/referenceLibrary/tmap-f3/hg19/hg19.fasta rodBind=[] nonDeterministicRandomSeed=false
downsampling_type=BY_SAMPLE downsample_to_fraction=null downsample_to_coverage=1000 baq=OFF baqGapOpenPenalty=40.0
performanceLog=null useOriginalQualities=false BQSR=null defaultBaseQualities=-1 validation_strictness=SILENT
unsafe=null num_threads=1 combined_sample_name= num_cpu_threads=null num_io_threads=null num_bam_file_handles=null read_group_black_list=null
pedigree=[] pedigreeString=[] pedigreeValidationType=STRICT allow_intervals_with_unindexed_bam=false logging_level=INFO log_to_file=null
help=false variant=(RodBinding name=variant source=/results/small_variants.sorted.vcf)
##phasing=none
##INFO=<ID=OID,Number=.,Type=String,Description="List of original Hotspot IDs">
.
.
##FORMAT=<ID=AO,Number=A,Type=Integer,Description="Alternate allele observation count">
.
.
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  Exome
chr1    871334  .   G   T   100.0   .   AO=12;DP=12;FAO=12;FDP=12;FR=.;FRO=0;FSAF=1;FSAR=11;FSRF=0;FSRR=0;FWDB=0.12468;
FXX=0;HRUN=2;LEN=1;MLLD=59.8209;QD=35.938;RBI=0.144227;REFB=0;REVB=0.0725005;RO=0;SAF=1;SAR=11;SRF=0;SRR=0;SSEN=0;SSEP=0;
SSSB=-3.23983E-8;STB=0.5;TYPE=snp;VARB=0.00930441;OID=.;OPOS=871334;OREF=G;OALT=T;OMAPALT=T;;dbsnp=rs4072383
GT:GQ:DP:FDP:RO:FRO:AO:FAO:SAR:SAF:SRF:SRR:FSAR:FSAF:FSRF:FSRR  1/1:5:12:12:0:0:12:12:11:1:0:0:11:1:0:0
```

Figure 4: Example VCF entries

Data format description: http://genome.ucsc.edu/FAQ/FAQformat.html#format3

3.8 Sequence generation

The sequence generation step, also known as base calling, determines the sequence of individual nucleotide bases in the DNA sample. The Torrent sequence generation algorithm, BaseCaller, runs automatically during the Torrent Suite pipeline and is optimized for Torrent data. During sequencing, each well in the sequencing chip contains a DNA template. As shown above in the nucleotide incorporation diagram, the DNA sample has one template (fixed) strand and a synthesized strand. (The base sequence of the template strand is not known before the run. The

bases shown in the nucleotide incorporation diagram are examples of how new bases bond to existing bases in the template strand.) The sample's synthesized strand grows in length whenever a base from a nucleotide flow is incorporated. To be incorporated into the growing strand, the base in the flow must be complementary to the next base on the template's fixed strand. The bases contained in each nucleotide flow's solution are known beforehand.

During each nucleotide flow over the chip, when a nucleotide is incorporated into a strand of DNA, a hydrogen ion is released. The ion release changes the pH of the solution and is detected by the chip's ion sensor. The raw pH value in each well is converted into a voltage and then into a digital representation of that voltage. If the nucleotide that floods the chip is not a match, no voltage change is recorded. Analysis of these data reveals the base incorporated during the nucleotide flow. This process is base calling for one position in the sequence.

The nucleotide release happens in each well of the chip. The millions of wells per chip and hundreds of flows per run make base calling a massively parallel operation. These measurements over the entire chip occur many times per second. The sequencing process also takes advantage of the successive improvements in semiconductor technology to increase throughput.

The Torrent Suite base calling algorithm is optimized for Torrent data and rarely requires tuning for a sequencing run. The base calling module uses fairly stringent filters that are designed to increase the accuracy of results. Base filters may be changed for specific scenarios, for instance, such as when the sequencing application requires maximizing the number of reads.

3.8.1 Sequence alignment

During sequence alignment (also known as mapping), base calls generated by the Torrent Suite analysis (in the BAM file) are aligned to a reference genome, and alignment metrics are also produced. The Torrent Suite includes the Torrent Mapping Alignment Program (TMAP), a sequence alignment software program that is developed specifically to meet Ion Torrent data mapping challenges. Ion Torrent data's particular qualities require special consideration during the alignment process for several reasons: Reads generated by Ion Sequencers are variable in length and are expected to increase over time. The principal error mode associated with Ion data relates to miscalling homopolymer lengths and results in insertion or deletion errors during alignment and post-processing. TMAP implements a two-stage mapping approach to maintain sensitivity and specificity while significantly reducing run time. In two-stage mapping, reads that do not align during the first stage are passed to the second stage with a new set of algorithms and/or parameters. The output of sequence alignment is a BAM file containing mapped reads. The BAM file is analyzed to extract various quality metrics, including quality estimates and read length estimates.

3.8.2 Variant Calling (Variant Caller plugin)

The Torrent Variant Caller (TVC) plugin is a secondary analysis software tool designed to call variants (SNPs, insertions, deletions, MNPs, and block substitutions) and is a component of the Torrent Suite Software package. This plugin accepts as input the aligned read output generated by Torrent Suite Software (a BAM file) and produces as output an annotated list of SNPs and indel variants called in a sample. Optionally TVC can be supplied with target regions files and hotspot files:

- **Target regions files** — Sequencing is restricted to specified chromosome regions that appear in the regions of interest file. The regions of interest file must be a Browser Extensible Data (BED) file, which is a tab-separated file format

- **Hotspot files** — A hotspots file contains a list of positions on the genome. TVC evaluates each listed position on the genome, and reports the filtering metrics for each position,

including positions that are not called as a variant. When a hotspot position receives a NOCALL rather than a reference call or a variant call, the filtering reasons in the VCF output file explain the reasons for the NOCALL. The hotspots file must be in BED format or Variant Call Format (VCF).

```
track name="AmpliSeqExome_Designed" description="Amplicon_Insert_AmpliSeqExome" ionVersion=4.0 type=bedDetail
chr1    68928    69134    OR4F5_1.1.3601    .    GENE_ID=OR4F5;Pool=1
chr1    69212    69428    OR4F5_1.1.11194   .    GENE_ID=OR4F5;Pool=2
chr1    69385    69563    OR4F5_1.2.19824   .    GENE_ID=OR4F5;Pool=3,6
chr1    69556    69740    OR4F5_1.2.28698   .    GENE_ID=OR4F5;Pool=4
chr1    69588    69810    OR4F5_1.2.30842   .    GENE_ID=OR4F5;Pool=5
chr1    861310   861531   SAMD11_8.7663     .    GENE_ID=SAMD11;Pool=6
```

bed file format example

3.8.3 Annotation of variant calls (Ion Reporter)

Ion Reporter Software can perform mapping, variant calling, and annotation, starting with the input of an unaligned BAM file, or perform only variant annotation starting with a variant file in VCF format, produced by the TVC plugin. In addition to the listing of the variant calls and positions, IR also annotates called variants with information from both publicly and private databases.

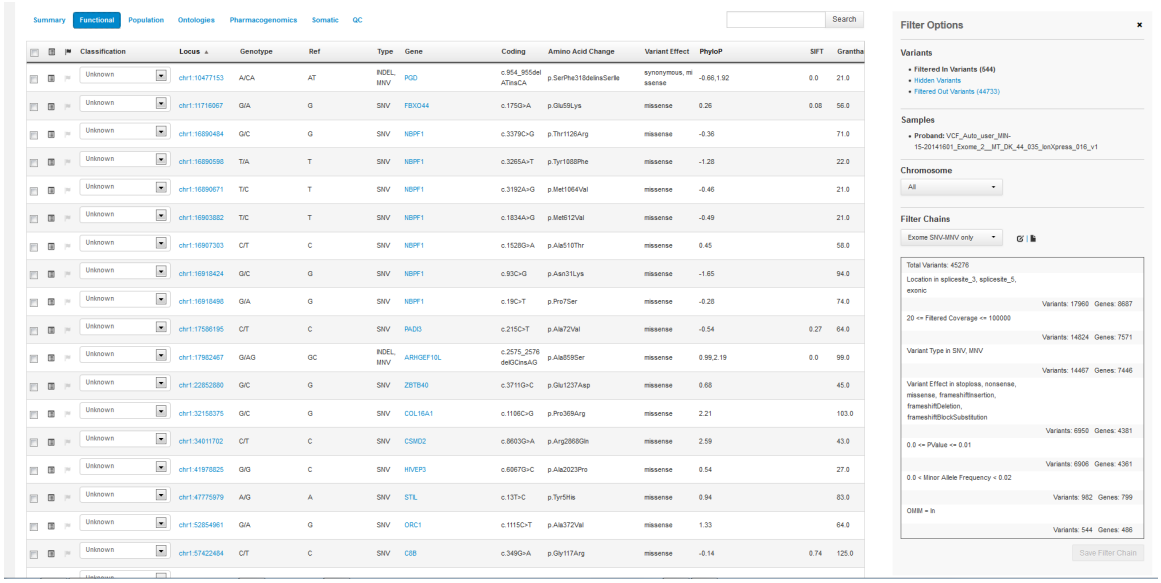


Figure 5: Ion Reporter Interface

The following annotation sources are packaged with IR:

- dbSNP— The Single Nucleotide Polymorphism Database, a free public-domain archive for simple genetic polymorphisms.

- COSMIC— The Catalogue of Somatic Mutations in Cancer, which contains information about somatic mutations in cancer, with more than 100,000 somatic mutations from approximately 400,000 tumors.
- OMIM (Online Mendelian Inheritance in Man®) is a comprehensive, authoritative, and timely compendium of human genes and genetic phenotypes. The full-text, referenced overviews in OMIM contain information on all known Mendelian disorders and over 12,000 genes. OMIM focuses on the relationship between phenotype and genotype.
- **5000 exomes** Population frequency information from the 5000 exomes project
- ClinVar Assessment of impact of the variant observed from NCBI ClinVar database
- DrugBank List of drugs known to target the gene(s) affected by the variant
- MAF Population frequency information from the 1000 genomes project
- **GeneModel** — Ensembl® or RefGene sources.
- **GenePanel** — Genomic regions panels that you download from AmpliSeq.com or your own custom panels.
- **SIFT scores** — A SIFT score predicts whether an amino acid substitution affects protein function.
- **PolyPhen-2 scores** — The PolyPhen-2 score predicts the possible impact of an amino acid substitution on the structure and function of a human protein.
- **Grantham scores** — The Grantham score attempts to predict the distance between two amino acids, in an evolutionary sense.
- **PhyloP** — PhyloP scores measure evolutionary conservation at individual alignment sites and report either slower evolution than expected or faster evolution than expected.
- **Gene Ontology** — The Gene Ontology project aims to standardize the representation of gene and gene product attributes across species and databases by providing a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data.
- **Pfam** —The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models.
-
- **ONCOMINE data** — Professionally curated data that is derived from vast amounts of publicly available cancer data.
- **Ingenuity® Variant Analysis Software** —An external annotation source (commercial), which can be used when a large number of new variants is detected in an analysis. Variant Analysis Software helps filter the large set down to the most interesting variants, based to pathways, literature citations, and other categories. The resulting variant set can then be imported back to IR.

3.8.4 Filtering Variants in Ion Reporter

A key challenge of using genome/exome sequencing to find novel disease genes for either Mendelian or complex traits, is how to identify disease-related alleles among the background of non-pathogenic polymorphism and sequencing errors. On average, exome sequencing identifies ~50,000 single nucleotide variants (SNVs), more than 95% of these variants are already known as polymorphisms in human populations. Strategies for finding causal alleles against this background vary, and various filtering steps and annotation sources are used.

Ion Reporter filtering approaches:

- **Evidence-based filtering:** The purpose of the evidence-based filtering is to separate the list of calls with the greatest evidence, such as :
    - P-value thresholds for the variant calls.
    - The read metric minor allele frequency (as described in dbSNP).
    - Variant types: SNV, indel, MNV, reference calls, NOCALL, CNV, and long deletions
- **Relevance-based filtering:** To filter variants based on context, we can filter by many annotations, including the following:
    - Gene Symbol
    - Gene Location such as intronic, exonic, utr, splicesite, etc.
    - Gene Function such as in unknown, missense, synonymous, etc.
    - Functional Prediction scores such as SIFT, PolyPhen, and Grantham
    - External annotation Sources such as dbSNP, OMIM, COSMIC, ClinVar, Gene Ontology etc
    - Hotspots
    - Variant type

The final steps of variant filtering involve manual curation, requiring expertise, knowledge of the patient's phenotype and family history, current knowledge and exploration of published variant-disease associations, and judgment, all of which can contribute to variability in the final variant interpretation. Relevance to disease-related risk will require comparison to a well-curated catalogue of variants that are known to influence risk of disease. General databases, such as the Human Gene Mutation Database (HGMD), or locus-specific databases like LOVD and pharmacogenomics databases like PharmGkb are commonly used.

3.8.5 Output files

A VCF file of detected variants, or a TSV, which is a tab-separated file of the variants. This file contains the variants' coordinates, annotation sources, and attribute-value pairs of annotation information.

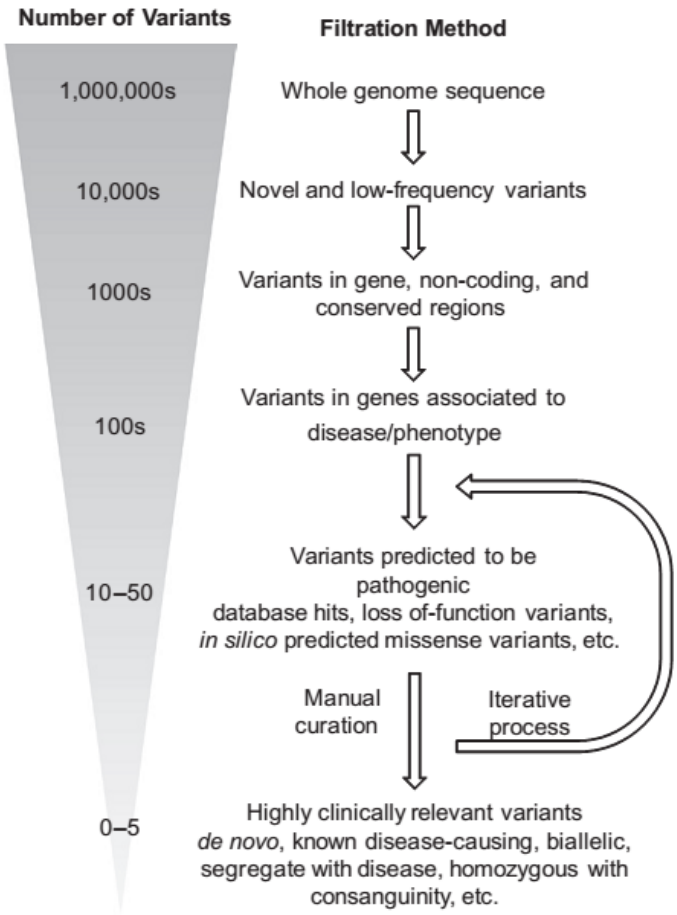


Figure 6: Variant analysis and filtration in whole genome sequencing

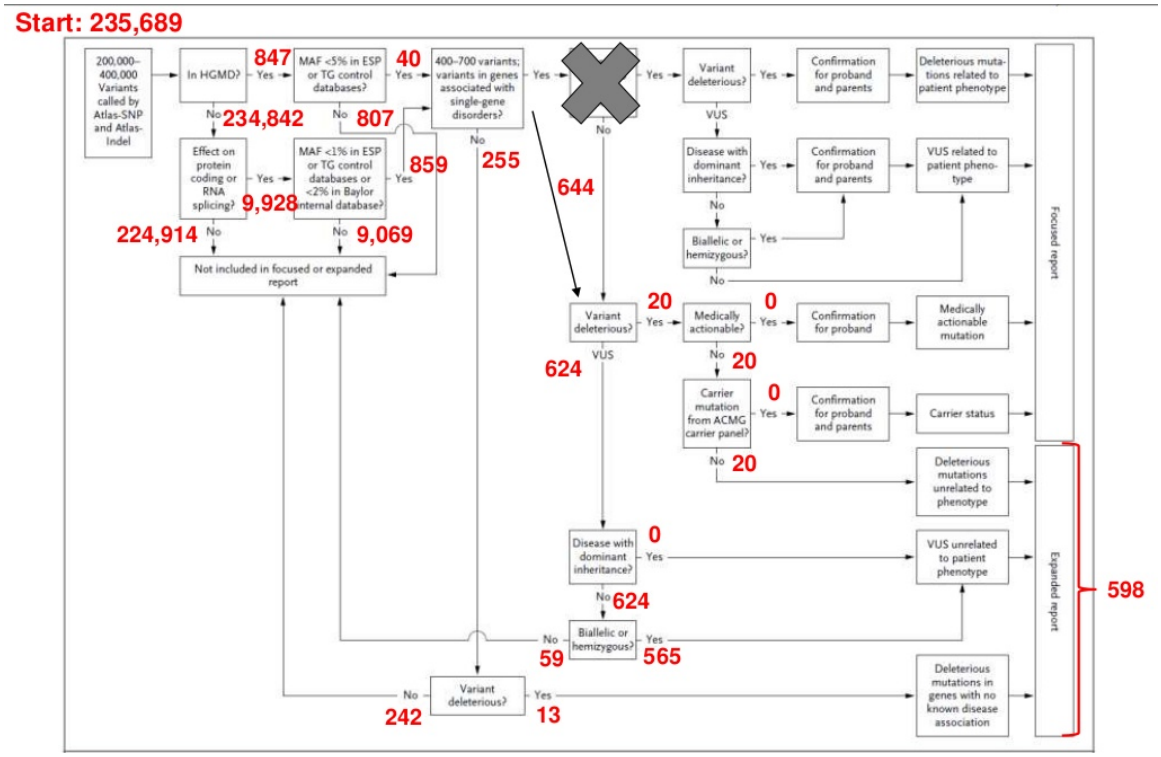


Figure 7: Exome Analysis example workflow

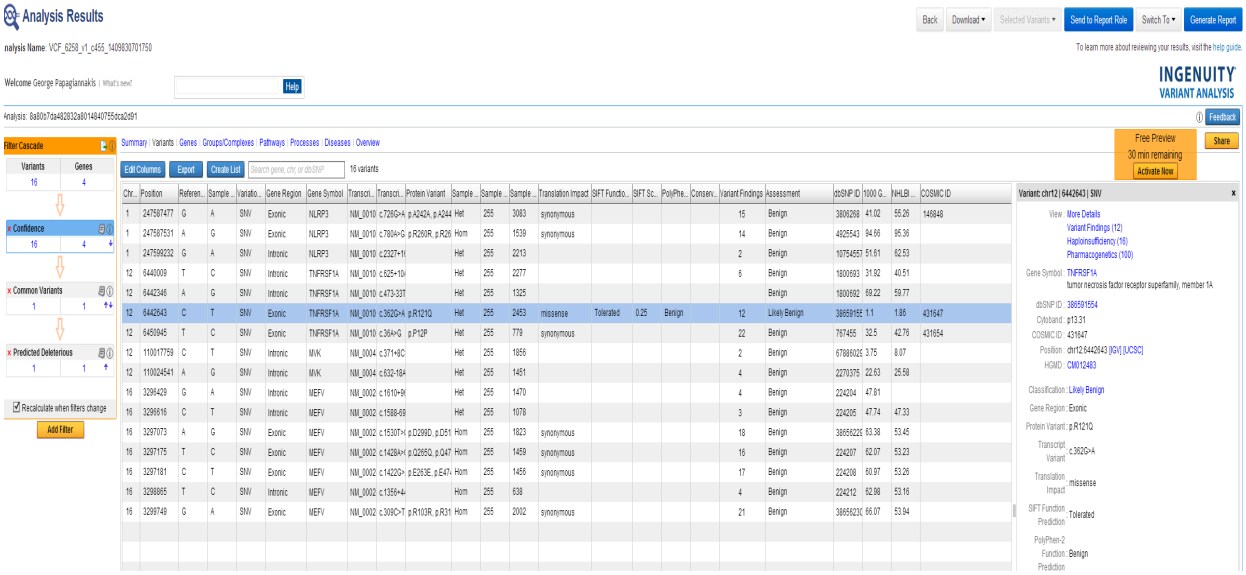


Figure 8: Ingenuity variant analysis example

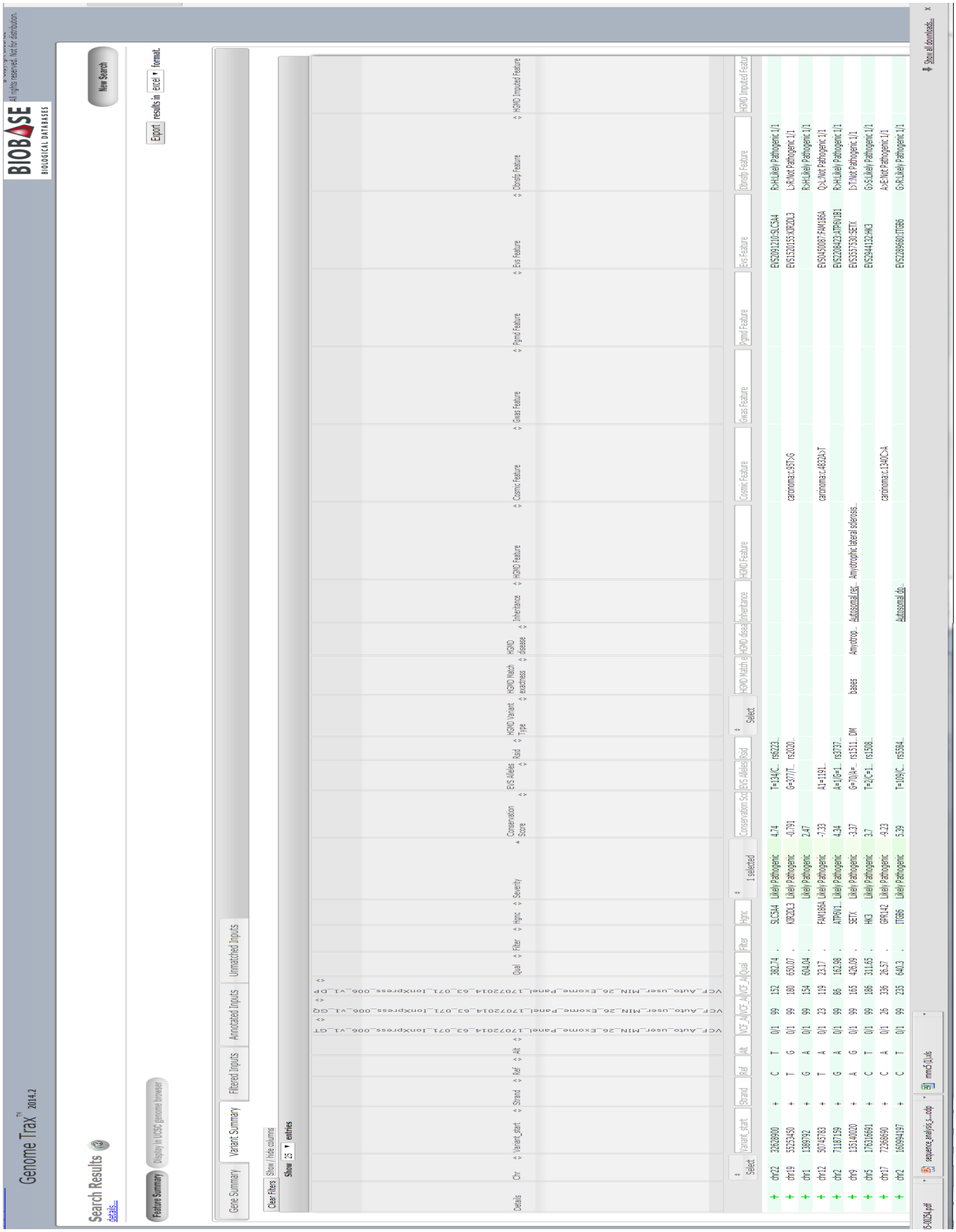


Figure 9: Variant annotation and filtering in HGMD

3.9 Diabetes Mellitus

id	Gene	Name	Type	Significance	SNP ID	Assembly	Location
1	HNF1B	NM_000458.2(HNF1B): c.1395C> G (p.Ser465Arg)	single nucleotide variant	Pathogenic	rs121918673	GRCh37	Chr 17, 36061127: 36061127
2	PAX4	NM_006193.2(PAX4): c.361C> T (p.Arg121Trp)	single nucleotide variant	Pathogenic	rs114202595	GRCh37	Chr 7, 127254587: 127254587
3	AKT2	NM_001626.5(AKT2): c.821G> A (p.Arg274His)	single nucleotide variant	Pathogenic	rs121434593	GRCh37	Chr 19, 40743886: 40743886
4	INSR	NM_000208.2(INSR): c.3572G> A (p.Arg1191Gln)	single nucleotide variant	Pathogenic	rs121913150	GRCh37	Chr 19, 7120718: 7120718
5	INSR	NM_000208.2(INSR): c.3034G> A (p.Val1012Met)	single nucleotide variant	Likely pathogenic, Pathogenic	rs1799816	GRCh37	Chr 19, 7125518: 7125518
6	GPD2	NM_001083112.2(GPD2): c.1904T> C (p.Phe635Ser)	single nucleotide variant	Pathogenic	rs121918407	GRCh37	Chr 2, 157435621: 157435621
7	SLC2A2	NM_000340.1(SLC2A2): c.589G> A (p.Val197Ile)	single nucleotide variant	Pathogenic	rs121909741	GRCh37	Chr 3, 170724960: 170724960
8	GCGR	NM_000160.4(GCGR): c.118G> A (p.Gly40Ser)	single nucleotide variant	Pathogenic	rs1801483	GRCh37	Chr 17, 79767715: 79767715
9	IRS1	IRS1, 3-BP DEL, GLY723	deletion	Pathogenic			
10	IRS1	NM_005544.2(IRS1): c.1823C> G (p.Thr608Arg)	single nucleotide variant	Pathogenic	rs104893642	GRCh37	Chr 2, 227661632: 227661632
11	MAPK8IP1	NM_005456.3(MAPK8IP1):	single nucleotide	Pathogenic	rs119489103	GRCh37	Chr 11, 45919710:

id	Gene	Name	Type	Significance	SNP ID	Assembly	Location
		c.176G> A (p.Ser59Asn)	variant				45919710
12	NEUROD1	NM_002500.4(NEUROD1): c.332G> T (p.Arg111Leu)	single nucleotide variant	Pathogenic	rs104893649	GRCh37	Chr 2, 182543256: 182543256
13	PDX1	NM_000209.3(PDX1): c.492G> T (p.Glu164Asp)	single nucleotide variant	Pathogenic, risk factor	rs80356661	GRCh37	Chr 13, 28498478: 28498478
14	ABCC8	NM_001287174.1(ABCC8): c.4138C> T (p.Arg1380Cys)	single nucleotide variant	Pathogenic	rs137852673	GRCh37	Chr 11, 17417462: 17417462
15	ABCC8	NM_001287174.1(ABCC8): c.1744C> G (p.Leu582Val)	single nucleotide variant	Pathogenic	rs137852674	GRCh37	Chr 11, 17452434: 17452434
16	HNF4A	NM_000457.4(HNF4A): c.1204G> A (p.Val402Ile)	single nucleotide variant	Pathogenic	rs137853337	GRCh37	Chr 20, 43057049: 43057049

Table 2: Clinvar genetic disease variations for Type 2 Diabetes Mellitus

disease	gene	chrom	genename	gdbid	omimid	amino	codon	codonAff	descr	hgvs	dbsnp
Diabetes, type 2	ISL1	5q11.2	ISL1 transcrip	376478	600366	Gln-Term	310	310	Gln310Term	928C>T	NULL
Diabetes, type 2	PAX4	7q32	Paired box ge	138170	167413	Arg-Trp	164	164	Arg164Trp	490C>T	rs1219177
Diabetes, type 2	PAX4	7q32	Paired box ge	138170	167413	Arg-Trp	133	133	Arg133Trp	397C>T	rs2233578
Diabetes, type 2	PAX4	7q32	Paired box ge	138170	167413	Arg-Trp	121	121	Arg121Trp	361C>T	rs1142025
Diabetes, type 2	PAX4	7q32	Paired box ge	138170	167413	Arg-Trp	37	37	Arg37Trp	109C>T	rs3515557
Diabetes, type 2	PAX4	7q32	Paired box ge	138170	167413	NULL	NULL	NULL	IVS7 as G-A -1	748-1G>A	NULL
Diabetes, type 2	MAPK8IP1	11p12-p11.2	Mitogen-acti	9956815	604641	Ser-Asn	59	59	Ser59Asn	176G>A	rs1194891
Diabetes, type 2	MTNR1B	11q21-q22	Melatonin re	NULL	600804	Trp-Leu	22	22	Trp22Leu	65G>T	NULL
Diabetes, type 2	MTNR1B	11q21-q22	Melatonin re	NULL	600804	Ala-Pro	42	42	Ala42Pro	124G>C	NULL
Diabetes, type 2	MTNR1B	11q21-q22	Melatonin re	NULL	600804	Ala-Thr	52	52	Ala52Thr	154G>A	NULL
Diabetes, type 2	MTNR1B	11q21-q22	Melatonin re	NULL	600804	Pro-Leu	95	95	Pro95Leu	284C>T	NULL
Diabetes, type 2	MTNR1B	11q21-q22	Melatonin re	NULL	600804	Arg-His	222	222	Arg222His	665G>A	NULL
Diabetes, type 2	MTNR1B	11q21-q22	Melatonin re	NULL	600804	Tyr-Ser	308	308	Tyr308Ser	923A>C	NULL
Diabetes, type 2	MTNR1B	11q21-q22	Melatonin re	NULL	600804	Arg-Trp	330	330	Arg330Trp	988C>T	NULL
Diabetes, type 2	ATP2A3	17p13.3	ATPase, Ca++	133715	601929	Ile-Leu	753	753	Ile753Leu	2257A>C	NULL
Diabetes, type 2	ATP2A3	17p13.3	ATPase, Ca++	133715	601929	Arg-Cys	674	674	Arg674Cys	2020C>T	rs9895012
Diabetes, type 2	ATP2A3	17p13.3	ATPase, Ca++	133715	601929	Val-Met	648	648	Val648Met	1942G>A	rs1473569
Diabetes, type 2	SREBF1	17p11.2	Sterol regula	249853	184756	Thr-Met	226	226	Thr226Met	677C>T	rs7093701
Diabetes, type 2	WFS1	4p16	Wolfram syn	434294	606201	Arg-Cys	653	653	Arg653Cys	1957C>T	NULL
Diabetes, type 2	MTNR1B	11q21-q22	Melatonin re	NULL	600804	Pro-Ser	36	36	Pro36Ser	106C>T	NULL
Diabetes, type 2	MTNR1B	11q21-q22	Melatonin re	NULL	600804	Gly-Ala	109	109	Gly109Ala	326G>C	NULL
Diabetes, type 2	MTNR1B	11q21-q22	Melatonin re	NULL	600804	Met-Val	120	120	Met120Val	358A>G	NULL
Diabetes, type 2	MTNR1B	11q21-q22	Melatonin re	NULL	600804	Tyr-Phe	141	141	Tyr141Phe	422A>T	NULL
Diabetes, type 2	MTNR1B	11q21-q22	Melatonin re	NULL	600804	Ile-Thr	223	223	Ile223Thr	668T>C	NULL
Diabetes, type 2	MTNR1B	11q21-q22	Melatonin re	NULL	600804	Ser-Gly	238	238	Ser238Gly	712A>G	NULL
Diabetes, type 2	MTNR1B	11q21-q22	Melatonin re	NULL	600804	Asp-Asn	246	246	Asp246Asn	736G>A	NULL
Diabetes, type 2	MTNR1B	11q21-q22	Melatonin re	NULL	600804	Phe-Val	250	250	Phe250Val	748T>G	NULL
Diabetes, type 2	MTNR1B	11q21-q22	Melatonin re	NULL	600804	Ala-Val	342	342	Ala342Val	1025C>T	NULL
Diabetes, type 2	IRS2	13q34	Insulin recep	9956176	600797	Leu-Val	647	647	Leu647Val	1939C>G	rs1378527
Diabetes, type 2	ADIPOQ	3q27	Adiponectin,	9956102	605441	Arg-His	55	55	Arg55His	164G>A	rs1436061
Diabetes, type 2	ADIPOQ	3q27	Adiponectin,	9956102	605441	Arg-His	131	131	Arg131His	392G>A	rs7868576
Diabetes, type 2	ADIPOQ	3q27	Adiponectin,	9956102	605441	Gly-Gly	15	15	Gly15Gly	45T>G	rs2241766
Diabetes, type 2	CACNA1E	1q25-q31	Calcium char	434408	601013	NULL	NULL	NULL	(A-G) +1317 t	NULL	rs4131571
Diabetes, type 2	DBI	2q12-q21	Diazepam bir	NULL	125950	Met-Val	71	71	Met71Val	211A>G	rs8192506
Diabetes, type 2	GFPT1	2p13	Glutamine-fr	136391	138292	NULL	NULL	NULL	(T-C) +43 to in	NULL	rs6720415
Diabetes, type 2	MTTP	4q24	Microsomal t	228961	157147	Ile-Thr	128	128	Ile128Thr	383T>C	rs3816873
Diabetes, type 2	PPARGC1A	4p15.1	Peroxisome p	9958452	604517	Thr-Thr	528	528	Thr528Thr	1584A>G	rs3755863
Diabetes, type 2	PPARGC1A	4p15.1	Peroxisome p	9958452	604517	Gly-Ser	482	482	Gly482Ser	1444G>A	rs8192678
Diabetes, type 2	RBP4	10q23-q24	Retinol-bindi	120342	180250	NULL	NULL	NULL	(G-A) -803 to i	NULL	rs3758539
Diabetes, type 2	SLC2A2	3q26.1-q26.2	Solute carrie	119995	138160	Thr-Thr	198	198	Thr198Thr	594G>A	rs5404
Diabetes, type 2	SOD3	4p15.3-p15.1	Superoxide d	125291	185490	Ala-Thr	58	58	Ala58Thr	172G>A	rs2536512
Diabetes, type 2	UTS2	1p36	Urotensin 2	9958587	604097	Thr-Met	21	21	Thr21Met	62C>T	rs228648
Diabetes, type 2	KCNJ11	11p15.1	Potassium in	7009893	600937	NULL	NULL	NULL	(C-T) +215 to t	NULL	rs5210
Diabetes, type 2	KCNJ11	11p15.1	Potassium in	7009893	600937	Ala-Ala	190	190	Ala190Ala	570T>C	rs5218
Diabetes, type 2	ABCC8	11p15.1	ATP-binding c	591370	600509	Lys-Lys	649	649	Lys649Lys	1947G>A	rs1799858

Table 3: Diabetes, type 2, variants annotation (Excel file)

3.10 Congestive Heart Failure

id	Symbol	Products	Description	Score	Implication
1	PLN		Phospholamban	20.87	GeneCards inferred via : Disorders, Summaries(show sections)
2	PDCD1		programmed cell death 1	19.44	GeneCards inferred via : Summaries (show sections)
3	VCL		vinculin	18.01	GeneCards inferred via : Disorders, Summaries(show sections)
4	ADRB1		adrenoceptor beta 1	16.46	GeneCards inferred via : Publications (show sections)
5	ADRA2C		adrenoceptor alpha 2C	15.04	GeneCards inferred via : Publications (show sections)
6	NPPB		natriuretic peptide B	13.57	GeneCards inferred via : Publications (show sections)
7	NPPA		natriuretic peptide A	12.14	GeneCards inferred via : Publications (show sections)
8	REN		renin	10.72	GeneCards inferred via : Publications, Disorders(show sections)
9	ACE		angiotensin I converting enzyme	9.29	GeneCards inferred via : Publications (show sections)
10	UTS2		urotensin 2	7.86	GeneCards inferred via : Publications, Disorders(show sections)
11	AMPD1		adenosine monophosphate deaminase 1	6.43	GeneCards inferred via : Publications (show sections)
12	CBR3		carbonyl reductase 3	5.00	GeneCards inferred via : Publications (show sections)

id	Symbol	Products	Description	Score	Implication
13	ADM		adrenomedullin	3.57	GeneCards inferred via : Publications (show sections)
14	FABP3		fatty acid binding protein 3, muscle and heart (mammary-derived growth inhibitor)	2.14	GeneCards inferred via : Publications (show sections)

Table 4: Genes related to Congestive Heart Failure

3.11 Osteoarthritis

id	Symbol	Products	Description	Score	Implication
1	COL2A1		collagen, type II, alpha 1	20.46	GeneCards inferred via : Publications, Disorders,Gene name, Summaries(show sections)
2	COL9A1		collagen, type IX, alpha 1	19.66	GeneCards inferred via : Summaries, Publications,Disorders (show sections)
3	ASPN		asporin	18.89	GeneCards inferred via : Publications, Disorders,Functions, Summaries(show sections)
4	FRZB		frizzled-related protein	18.12	GeneCards inferred via : Publications, Disorders,Summaries (show sections)
5	MMP13		matrix metallopeptidase 13 (collagenase 3)	17.35	GeneCards inferred via : Publications, Disorders,Summaries (show sections)
6	CTSK		cathepsin K	16.57	GeneCards inferred via : Publications, Disorders,Summaries (show sections)
7	LRCH1		leucine-rich repeats and calponin homology (CH) domain containing 1	15.80	GeneCards inferred via : Publications, Disorders,Summaries (show sections)
8	CTSA		cathepsin A	15.03	GeneCards inferred via : Summaries (show sections)
9	TNFAIP6		tumor necrosis factor, alpha-induced protein 6	14.27	GeneCards inferred via : Publications, Disorders,Summaries (show sections)
10	CTSL		cathepsin L	13.50	GeneCards inferred via : Publications, Disorders,Summaries (show sections)
11	CTSV		cathepsin V	12.72	GeneCards inferred via : Summaries (show sections)
12	UQCC1		ubiquinol-cytochrome c reductase complex assembly factor 1	11.95	GeneCards inferred via : Summaries (show sections)

id	Symbol	Products	Description	Score	Implication
13	OA12		Osteoarthritis QTL 12	11.17	GeneCards inferred via : Publications, Gene name(show sections)
14	OA14		Osteoarthritis QTL 14	10.40	GeneCards inferred via : Publications, Gene name(show sections)
15	OA17		Osteoarthritis QTL 17	9.63	GeneCards inferred via : Publications, Gene name(show sections)
16	OA18		Osteoarthritis QTL 18	8.86	GeneCards inferred via : Publications, Gene name(show sections)
17	OA15		Osteoarthritis QTL 15	8.09	GeneCards inferred via : Publications, Gene name(show sections)
18	OA19		Osteoarthritis QTL 19	7.32	GeneCards inferred via : Publications, Gene name(show sections)
19	OA20		Osteoarthritis QTL 20	6.55	GeneCards inferred via : Publications, Gene name(show sections)
20	OA21		Osteoarthritis QTL 21	5.78	GeneCards inferred via : Publications, Gene name(show sections)
21	OA22		Osteoarthritis QTL 22	5.01	GeneCards inferred via : Publications, Gene name(show sections)
22	OA23		Osteoarthritis QTL 23	4.25	GeneCards inferred via : Publications, Gene name(show sections)
23	OA25		Osteoarthritis QTL 25	3.48	GeneCards inferred via : Publications, Gene name(show sections)
24	OA29		Osteoarthritis QTL 29	2.71	GeneCards inferred via : Publications, Gene name(show sections)
25	OA30		Osteoarthritis QTL 29	1.94	GeneCards inferred via : Publications, Gene name(show sections)

id	Symbol	Products	Description	Score	Implication
26	OA6		Osteoarthritis QTL 6	1.17	GeneCards inferred via : Publications, Gene name(show sections)

Table 5: Genes related to Osteoarthritis

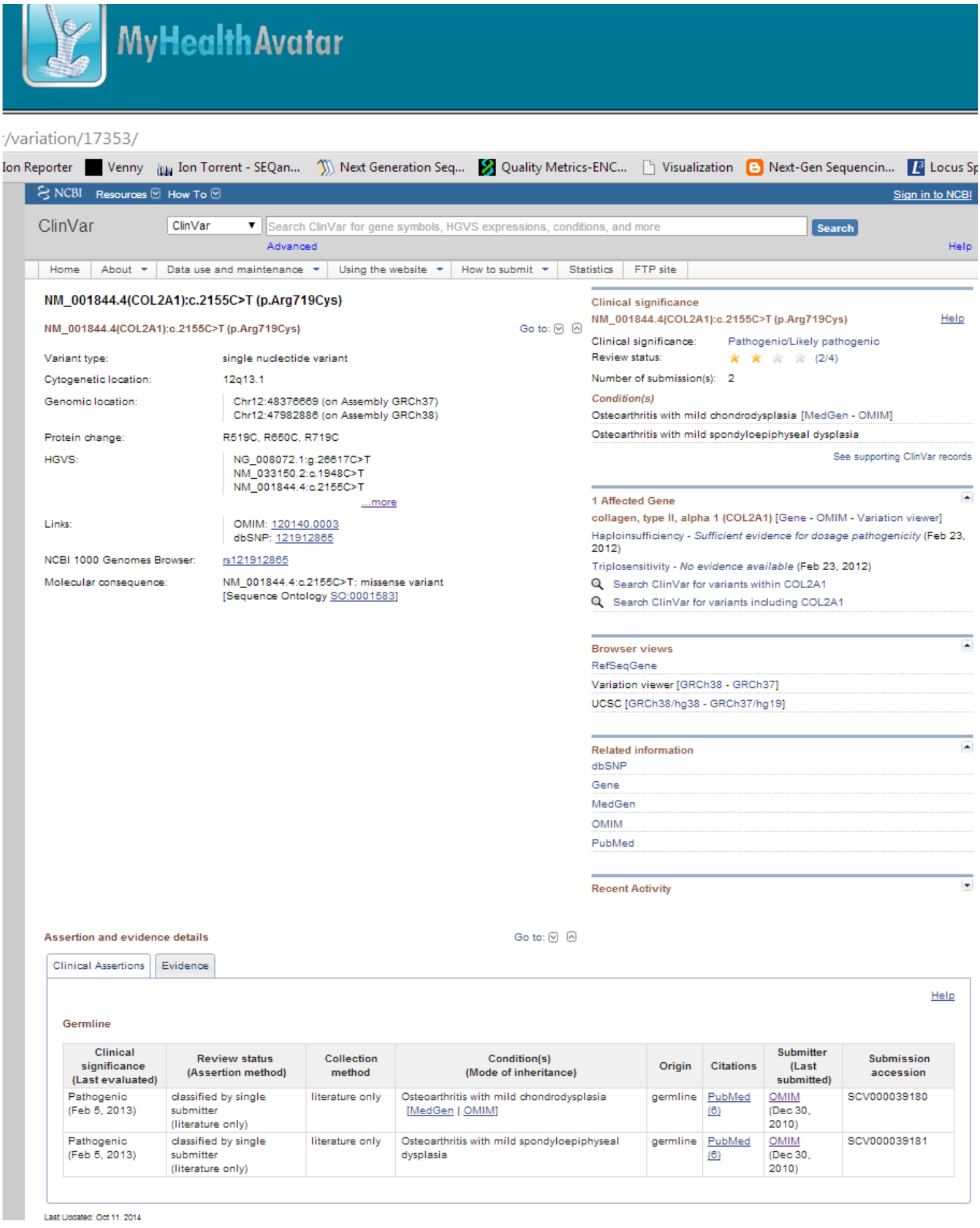


Figure 10: ClinVar COL2A1 annotation

disease	gene	chrom	genename	gdbid	omimid	amino	codon	descr	hgvs	hgvsAll	dbsnp
Osteoarthriti	ADAM12	10q26.3	ADAM metall	NULL	602714	Gly-Arg	48	Gly48Arg	142G>C	142GtoC | G4	rs3740199
Osteoarthriti	COL11A1	1p21	Collagen XI a	120595	120280	Pro-Gln	974	Pro974Gln	2921C>A	2921CtoA | P	rs78046647
Osteoarthriti	COL11A1	1p21	Collagen XI a	120595	120280	Phe-Ile	47	Phe47Ile	139T>A	139TtoA | F4	rs143159512
Osteoarthriti	COL11A2	6p21.3	Collagen XI a	119788	120290	Arg-Trp	539	Arg539Trp	1615C>T	1615CtoT | R	rs145499142
Hip osteoartl	FRZB	2qter	Frizzled-relat	5885889	605083	Arg-Gly	324	Arg324Gly	970C>G	970CtoG | R3	rs7775
Osteoarthriti	IDH1	2q33.3	Isocitrate de	119325	147700	Tyr-Cys	183	Tyr183Cys	548A>G	548AtoG | Y1	rs34599179
Osteoarthriti	MATN3	2p24-p23	Matrilin 3	9836486	602109	Thr-Met	303	Thr303Met	908C>T	908CtoT | T3(	rs77245812
Osteoarthriti	COL11A2	6p21.3	Collagen XI a	119788	120290	NULL	NULL	ins 1 bp non-	3312+2dupT	3312+2dupT	NULL
Osteoarthriti	COL2A1	12q13.11-q1	Collagen II al	119063	120140	Gly-Ser	1176	Gly1176Ser	3526G>A	3526GtoA | G	NULL
Osteoarthriti	COL2A1	12q13.11-q1	Collagen II al	119063	120140	Gly-Gly	858	Gly858Gly	2574C>T	2574CtoT | G	rs141423593
Osteoarthriti	COL2A1	12q13.11-q1	Collagen II al	119063	120140	Arg-Cys	719	Arg719Cys	2155C>T	2155CtoT | R	rs121912865
Osteoarthriti	CALM1	14q24-q31	Calmodulin 1	127560	114180	NULL	NULL	(C-T) +31 to tr	NULL	NULL	rs12885713
Aneurysms-o	SMAD3	15q22.33	SMAD, mothe	4642785	603109	NULL	104	del 1 bp codc	313delG	313delG | 31	NULL
Aneurysms-o	SMAD3	15q22.33	SMAD, mothe	4642785	603109	NULL	179	ins 1 bp codc	539dupC	539dupC | 53	NULL
Osteoarthriti	SMAD3	15q22.33	SMAD, mothe	4642785	603109	Asn-Ile	197	Asn197Ile	590A>T	590AtoT | N1	NULL
Aortic aneury	SMAD3	15q22.33	SMAD, mothe	4642785	603109	NULL	246	del 2 bp codc	741_742delA	741_742delA	NULL
Aortic aneury	SMAD3	15q22.33	SMAD, mothe	4642785	603109	Thr-Ile	261	Thr261Ile	782C>T	782CtoT | T2(	NULL
Aneurysms-o	SMAD3	15q22.33	SMAD, mothe	4642785	603109	Pro-Leu	263	Pro263Leu	788C>T	788CtoT | P2	NULL
Aortic aneury	SMAD3	15q22.33	SMAD, mothe	4642785	603109	Arg-Trp	287	Arg287Trp	859C>T	859CtoT | R2	NULL
Aneurysms-o	SMAD3	15q22.33	SMAD, mothe	4642785	603109	Ala-Pro	349	Ala349Pro	1045G>C	1045GtoC | A	NULL
Aneurysms-o	SMAD3	15q22.33	SMAD, mothe	4642785	603109	NULL	360	ins 1 bp codc	1080dupT	1080dupT | 1	NULL
Osteoarthriti	GDF5	20q11.2	Growth differ	433948	601146	NULL	NULL	(C-T) +45 to tr	NULL	NULL	rs143383

Table 6: Osteoarthritis annotated variants example (Excel file)

5 Conclusion

This deliverables was mainly focused on the datasets that required for the demonstration purposes by the avatar system. Although these datasets will be collected and stored, we will focus mainly on the imaging dataset for the osteoarthritis disease in order to be used by the corresponding use case (UC-OG) described in detail at [4].

6 References

[1] Oleg S. Pianykh. Digital Imaging and Communications in Medicine (DICOM): A Practical Introduction and Survival Guide. 2nd ed. Heidelberg: Springer Berlin, 2009.

[2] The DICOM Standard – NEMA, http://dicom.nema.org/standard.html

[3] The Project dcm4che.org, http://www.dcm4che.org/

[4] Deliverable 7.1, Description of scenarios and use cases for MyHealthAvatar

Appendix 1 – Abbreviations and acronyms

DCE-MRI	Dynamic paramagnetic Contrast Enhancement-MRI
DICOM	Digital Imaging and COmmunications in Medicine
DWI-MRI	Diffusion Weighted Imaging-MRI
HR-MRI	High Resolution morphometric-MRI
MHA	MyHealthAvatar
MRI	Magnetic Resonance Imaging
NGS	Next Generation Sequencing
PACS	Picture Archiving and Communication System
VIBE	Volume Interpolated Breath-Hold Examination