Plan Overview

A Data Management Plan created using DMPonline.be

Title: High-resolution spatial maps of the tumor microenvironment during checkpoint immunotherapy

Creator: Heesoo Song

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Project abstract:

Recent advances in imaging-based spatial technologies allow researchers to detect expression of 500-1000 genes at single-cell level within the tissue context. This allows tumor heterogeneity, both in terms of the cancer cells and their microenvironment (TME), to be explored in situ. State-of-the-art methods to detect heterogeneity use scRNAseq data generated on dissociated cells. To recover the same resolution at the spatial level, integration of both methods is therefore needed. In my host lab, Lodi et al. developed a unique pan-cancer scRNAseq atlas covering >200 tumors from 9 different cancer types, characterizing TME heterogeneity at unprecedented resolution. The aim of my project is to construct a bio-informatics pipeline that can integrate these scRNAseq data to spatially resolve the TME at a similar resolution as for scRNAseq data. Particularly, I will measure expression of 500 selected genes in tumor sections using MERFISH to reconstruct cellular neighborhoods and characterize underlying cell-cell communication networks. Existing methods in each analysis step of the pipeline, including cell segmentation, data integration, construction of neighborhoods and communication networks, will be benchmarked. Once optimized, I will apply the pipeline to serially sampled tumor biopsies exposed to checkpoint immunotherapy (ICB) and monitor spatial dynamics in the TME during treatment. This will uncover novel spatial neighboorhoods within the TME determining response to ICB.

ID: 213403

Start date: 01-11-2024

End date: 31-10-2028

Last modified: 17-04-2025

High-resolution spatial maps of the tumor microenvironment during checkpoint immunotherapy Application DMP

Questionnaire

The questions in this section should only be answered if you are currently applying for FWO funding.
Are you preparing an application for funding?

- No

High-resolution spatial maps of the tumor microenvironment during checkpoint immunotherapy
DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- Not applicable

High-resolution spatial maps of the tumor microenvironment during checkpoint immunotherapy
GDPR

GDPR

Have you registered personal data processing activities for this project?

- Yes

High-resolution spatial maps of the tumor microenvironment during checkpoint immunotherapy
FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .csv,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	

Spatial transcriptomics data	Imaging-based spatial transcriptomics data from tumor samples collected from patients with breast cancer during immunotherapy	Generate new data	Physical + Digital	Experimental	Tabular data: comma-separated value files (.csv), tab-delimited file (.tab), delimited text (.txt), MS excel (.xls/.xlsx) Text files: Plain text data (Unicode, .txt), MS Word (.doc/.docx), Adobe Portable Document Format (.pdf), LaTex (.tex) format R/Python scripts (.R, .py) Scanpy and Seurat downstream analysis objects (.h5ad, .rds) Digital images in raster formats: uncompressed TIFF (.tif/.tiff), JPEG (.jpg), PNG (.png), Adobe Portable Document Format (.pdf), bitmap (.bmp), .gif Digital images in vector formats: scalable vector graphics (.svg), Adobe Illustrator (.ai) Spatial transcriptomics visualizer object: Vizgen (.vzg/.vzg2) Metadata: textual/tabular data (.rtf, .xml, .txt, .xls, .hdf5, .parquet)	<10TB	10 FFPE tumor samples

Sequencing data (scRNA-seq)	Single-cell profiling data from tumor samples collected from cancer patients during immunotherapy	Reuse existing data	Digital	Experimental Compiled/aggregated data	Tabular data: comma-separated value files (.csv), tab-delimited file (.tab), delimited text (.txt), MS excel (.xls/.xlsx) Text files: Plain text data (Unicode, .txt), MS Word (.doc/.docx), Adobe Portable Document Format (.pdf), LaTex (.tex) format R/Python scripts (.R, .py) Scanpy and Seurat downstream analysis objects (.h5ad, .rds) Digital images in raster formats: uncompressed TIFF (.tif/.tiff), JPEG (.jpg), PNG (.png), Adobe Portable Document Format (.pdf), bitmap (.bmp), .gif Digital images in vector formats: scalable vector graphics (.svg), Adobe Illustrator (.ai) Next generation sequencing raw data: binary base call format (.bcl), .fastq (zipped as .gz) Structural variations data: .vcf, .bcf Read/UMI count data: .tsv, .rds Coverage data: .bed, .bg, .bedGraph, .bw, .bigwig Sequence alignment data: .bam Metadata: textual/tabular data (.rtf, .xml, .txt, .xls, .hdf5, .parquet)	<10 TB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

The data compilation is created by Lodi et al. which is under review in Cell Reports medicine. Each dataset is generated by Diether Lambrechts Lab, and the raw sequencing reads are deposited under restricted access in the European Genomephenome Archive (EGA). Data accession number included in compiled scRNA-seq data are:

- EGAS00001004809
- E-MTAB-8107
- EGAS00001004871
- EGAS00001007547
- EGAS00001004987
- EGA50000000033
- EGAS00001006488
- E-MTAB-6149
- E-MTAB-6653

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes, human subject data

ALBERT:
ALBERT is a study with secondary use of HBM considering patients with early TNBC treated at UZ Leuven between 2016 and 2024. Ethical approval for the use of samples and patient data is already in place, and was approved by the Ethics Committee Research UZ/KU Leuven.
Sponsor: UZ Leuven
Collaborator: Prof. Dr. Giuseppe Floris
S-number: S68408
Approval date: 12/03/2024

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes

We will generate imaging-based spatial transcriptomics data from breast cancer patient samples collected in ALBERT study. The patient samples that we will receive are coded with a study sample ID (the coding key remains with the oncologists, UZ Leuven). This code does not carry any personal identifiers, keeping the identity of the study participant private and confidential. Access to the coding key is necessary to link any data or biological samples back to a subject identifier. In addition, when the samples are processed in our lab, they also receive a DILA ID (double-coded) and this DILA ID is further used in the downstream analyses. We will also receive pseudonymized patient data linked to the sample IDs. The clinical patient data will include the following information:

- patient number, sample ID, sample type (pre/on-treatment), biopsy type, gender, inclusion date, age at diagnosis, cohort (upfront surgery or pre-treated with chemo), histological type, tumour grade, Ki67 at core biopsy, Ki67 at resection biopsy, pTNM, cTNM, HER2 status, TILs at core biopsy, TILs at resection biopsy, pre/post-menopausal state, BRCA status, breast cancer subtype, histological type, previous medical history, concomitant medication, BMI.

The spatial transcriptomics data that will be generated within this project will be correlated with the pseudonymized clinical data. The spatial transcriptomics data and the associated pseudonymized patient data are defined as sensitive personal data and will only be processed in accordance with the institutional SOPs, the principles of the General Data Protection Regulation (GDPR) 2016/679 and the Belgian privacy law. These procedures include procedures for pseudonymization, data storage and data protection. This study will only use historical samples collected for diagnostic reasons (=residuary material). Therefore, no additional consent is needed. Patients will be informed that their sample will be used in this study through the MyNexusHealth app with a chance to opt out of the study. The data will be processed and stored on the institutional password protected IT infrastructure which is protected by a genuine user authentication system relying on username and password. Access to the data as well as the access level will be limited on a project need and individual basis. Only the researchers working on the project have access to these data. Due to the sample labeling as protective measure, the researchers are not able to decipher the identity of the donor.
All data that will be collected and the strategy to guarantee the privacy of the study participants are specified in the research protocol approved by the ethical committee.

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- Yes

We do not exclude that the proposed work could result in research data with potential for tech transfer and valorization. Both VIB and KU Leuven have a policy to actively monitor research data for such potential. If there is substantial potential, the invention will be thoroughly assessed, and in a number of cases the invention will be IP protected (mostly patent protection or copyright protection). As such the IP protection does not withhold the research data from being made public. In the case a decision is taken to file a patent application it will be planned so that publications need not be delayed.

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- Yes

The work described in the current proposal is included as the translational research part of the ALBERT study. In the current proposal, we will use patient samples and data collected at UZ Leuven between 2016 and 2024. Therefore, there is a Material and Data Transfer Agreement between the legal entities of KU Leuven (> UZ Leuven) and VIB. This existing agreements between VIB and KU Leuven do not restrict publication of data.

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- Yes

Parties have expressly agreed that any and all data as collected and prepared in the context of this study shall be the joint property of UZ Leuven / KU Leuven and VIB.

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

- We have an electronic (password-protected) MTD file (.xls format) stored on the password-protected KU Leuven shared drive, giving an overview of all samples that are processed in our lab with specific notes referring to specific datasets. These notes describe the biological/clinical samples used, the physical storage location of the (processed) samples, experimental setup and protocols used, sequences/omics/imaging data generated, links to the specific computer location and the specific names of the respective datasets. This MTD describes the connection between lab samples, sample IDs and files on our data storage, so that data files, lab samples, and experimental notes remain properly linked to the same sample ID.
- The data file names that are being generated will be named according to a standard procedure, so that all the name of all files in a given dataset have the same format. All changes in the files will be recorded. All names start with the date (and time if applicable), followed by the project acronym, a short but specific descriptive name (e.g. type of experiment) and a version number (containing leading zeros as needed) if applicable. Names only contained letters, numbers and underscores. Dots will be only used for version control indicators (minor revisions indicated by decimal numbers, and major revisions by whole numbers): YYYMMDD_Project_Experiment_version.format.
- Research methods and practices (SOPs) will be fully documented. When the wet lab techniques, scripts, algorithms and software tools are finalized, they will be additionally described in manuscripts and/or on GitHub.
- Raw spatial transcriptomics data (staining images, segmentation, and read count data matrices; each named with their DILA ID) will be stored on the Vlaamse Super computer (VSC), ordered in folders per sequencing-run, including an .xlsx file with the sample sheet information containing the DILA IDs sequenced in that run and the sequencing run information per DILA ID (Illumina sequencer, lane and index information). The name of the folder will contain the date of the sequencing run and the Illumina sequencer used. When data are published, raw and processed sequencing data will be uploaded on a public repository (e.g. EGA) with appropriate access control if required, to enable sharing and long-term validity of the data. Any data shared will only be released prior to a Data Transfer Agreement that will have to include the necessary conditions to guarantee protection of personal data (according to European GDPR law). Double/triple-coded read count data matrix (linked to double/triple-coded human data) will be available on our website (https://lambrechtslab.sites.vib.be/en/data-access). Raw imaging data (microscopy files, imaging protocol details, count matrices) will be stored on password-protected and backed up VIB-KU Leuven IT infrastructure in folders per imaging run. The name of the folder will contain the date of the imaging run and the microscope and imaging protocol used.
- Raw imaging data (microscopy files, imaging protocol details, count matrices) will be stored on password-protected and backed up VIB-KU Leuven IT infrastructure in folders per imaging run. The name of the folder will contain the date of the imaging run and the microscope and imaging protocol used.
- Manuscripts: metadata information will be submitted alongside the final version of the manuscript, including the names, titles, email addresses, ORCIDs and affiliations of all authors. Upon publication, this metadata information will be also submitted to bibliographic databases such as Medline, Web of Science, BioRxiv, KU Leuven Lirias. All manuscripts will be assigned a unique Digital Object Identifer (DOI) by the publisher. Manuscripts are given a descriptive title, and are accompanied by keywords provided by the authors in order to maximize their findability.

All data will be processed and (temporarily) stored on secured, password-protected and backed up servers of VIB-KU Leuven (managed by ICT of the Biomedical Sciences Group) or on the VSC.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- Yes

Imaging data also requires specific metadata when submitted to public repositories such as OMERO or SpatialDB.
Data documentation will be tailored to their ultimate deposition in public repositories, with spreadsheet headers corresponding to fields required by these public repositories. Technical and analytical methods used to generate the data will be documented in sufficient detail to allow for independent reproduction. These will include analysis package version numbers, analysis kit, disease status, treatment type and duration, organism, genome build.... When depositing data in a repository, the final dataset will be accompanied by this information in the file format that the repository provides. This will allow the data to be understood by other members of the laboratory and add context to the dataset for future reuse.

3. Data storage & back-up during the research project

Where will the data be stored?

All electronical data collected and generated during the project will be processed and (temporarily) stored on secured, password-protected and backed up servers of VIB-KU Leuven (managed by ICT of the Biomedical Sciences Group).
The sequencing data generated during the project will either be stored on VIB-KU Leuven servers or on the Flemish Supercomputer Centre (VSC), initially in the staging and archive area, and later only in the archive area (archive is mirrored).
Raw and processed data will be submitted to a public repository (e.g. EGA) with appropriate access control if required, to enable sharing and long-term validity of the data. Double/triple-coded read count data matrix (linked to double/triple-coded human data) will be available on our website (https://lambrechtslab.sites.vib.be/en/data- access).
All patient samples and their derivatives will be stored in labeled tubes or SBS plates in -20°C or -80°C freezers purchased by our own funding. The samples will be registered and handled according to the UZ Leuven Biobank guidelines, in compliance with the Belgian law on human body material (dd 19-12-2008).

How will the data be backed up?

KU Leuven drives are automatically (daily) backed up using KU Leuven services according to the following scheme:

- Data stored on the “L-drive” is backed up daily using snapshot technology, where all incremental changes in respect of the previous version are kept online; the last 14 backups are kept.
- Data stored on the “J-drive” is backed up hourly, daily (every day at midnight) and weekly (at midnight between Saturday and Sunday); in each case the last 6 backups are kept.
- Data stored on the digital vault is backed up using snapshot technology, where all incremental changes in respect of the previous version are kept online. As standard, 10% of the requested storage is reserved for backups using the following backup regime: an hourly backup (at 8 a.m., 12 p.m., 4 p.m. and 8 p.m.), the last 6 of which are kept; a daily backup (every day) at midnight, the last 6 of which are kept; and a weekly backup (every week) at midnight between Saturday and Sunday, the last 2 of which are kept.
- Incremental backups are done daily from one 20 TB QNAP NAS to a second 20 TB QNAP NAS.

All sequencing data stored on the Flemish Supercomputer Centre (VSC) will be transferred on a regular basis to the archive area which is backed up.
Data is stored on EGA/our website for the purpose of data sharing.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

There is sufficient storage and back-up capacity on all VIB-KU Leuven servers:

- The “L-drive” is an easily scalable system, built from General Parallel File System (GPFS) cluster with NetApp e-series storage systems, and a CTDB samba cluster in the front-end.
- The “J-drive” is based on a cluster of NetApp FAS8040 controllers with an Ontap 9.1P9 operating system.
- The Staging and Archive on VSC are also sufficiently scalable (petabyte scale).

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

Since we are working with personal data (pseudonymized genome data and associated pseudonymized patient data), the data will be processed and stored on the VIB-KU Leuven IT infrastructure which is protected by a genuine user authentication system relying on username and password: Both the “L-drive” and “J-drive” servers are accessible only by laboratory members, and are mirrored in the second ICTS datacenter for business continuity and disaster recovery so that a copy of the data can be recovered within an hour. Access to the digital vault is possible only through using a KU Leuven user-id and password, and user rights only grant access to the data in their own vault. Sensitive data transfer will be performed according to the best practices for “Copying data to the secure environment” defined by KU Leuven. The operating system of the vault is maintained on a monthly basis, including the application of upgrades and security patches. The server in the vault is managed by ICTS, and only ICTS personnel (bound by the ICT code of conduct for staff) have administrator/root rights. A security service monitors the technical installations continuously, even outside working hours.
Access to the data as well as the access level will be limited on a project need and individual basis. Only the researchers working on the project has access to these data. Due to the sample labeling as protective measure, the researchers are not able to decipher the identity of the donor.
No personal data will be stored on the VSC nor local drives. The coding key to patient information of linked pseudonymized data will be kept with the oncologists of UZ Leuven.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

The total estimated cost of data storage during the project is ~2500 EUR. This estimation is based on the following costs:

- €113,84/TB/Year for the “L- drive”
- €519/TB/Year for the “J-drive”
- €35/TB/Year for VSC staging area and archive area.
- €1913,58/TB/Year for desktop file storage on file shares with very high input/output speed
- €450,76/TB/Year for desktop file storage on file shares with normal input/output speed

We need up to 20TB of data storage, but the storage after the project is smaller because during the project a large working space is needed, and post-publication data are made accessible via open access platforms. Data storage and backup costs are covered by our own funding. Electricity costs for the freezers present in the labs are included in general lab costs.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

The work performed in this project is included as the translational research part of the clinical study involving checkpoint immunotherapy (i.e. the ALBERT study: S68408). The datasets, collected in the context of this study, will be archived for 20-25 years in a safe, secure & sustainable way for purposes of reproducibility verification, and potential reuse, in accordance with the UZ/KU Leuven policy and the applicable laws, rules and legislation in relation to clinical trials/studies. Subsequently, they may be kept for an additional period of time, for the described scientific purposes of the clinical studies or for any legal reason (change of obligations with regard to storage, for example).

Where will these data be archived (stored and curated for the long-term)?

As a general rule, datasets will be made openly accessible, whenever possible via existing platforms that support FAIR data sharing (www.fairsharing.org), at the latest at the time of publication or preprint deposition.
For all other datasets, long term storage will be ensured as follows:

- Large (sequencing) data will be stored on VSC archive
- Small digital files will be stored on the “L-drive” and on the digital vault.
- Developed algorithms and software will be stored on VSC archive and/or L-drive, as well on public repositories such as Github.com

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

The storage after the project is smaller because during the project a large working space is needed, and post-publication data are made accessible via open access platforms. Data storage and backup costs are covered by our own funding. Electricity costs for the freezers present in the labs are included in general lab costs.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository
- Yes, in a restricted access repository (after approval, institutional access only, …)
- Other, please specify:

The PI in the present project is committed to publish research results to communicate them to peers and to a wide audience. As a general rule all research outputs related to publications will be made openly accessible, whenever possible via existing platforms that support FAIR data sharing (www.fairsharing.org). Metadata will contain sufficient information to support data interpretation and reuse, and will be conform to community norms. These repositories clearly describe their conditions of use (typically under a Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication, a Creative Commons Attribution (CC-BY), or an ODC Public Domain Dedication and Licence, with a material transfer agreement when applicable). Interested parties will thereby be allowed to access data directly, and they will give credit to the authors for the data used by citing the corresponding DOI.
We aim at communicating our results in top journals that require full disclosure upon publication of all included data, either in the main text, in supplementary material or in a separate data repository. All research outputs supporting publications will be made openly accessible at the latest at the time of publication (or preprint deposition) via the required link in the publication or upon reasonable request and after an embargo period after publication. Before the end of the embargo or in cases where sharing the post-print is not allowed due to copyright agreements, a pre-print version of the manuscript will be made available through our institutional repository, Lirias. Depending on their nature, some data may be made available prior to publication, either on an individual basis to interested researchers and/or potential new collaborators, or publicly via repositories (e.g. negative data). For data shared directly by the PI (and approval of the 3 party if necessary), a material/data transfer agreement (and a non-disclosure agreement if applicable) will be concluded with the beneficiaries in order to clearly describe the types of reuse that are permitted.
The work performed in this project used tumor samples from patients participating in clinical studies. In case of human personal data, only anonymized data will be published or deposited in open access repositories. In case of pseudonymized personal data, personal data will only be published after de-identification and identifiers will not be published. Pseudonymized data, or anonymized data containing unique data (e.g. raw sequencing data) will only be uploaded to a restricted access repository such as EGA, and may only be used by third parties where there is ethical approval, and where the data processing is in line with the patient consent and under data sharing agreement.
Data that will be made available after the end of the project:

- Double/triple-coded raw and processed sequencing data (linked to double-coded patient data) will be deposited in open access repositories with restricted access control such as the EBI European Genome-phenome Archive (EGA). The EGA is a repository for personally identifiable genetic and phenotypic data. Accession to these data will only be available upon reasonable request via our institutional data access committee (DAC) and if necessary (DAC) a data transfer agreement will be concluded with the beneficiaries in order to describe the types of reuse that are permitted and to include the necessary conditions to guarantee protection of personal data (according to European GDPR law). The double/triple-coded read count data matrix (linked to double/triple-coded patient data) will be available on our website (https://lambrechtslab.sites.vib.be/en/data-access). Note: Personal data will be double/triple coded and no reference to subject name will be made.
- All protocols used to generate published data will be described in the corresponding manuscript(s), and the related documentation will be included as supplementary information. These data and all other documents (raw data) deposited in the electronic laboratory notebook are accessible to the PI and the research staff and will be made available upon request.
- All the relevant algorithms, scripts and software tools will be described in manuscripts and/or on GitHub (https://github.com).
- The results will be published as BioRxiv preprints and/or as Open Access in peer reviewed journal.

If access is restricted, please specify who will be able to access the data and under what conditions.

Double/triple-coded raw and processed sequencing data (linked to double-coded patient data) will be deposited in open access repositories with restricted access control such as the EBI European Genome-phenome Archive (EGA). Accession to these data will only be available upon reasonable request via our institutional data access committee (DAC) and if necessary (DAC) a data transfer agreement will be concluded with the beneficiaries in order to describe the types of reuse that are permitted and to include the necessary conditions to guarantee protection of personal data (according to European GDPR law). The double/triple-coded read count

data matrix (linked to double/triple-coded patient data) will be available on our website (https://lambrechtslab.sites.vib.be/en/data-access).

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- Yes, Intellectual Property Rights
- Yes, Privacy aspects
- Yes, Ethical aspects

In general, personal data will only be published after de-identification and identifiers will not be published. If despite all efforts it is not possible to protect the identities of subjects even after removing all identifiers, personal data will not be made public.

In order to respect the patient's privacy, tumour samples will only be available to the research and technical staff involved in the project, not to other groups, studies or purpose, unless ethical approval is granted.

We aim at communicating our results in top journals that require full disclosure of all included data, or restricted access through a repository with appropriate access control (e.g. EGA). Additional material or information could be shared upon simple request following publication, unless we identify valuable IP, in which case we will first protect commercial exploitation, either through patenting or via an MTA that restricts the material from commercial use.

The work described in the current proposal is included as the translational research part of the clinical study. In the current proposal, we will use retrospective tumour samples collected in the ALBERT study at UZ Leuven. Therefore, we have a Material and Data Transfer Agreement (MTA) between the legal entities of KU Leuven (> UZ Leuven) and VIB.

Since the ALBERT study will only use historical samples collected for diagnostic reasons (=residuary material), no additional consent is needed. Patients will be informed that their sample will be used in this study through the MyNexusHealth app with a chance to opt out of the study.

Where will the data be made available? If already known, please provide a repository per dataset or data type.

Whenever possible, datasets and appropriate metadata will be made publicly available through repositories that support FAIR data sharing. Personal data will be double coded and no reference to subject name will be made. Sharing policies for specific research outputs are detailed below:

- Double/triple-coded raw sequencing data (linked to double-coded patient data) will be deposited in open access repositories with restricted access control such as the EBI European Genome-phenome Archive (EGA). The EGA is a repository for personally identifiable genetic and phenotypic data. Sequencing data at EGA will only be available upon reasonable request via our institutional data access committee and if necessary a material transfer agreement will be concluded with the beneficiaries in order to describe the types of reuse that are permitted. The double/triple-coded read count data matrix (linked to double/triplecoded patient data) will be available on our website (https://lambrechtslab.sites.vib.be/en/data- access).
- Double/triple-coded patient data: Upon publication, all double/triple-coded patient details supporting a manuscript will be made publicly available as supplemental information.
- Research documentation: All protocols used to generate published data will be described in the corresponding manuscript(s), and the related documentation will be included as supplementary information. These data and all other documents (raw data) deposited in the E-Notebook are accessible to the PI and the research staff and will be made available upon request.
- Manuscripts: All scientific publications will be shared openly. Manuscripts submitted for publication will be deposited in a pre-print server such as bioRxiv. At the time of publication, research results will be summarized on the PI's website (https://lambrechtslab.sites.vib.be/) and post-print pdf versions of publications will be made available there if allowed by copyright agreements, possibly after an embargo as determined by the publisher. Before the end of the embargo or in cases where sharing the post-print is not allowed due to copyright agreements, a pre-print version of the manuscript will be made available. (Pre-print) publications will also be automatically added to our institutional repository, Lirias 2.0, based on the authors name and ORCID ID.
- Algorithms, scripts and software: All the relevant algorithms, scripts and software toosls driving the project will be described in manuscripts and/or on GitHub (https://github.com) and/or on our interactive webserver (http://blueprint.lambrechtslab.org).
- Extra data that do not support publication will be either deposited in an open access repository or made available upon request by email. Data will be reused by transfer via Belnet Filesender or secure copy.

When will the data be made available?

As a general rule all research outputs will be made openly accessible at the latest at the time of publication (or preprint deposition). No embargo will be foreseen unless imposed e.g. by pending publications, potential IP requirements – note that patent application filing will be planned so that publications need not be delayed.

Which data usage licenses are you going to provide? If none, please explain why.

As detailed above, metadata will contain sufficient information to support data interpretation and reuse, and will be conform to community norms. These repositories clearly describe their conditions of use (typically under a Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication or an ODC Public Domain Dedication and License, with a material transfer agreement when applicable). Interested parties will thereby be allowed to access data directly, and they will give credit to the authors for the data used by citing the corresponding DOI. A CC-BY license will be opted for when possible. For data shared directly by the PIs (and approval of the 3rdparty if necessary), a material transfer agreement (and a non-disclosure agreement if applicable) will be concluded with the beneficiaries in order to clearly describe the types of reuse that are permitted.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

cfr. supra

What are the expected costs for data sharing? How will these costs be covered?

We minimized data management costs by implementing standard operating procedures (SOPs) e.g. for metadata collection and file storage and organization from the start of the project, and by using free-to-use data repositories and dissemination facilities whenever possible. Data management costs will be covered by own funding.
Usually, there is no charge for sharing data with third parties. For DTA of privacy-sensitive data, a quid pro quo in the form of co-authorship is usually requested.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

(Meta)data will be documented by the research and technical staff at the time of data collection and analysis, by taking careful notes in the (electronic) notebook that refer to specific datasets and by maintaining metadata sheets that preserve the connection between lab samples, sample and patient IDs, the specific names of the respective datasets and a link to the sequencing runs.

Who will manage data storage and backup during the research project?

The research and technical staff will ensure data storage and back up, with support from ICTS, gbiomed-IT staff, and UZ-IT staff. The project coordinator will regularly verify these protocols are followed.

Who will manage data preservation and sharing?

The PI, Diether Lambrechts, is in the end responsible for data preservation and sharing, with support from ICT of the of the Biomedical Sciences Group and his research team.

Who will update and implement this DMP?

The PI, Diether Lambrechts, bears the end responsibility of updating & implementing this DMP, supported by his research manager and his research team.