PrECISE

D8.2

Data Management Plan

Project number:	668858
Project acronym:	PrECISE
Project title:	PrECISE: Personalized Engine for Cancer Integrative Study and Evaluation
Start date of the project:	1st January, 2016
Duration:	36 months
Programme:	H2020-PHC-02-2015

Deliverable type:	Report
Deliverable reference number:	PHC-668858 / D8.2 / 1.0
Work package contributing to the deliverable:	WP 8
Due date:	June 2016 – M06
Actual submission date:	8th July, 2016

Responsible organisation:	UZH
Editor:	Peter Wild
Dissemination level:	PU
Revision:	1.0

Abstract:	This document constitutes the Data Management Plan (DMP) of the PrECISE project, explaining how the project plans to manage datasets. It is not a fixed document, but evolves during the lifespan of the project.
Keywords:	Prostate cancer, omics, next generation sequencing, proteomics, genomics, computation, bioinformatics, patient outcome

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 668858.

This work was supported (in part) by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 15.0324-2. The opinions expressed and arguments employed therein do not necessarily reflect the official views of the Swiss Government.

Editor

Peter Wild (UZH)

Contributors and Reviewers (ordered according to beneficiary numbers)

Martina Truskaller, Sandra Lattacher (TEC)

Luis Tobalina (UKAACHEN)

Ulrich Wagner, Qing Zhong (UZH)

Zsolt Torok (ABT)

Disclaimer

The information in this document is provided “as is”, and no guarantee or warranty is given that the information is fit for any particular purpose. The users thereof use the information at their sole risk and liability.

Executive Summary

“Data” is defined as materials generated or collected during the course of conducting research. Many variables govern what constitutes “data” and the management of data, and each project has its own culture regarding data. This data management plan (DMP) provides an instruction on how the data generation and publication process is handled within the PrECISE project.

This document describes how data are managed and handled within the PrECISE project among consortium members. In short, clinical and genetic data are produced by UZH, whereas proteomic data are provided by ETH Zurich.

This data management plan ensures that the data is documented properly and IPR rules, defined within the project consortium, are respected properly. This DMP also handles the data archiving and preservation.

Further details can be found in the elaborated sections. Please note that this document is planned to be kept as a living document and is meant to be updated on a regular basis.

Table of content

List of tables

List of figures

Chapter 1 Introduction

In a research project it is common that several partners work together and produce a lot of data related to the project. Therefore, it is important to specify in an early stage of the project what data will be generated, how it will be shared between the project partners and if it will be publicly available. A data management plan (DMP) is a tool which should assist in managing the data created during the project.

In general, the DMP of the PrECISE project will specify what data is already available and what data will be generated, collected, and processed during the project. It should also provide information whether and how data will be exploited and open for public and re-use. The DMP includes information on what standards and methodologies will be used and how the data will be handled during and after the research project (how the data will be curated and preserved).

The DMP is not a fixed document. It will evolve and gain more precision and substance during the lifespan of the PrECISE project. The first version of the DMP, including information from the first six months of the project, includes the following:

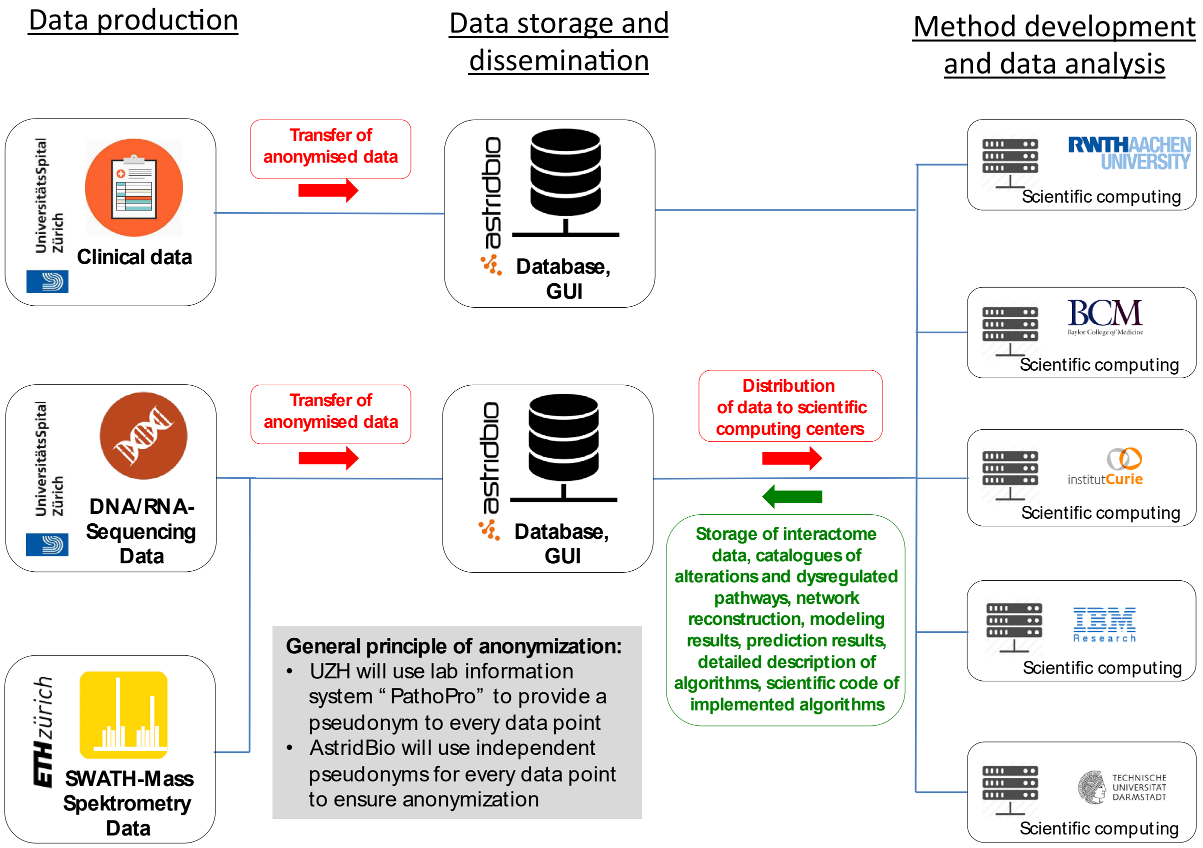
- Methodology (Chapter 2): Data production, storage, dissemination and anaylsis
- Data Generation (Chapter 3): Data set description, research data identification
- Processing and explanation of generated data (Chapter 4): Information about the entity, which is responsible for the data, how the data is collected, an identification of the end-users of the data, and research data identification.
- Accessibility – Data sharing, archiving and preserveration (Chapter 5)

Chapter 2 Methodology

The DMP describes how data are managed and handled within the PrECISE project among consortium members. In short, clinical and genetic data are produced by UZH, whereas proteomic data are provided by ETH Zurich.

Every data point is anonymized, using a two step process: First, a pseudonym is given by the lab information system PathPro at UZH, where only UZH has access. Second, another independent ID is given to each data point on the PrECISE server by AstridBio. Two PrECISE servers will be set up by AstridBio: one for clinical data collection and storage and another one for experimental (genomics, proteomics) data collection and storage. and the key to this second pseudonym is only known by AstridBio. As a result, distribution of data to the other consortium members can be regarded anonymized. Other consortium members, other than UZH, only have access to coded (ETHZ, AstridBio) or anonymized data (all remaining groups). The relevant contributions and data flows are summarizes in the figure below: Data will be provided by UZH and ETHZ. Data storage and dissemination will be done by AstridBio. The consortium members on the right side of the figure will provide algorithm outlines, code and analysis results as stated in the grant proposal. During the project, data and algorithms will shared within members of the project. Upon publication, anonymized data and algorithms/ code will be made publicly available.

In Figure 1, data production, storage, dissemination and analysis is summarized for the PrECISE project.


Data production
Data storage and dissemination
Method development and data analysis
UniversitätsSpital Zürich
Clinical data
Transfer of anonymised data
astridbio
Database, GUI
UniversitätsSpital Zürich
DNA/RNA-Sequencing Data
Transfer of anonymised data
astridbio
Database, GUI
Distribution of data to scientific computing centers
ETH zürich
SWATH-Mass Spektrometry Data
RWTH AACHEN UNIVERSITY
Scientific computing
BCM Baylor College of Medicine
Scientific computing
institut Curie
Scientific computing
IBM Research
Scientific computing
TECHNISCHE UNIVERSITÄT DARMSTADT
Scientific computing

Storage of interactome data, catalogues of alterations and dysregulated pathways, network reconstruction, modeling results, prediction results, detailed description of algorithms, scientific code of implemented algorithms

General principle of anonymization:

- UZH will use lab information system " PathoPro" to provide a pseudonym to every data point
- AstridBio will use independent pseudonyms for every data point to ensure anonymization

Figure 1: PrECISE data production, storage, dissemination and analysis

Chapter 3 Data Generation

The data generation, which is illustrated in Table 1, is described below.

Data set reference and name:

Identifier for the data set to be produced.

Data set description:

Description of the data that will be generated or collected, its origin (in case it is collected), nature and scale to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration of reuse.

Research Data Identification

The boxes (D, A, AI, U and I) symbolize a set of questions that should be clarified for all datasets produced in this project.

Discoverable:

Are the data and associated software produced and/or used in the project discoverable (and readily located), identifiable by means of a standard identification mechanism (e.g. Digital Object Identifier).

Accessible:

Are the data and associated software produced and/or used in the project accessible and in what modalities, scope, licenses (e.g. licencing framework for research and education, embargo periods, commercial exploitation, etc.)

Assessable and intelligible:

Are the data and associated software produced and/or used in the project assessable for and intelligible to third parties in contexts such as scientific scrutiny and peer review (e.g. are the minimal datasets handled together with scientific papers for the purpose of peer review, are data provided in a way that judgements can be made about reliability and the competence of those who created them)?

Useable beyond the original purpose for which it was collected:

Are the data and associated software produced and/or used in the project usable by third parties even long time after the collection of the data (e.g. is the data safely stored in certified repositories for long term preservation and curation; is it stored together with the minimum software, metadata and documentation to make it useful; is the data useful for the wider public needs and usable for the likely purposes of non-specialists)?

Interoperable to specific quality standards:

Are the data and associated software produced and/or used in the project interoperable allowing data exchange between researchers, institutions, organisations, countries, etc. (e.g. adhering to standards for data annotation, data exchange, compliant with available software applications, and allowing re-combinations with different datasets from different origins?)

It is recommended to make an "x" to each applicable box and explain it literally in more detail afterwards.

Data Nr.	Res-pon-sible Bene-ficiary	Data set reference and name	Data set description			Research data identification				
			End user (e.g. university, research organization, SME's, scientific publication)	Existence of similar data (link, information)	Possibility for integration and reuse (Y/N) + information	D[1]	A[2]	AI[3]	U[4]	I[5]
1	UZH	PPM39_rnaSeq	Members of PhosphoNetPPM project in the framework of SystemsX.ch initiative; members of the PrECISE project	Similar data was produced in the framework of The Cancer Genome Atlas project (http://cancergenome.nih.gov/cancersselected/prostatecancer)	Yes. Data can be integrated with information on DNA and proteins from the same samples and compared to /integrated with data from TCGA	Digital Object Indentifiers (DOIs) will be provided to ensure that data and algorithms are discoverable.	Yes, see Chapter 5	Data is being presently analysed and assembled for publication in a high-impact peer-reviewed journal.	Yes, see Chapter 5	Data formats adhere to commonly used scientific standards (e.g. BAM format, VCF files etc.)
2	UZH	PPM39_exomeSeq	Members of PhosphoNetPPM project in the framework of SystemsX.ch initiative; members of the PrECISE project	Similar data was produced in the framework of The Cancer Genome Atlas project (http://cancergenome.nih.gov/cancersselected/prostatecancer)	Yes. Data can be integrated with information on RNA and proteins from the same samples and compared to /integrated with data from TCGA	Digital Object Indentifiers (DOIs) will be provided to ensure that data and algorithms are discoverable.	Yes, see Chapter 5	Data is being presently analysed and assembled for publication in a high-impact peer-reviewed journal.	Yes, see Chapter 5	Data formats adhere to commonly used scientific standards (e.g. BAM format)
3	UZH/ETHZ	PPM39_swathMS	Members of PhosphoNetPPM	-	Yes. Data can be integrated	Digital Object Indentifiers	Yes, see	Data is being presently	Yes, see Chapter 5	Data formats will be

[1] Discoverable

[2] Accessible

[3] Assessable and intelligible

[4] Usable beyond the original purpose of which it was collected

[5] Interoperable to specific quality standards

Data Nr.	Res-pon-sible Bene-ficiary	Data set reference and name	Data set description			Research data identification				
			End user (e.g. university, research organization, SME's, scientific publication)	Existence of similar data (link, information)	Possibility for integration and reuse (Y/N) + information	D[1]	A[2]	AI[3]	U[4]	I[5]
			project in the framework of SystemsX.ch initiative; members of the PrECISE project		with information on DNA and RNA from the same samples and compared to /integrated with data from TCGA	(DOIs) will be provided to ensure that data and algorithms are discoverable.	Chapter 5	analysed and assembled for publication in a high-impact peer-reviewed journal.		developed as part of this study.
4	UZH	PPM39_clinico-pathologic	Members of PhosphoNetPPM project in the framework of SystemsX.ch initiative; members of the PrECISE project	-	Yes. Data can be integrated with information on DNA, RNA and protein from the same samples.	Digital Object Indentifiers (DOIs) will be provided to ensure that data and algorithms are discoverable.	Yes, see Chapter 5	Data is being presently analysed and assembled for publication in a high-impact peer-reviewed journal.	Yes, see Chapter 5	Data formats adhere to commonly used scientific standards (WHO/ISUP2016, GCP)
5	UZH	PPP500_swathMS	Members of the PrECISE project	-	Yes. Data can be integrated with clinico-pathological information from the same samples.	Digital Object Indentifiers (DOIs) will be provided to ensure that data and algorithms are discoverable.	Yes, see Chapter 5	Data is being presently analysed and assembled for publication in a high-impact peer-reviewed journal.	Yes, see Chapter 5	Data formats will be developed as part of this study.
6	UZH	PPP500_	Members of the	-	Yes. Data can	Digital Object	Yes,	Data is being	Yes, see	Data formats

Data Nr.	Responsible Beneficiary	Data set reference and name	Data set description			Research data identification				
			End user (e.g. university, research organization, SME's, scientific publication)	Existence of similar data (link, information)	Possibility for integration and reuse (Y/N) + information	D[1]	A[2]	AI[3]	U[4]	I[5]
		clinico-pathologic	PrECISE project		be integrated with proteomic information from the same samples.	Indentifiers (DOIs) will be provided to ensure that data and algorithms are discoverable.	see Chapter 5	presently analysed and assembled for publication in a high-impact peer-reviewed journal.	Chapter 5	adhere to commonly used scientific standards (WHO/ISUP2016, GCP).
7	UZH	CRPC_heterogeneity_ampliSeq	Members of the PrECISE project	Similar data was produced in the framework of The Cancer Genome Atlas project (http://cancergenome.nih.gov/cancersselected/prostatecancer)	Yes. Data can be integrated with proteomic information from the same patients.	Digital Object Indentifiers (DOIs) will be provided to ensure that data and algorithms are discoverable.	Yes, see Chapter 5	Data is being presently analysed and assembled for publication in a high-impact peer-reviewed journal.	Yes, see Chapter 5	Data formats adhere to commonly used scientific standards (e.g. BAM format)

Table 1: PrECISE data generation

Chapter 4 Processing and explanation of generated data

The following sections provide some additional information to the listed data introduced in Chapter 3. This information includes the entity, which is responsible for the data, how the data is collected, an identification of the end-users of the data, and research data identification.

4.1 PPM39_rnaSeq

This data set consists of RNA sequencing data for 39 prostate cancer patients of the University Hospital Zurich. More specifically, for all 39 patients, the RNA of prostate cancer tissue and the matching normal tissue was sequenced. In addition, for 27 of the patients, a second cancer tissue sample from the same prostate tumour was analysed the same way.

4.1.1 Responsible Beneficiary

The data was produced by UZH.

4.1.2 Gathering Process

RNA sequencing was performed at the Functional Genomics Center Zurich. RNAseq libraries were generated using the TruSeq RNA stranded kit with PolyA-enrichment (Illumina, San Diego, CA, USA). Libraries were sequenced in paired-end mode on an Illumina HiSeq 2500 platform. After base calling, de-multiplexing and in-silico adaptor-trimming, the data was aligned to the hg19 reference genome (UCSC version) using STAR aligner v.2.2.3 (Dobin et al., 2013). Expression values were calculated using the FeatureCount software (Liao et al, 2014).

4.1.3 End-User of the Data

Biofinformatics researchers of the PhosphonetPPM (SystemsX.ch) project and the PrECISE project.

4.1.4 Research Data Identification

Within the scientific community, several initiatives have been launched to cover the minimal standards for the annotation of a sequencing project, such as MIGS or MISEEQE. However these standards are rather preliminary and quite extensive.

For practical reasons, we decided to follow the example of the public data repository Gene Expression Omnibus (GEO), which requires a minimal set of metadata to be uploaded with every sequencing experiment that is submitted to the archive. This set contains descriptive information and protocols for the overall study and individual samples, and references to processed and raw data file names. A template for this set can be found under links (see bibliography, links - Chapter 8). The most important part of the metadata belonging to each data file is covered by the clinical data (e.g. age, tissue type, survival data, time to relapse, BSA levels etc.), which allows for setting individual samples into context with each other.

A pseudonym will be provided by the lab information system PathPro of UZH (format: GXX.XXXX_TA1, _TA2, _No, _Blood) to share data with AstridBio and other members of the consortium. Only UZH can unblind these data. This numbering system provides a unique indentifyer to each sample, linking all omics, clinical and pathological data. This PathoPro identifier will be replaced by AstridBio when setting up the data on the PrECISE server.

4.2 PPM39_exomeSeq

This data set consists of DNA sequencing data for 39 prostate cancer patients of the University Hospital Zurich. More specifically, for all 39 patients, the DNA of prostate cancer tissue, the matching normal tissue and blood was sequenced. In addition, for 27 of the patients, a second cancer tissue sample from the same prostate tumour was analysed the same way.

4.2.1 Responsible Beneficiary

The data was produced by UZH.

4.2.2 Gathering Process

DNA samples were sequenced in paired-end mode on an Illumina HiSeq 2500 platform. After base calling, de-multiplexing and in-silico adaptor-trimming (using Trimmomatic, Bolder et al., 2014), the data was aligned to the hg19 reference genome (UCSC version) using the Bowtie2 software (Langmead et al., 2012). Bam files containing the mapped reads were preprocessed in the following way: Indel information was used to realign individual read using the RealignerTargetCreator and IndelRealigner option of the Genome Analysis Tools Kit (McKenna et al., 2010). Mate-pair information between mates was verified and fixed using Picard tools (http://picard.sourceforge.net) and single base were recalibrated using GATK BaseRecalibrator. After preprocessing, variant calling was carried out by comparing normal or tumor prostate tissue samples with matched blood samples using the programs MuTect (Cibulskis et al., 2013) and , independently, Strelka (Saunders et al., 2012). Somatic variants that were not detected by both programs were filtered out using CLC Genomics Workbench (CLC Genomics Workbench 8.0.3, https://www.qiagenbioinformatics.com) as well as those that had an entry in the dbsnp (Sherry et al., 2001) common database or those that represented synonymous variants without predicted effects on splicing.

4.2.3 End-User of the Data

Biofinformatics researchers of the PhosphonetPPM (SystemsX.ch) project and the PrECISE project.

4.2.4 Research Data Identification (see also 4.1.4)

A pseudonym will be provided by the lab information system PathPro of UZH (format: GXX.XXXX_TA1, _TA2, _No, _Blood) to share data with AstridBio and other members of the consortium. Only UZH can unblind these data. This numbering system provides a unique indentifyer to each sample, linking all omics, clinical and pathological data. This PathoPro identifier will be replaced by AstridBio when setting up the data on the PrECISE server.

4.3 PPM39_swathMS

This data set consists of Swath-MS proteomics data for 39 prostate cancer patients of the University Hospital Zurich. More specifically, for all 39 patients, the protein lysates of prostate cancer tissue, and the matching normal tissue was analysed. In addition, for 27 of the patients, a second cancer tissue sample from the same prostate tumour was analysed the same way.

4.3.1 Responsible Beneficiary

The data was produced by the Institute for Molecular Systems Biology of the ETH Zurich (ETHZ).

4.3.2 Gathering Process

Production of data has been described in detail by Tiannan et al., Nature Medicine 2015.

4.3.3 End-User of the Data

Biofinformatics researchers of the PhosphonetPPM (SystemsX.ch) project and the PrECISE project.

4.3.4 Research Data Identification

A pseudonym will be provided by the lab information system PathPro of UZH (format: GXX.XXXX_TA1, _TA2, _No) to share data with AstridBio and other members of the consortium. Only UZH can unblind these data. This numbering system provides a unique indentifyer to each sample, linking all omics, clinical and pathological data. This PathoPro identifier will be replaced by AstridBio when setting up the data on the PrECISE server.

4.4 PPM39_clinico-pathologic

This data set consists of clinical follow-up data (age at diagnosis, Progression-free survival time and status) and pathological data (UICC TNM classification, Gleason grade, WHO/ISUP 2016 grade group), for 39 prostate cancer patients of the University Hospital Zurich. More specifically, for all 39 patients, prostate cancer tissue was graded. In addition, for 27 of the patients, a second (lower grade) cancer tissue area from the same prostate tumour was selected and grade the same way.

4.4.1 Responsible Beneficiary

The data was produced by UZH.

4.4.2 Gathering Process

The 2016 WHO/ISUP classification of urological tumors was used to stage and grade the respective lesions. Clinical follow-up data of patients of where retrieved from the Zurich proCOC (**Pro**state **C**ancer **O**utcomes **C**ohort) study.

4.4.3 End-User of the Data

Biofinformatics researchers of the PhosphonetPPM (SystemsX.ch) project and the PrECISE project.

4.4.4 Research Data Identification

A pseudonym will be provided by the lab information system PathPro of UZH (format: GXX.XXXX) to share data with AstridBio and other members of the consortium. Only UZH can unblind these data. This numbering system provides a unique indentifyer to each sample, linking all omics, clinical and pathological data. This PathoPro identifier will be replaced by AstridBio when setting up the data on the PrECISE server.

4.5 PPP500_swathMS

This data set consists of Swath-MS proteomics data for 500 prostate cancer patients of the University Hospital Zurich. More specifically, for all 500 patients, the protein lysates of prostate cancer tissue and matching normal tissue was analysed.

4.5.1 Responsible Beneficiary

The data was produced by the Institute for Molecular Systems Biology of the ETH Zurich (ETHZ).

4.5.2 Gathering Process

Production of data has been described in detail by Tiannan et al., Nature Medicine 2015.

4.5.3 End-User of the Data

Biofinformatics researchers of the PhosphonetPPM (SystemsX.ch) project and the PrECISE project

4.5.4 Research Data Identification

A pseudonym will be provided by the lab information system PathPro of UZH (format: GXX.XXXX_TA1, _TA2, _No) to share data with AstridBio and other members of the consortium. Only UZH can unblind these data. This numbering system provides a unique indentifyer to each sample, linking all omics, clinical and pathological data. This PathoPro identifier will be replaced by AstridBio when setting up the data on the PrECISE server.

4.6 PPP500_clinico-pathologic

This data set consists of clinical follow-up data (age at diagnosis, Progression-free survival time and status) and pathological data (UICC TNM classification, Gleason grade, WHO/ISUP 2016 grade group), for 500 prostate cancer patients of the University Hospital Zurich. More specifically, for all 500 patients, prostate cancer tissue was graded.

4.6.1 Responsible Beneficiary

The data was produced by UZH.

4.6.2 Gathering Process

The 2016 WHO/ISUP classification of urological tumors was used to stage and grade the respective lesions. Clinical follow-up data of patients of where retrieved from the Zurich proCOC (**Pro**state **C**ancer **O**utcomes **C**ohort) study.

4.6.3 End-User of the Data

Biofinformatics researchers of the PhosphonetPPM (SystemsX.ch) project and the PrECISE project.

4.6.4 Research Data Identification

A pseudonym will be provided by the lab information system PathPro of UZH (format: GXX.XXXX) to share data with AstridBio and other members of the consortium. Only UZH can unblind these data. This numbering system provides a unique indentifyer to each sample, linking all omics, clinical and pathological data. This PathoPro identifier will be replaced by AstridBio when setting up the data on the PrECISE server.

4.7 CRPC_heterogeneity_ampliSeq

This data set consists of DNA sequencing data for 10 prostate cancer patients of the University Hospital Zurich. More specifically, for all 10 patients, the genomic DNA of 10 different tumour areas (CRPC) and of a matching normal tissue area will be sequenced.

4.7.1 Responsible Beneficiary

The data will be produced by UZH.

4.7.2 Gathering Process

DNA samples will sequenced using Amplicon sequencing on the Ion Torrent Proton System (genes *PTEN, AR, TP53, SPOP, FOXA1*).

4.7.3 End-User of the Data

Biofinformatics researchers of the PhosphonetPPM (SystemsX.ch) project and the PrECISE project.

4.7.4 Research Data Identification (see also 4.1.4)

A pseudonym will be provided by the lab information system PathPro of UZH (format: GXX.XXXX_TA1-12, _No) to share data with AstridBio and other members of the consortium. Only UZH can unblind these data. This numbering system provides a unique indentifyer to each sample, linking all omics, clinical and pathological data. This PathoPro identifier will be replaced by AstridBio when setting up the data on the PrECISE server.

Chapter 5 Accessibility – Data sharing, archiving and preservation

We will store clinical and experimental data on separate servers. Experimental data will be accessible for up- and download using encrypted standard file transfer protocols. We will separate data produced within the PrECISE project and data uploaded from other data repositories (e.g. TCGA data). The respective meta-data will be stored together with the clinical data. Clinical data will be accessible via web applications using encryption. The system for the data management will be based on the SmartBioBank software (http://www.smartbiobank.com). This software allows for integration of clinical and experimental/genomic data and enables data sharing and database merging between independent research groups. The clinical data and the meta-data should be searchable via a web-based interface.

In the beginning of the project, data will be shared exclusively between members of the PrECISE project. As a sharing platform for small amounts of data up to 100MB the already existing project SVN repository https://precise.technikon.com is used. This allows easy synchronization between the partners as well as data versioning. It has to be noted that only project partners have access to the project SVN. A data transfer agreement has to be signed for getting access to unpublished data (see Chapter 9).

The PrECISE Executive Board will make decisions on the publication of the data. Most journals in the biomedical field require the submission of genomic data together with standard meta-data to public repositories. This means that in case of publishing our findings in scientific journals, we will make the data available to the public. In this case, the data sets will get a persistent identifier.

External researchers will be able to apply for access to the PrECISE data before the general release to the public. In this case an application form needs to be filled detailing a research plan, ethical approval etc.. The application form will be available on the PrECISE project website together with information on approval process and data content of the PrECISE database. The Executive board can grant access to the entire database or restrict access to selected data sets. Non-profit organizations should cover the expenses of the consortium related to the data migration etc. For-profit organizations will have to pay an additional fee for data access.

The data can be used to discover new biomarkers or sets of biomarkers, which allow for improved diagnostics of prostate cancer. In addition, different levels of data were acquired of a complex biological systems transitioning from one developmental stage (i.e. normal tissue) to another (i.e. tumour tissue). Integrating these data sets may result in important discoveries that will elucidate the dynamics of biological systems that fails to control homeostasis resulting in disease.

No data will be deleted. Instead, preservation of raw and meta-data will be reached by two efforts: first, setup of a data server by AstridBio; second, publication of raw and meta-data in journals specialized in long-term preservation of biomedical data (e.g. Nature Scientific Data using the DataVerse platform). Through this, no costs (time/effort) to prepare the data for sharing/preservation are involved, and data will be publicly available for a very long time period (decades).

Chapter 6 Summary and Conclusion

This data management plan outlines the handling of data generated within the PrECISE project, during and after the project lifetime. As the deliverable will be kept as a living document it will be regularly updated by the consortium (the next time in M18 within the periodic report). The partners put into write their plans and guarded expectations regarding valuable and publishable data.

Chapter 7 List of Abbreviations

Abbreviation	Explanation
WHO	World Health Organization
ISUP	International Society for Urologic Pathology
GCP	Good Clinical Practice
UZH	University of Zurich
ETHZ	ETH Zurich
CRPC	Castration resistant prostate cancer

Chapter 8 Bibliography

References

- Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30:2114-21120
- Cibulskis K1, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 31:213-219
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15-21. PMID: 23104886
- Langmead B, Salzberg S. (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods, 9:357-359
- Liao Y, Smyth GK and Shi W (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7):923-30. PMID: 24227677
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Research 20:1297-303
- Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK (2012) Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 28:1811-1817
- Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29:308-311

Links

- http://www.ncbi.nlm.nih.gov/geo/info/examples/seq_template_v2.1.xls

Chapter 9 Appendix – Data transfer agreement

DATA TRANSFER AGREEMENT

This agreement (hereinafter referred to as "Agreement") is made and entered by and between:

UNIVERSITAET ZUERICH

acting on behalf of its Institut für Klinische Pathologie, Schmelzbergstrasse 12, 8091 Zurich, Switzerland ("UZH")

and

XXX ("Recipient")

Hereinafter jointly referred to as "Parties" and individually as "Party";

WHEREAS

a) The Parties are collaborating in the course of the EU-H2020 project entitled "PrECISE".

b) UZH is the exclusive owner of Clinical and Genomics Data on prostate cancer ("DATA") specified in Annex 1 to this Agreement which is Background included according to the Consortium Agreement of PrECISE signed on Dec 11, 2015;

c) RECIPIENT, through XXX hereinafter referred to as "RECIPIENT SCIENTIST", has requested UZH, through Prof. Peter Wild, hereinafter referred to as "UZH SCIENTIST", to provide RECIPIENT with the DATA for use by RECIPIENT'S SCIENTIST for the purpose of the RESEARCH PLAN;

d) UNIVERSITY is willing, subject to the terms and conditions of this Agreement, to provide the DATA to RECIPIENT.

I. Definitions

1. DATA: The data being transferred under this Agreement as specified in Annex I to this Agreement.

2. RESEARCH PLAN: means the description of the research as defined in Annex 1 of the PrECISE Grant Agreement..

3. EFFECTIVE DATE: The date of last signing of this Agreement.

4. RESULTS: All Results (as defined in the PrECISE Grant Agreement) arising from the use of the DATA by RECIPIENT.

II. Terms and Conditions of this Agreement:

1. The DATA provided is and remains the property of UNIVERSITY.

2. In accordance with Article 10.8 of the Consortium Agreement The RECIPIENT and the RECIPIENT SCIENTIST agree that the DATA are Confidential Information of UZH and shall solely be used for implementation of the Action and shall as soon as the Action will be concluded or this Agreement will expire or be terminated for whatever reason, be returned to UZH, if possible, or be destroyed with required care. No Access rights for Exploitation of DATA are granted.

3. RECIPIENT'S SCIENTIST shall keep UZH SCIENTIST informed of the RESULTS. For RESULTS, the terms and conditions of the PrECISE Consortium Agreement shall apply unless otherwise agreed in this Agreement.

4. RECIPIENT will refrain from publishing the RESULTS until the publication by UZH of the results of the study in which DATA was gained.

   Thereafter RECIPIENT will be free to publish and disclose the RESULTS but agrees to submit the proposed disclosure to UZH for review at least fifteen (15) days prior to the scheduled submission for publication or disclosure. If UZH believes that the publication or disclosure contains INFORMATION of UZH, UZH has the right to request for deferral of the publication for up to sixty (60) days from the date of submission of the documents to UZH. Any such INFORMATION will be removed from the publication or disclosure. UZH also has the right to provide comments on the manuscript and both Parties shall discuss in good faith to incorporate such comments in the publication or disclosure.

   All publications of the RESULTS must include at least four co-authors of UZH, in accordance with generally recognized principles of scientific collaborations.

5. This Agreement will become effective on the Effective Date and will terminate upon termination or termination of the participation of RECIPIENT of the PrECISE Consortium Agreement. Parties can terminate this Agreement by giving a three (3) months prior written notice. Any clauses that will be expected or intended by its nature to survive the termination or the expiration of this Agreement, shall survive the termination or the expiration of this Agreement.

IN WITNESS WHEREOF, the parties have executed this Agreement, in duplicate originals, as of the Effective Date.

UNIVERSITAET ZUERICH	**XXX**
Date:___________________	Date: ___________________
By:____	By: ____

ANNEX I

DATA:

____

ANNEX II

____