Paper DH14

Transcriptomics – A Demanding Data Analysis that Expands the Limits of Regulated Clinical Data Handling

Lea Yoshida Hildebrandt, Novo Nordisk, Søborg, Denmark

ABSTRACT

High-throughput molecular profiling aka "omics data" has been used for varied exploratory purposes recently within the pharmaceutical industry, maturing these technologies for use as endpoints in clinical studies. However, including novel data modalities puts new demands on data handling and infrastructure. At Novo Nordisk, we recently achieved a breakthrough in clinical development to include transcriptomics data in an interventional clinical study. The computational needs were significantly higher than those of our established computing environment. As a result, we split the data process into two parts, to partly run in an AWS high performance computing environment. We therefore tested combining CDISC with BioCompute, a non-CDISC standard supported by FDA. Analysis was based on R-packages developed by the scientific community – highlighting the need for fit-for-purpose architecture within clinical development. Our solutions initiate bridging between the world of exploratory studies with the regulatory demands for a primary endpoint in a clinical study.

INTRODUCTION

The bioinformatics technologies for drug discovery and understanding are continuously expanding, and omics is one area which within the last 10 years has seen a steep increase in accessibility and understanding. Omics covers a general field of research where all components of a certain biosample (ie. blood, cerebral spinal fluid, tissue sample) are measured simultaneously. The most famous omics technology is probably gen-omics – the study of whole genomes of organisms. Other major omics technologies to mention are proteomics, where all proteins of a certain biosample are measured simultaneously, and transcriptomics, the study of the transcripts that are expressed in each cell in the sample. The technologies also have synergy with each other, where multi-omics studies offer even stronger evidence generation and hypothesis testing[1].

This work is about a clinical study which includes transcriptomics and proteomics – as part of several endpoints included in the protocol. There were 3 omics modalities included in the study – two types of transcriptomics: single cell RNA sequencing (scRNAseq) and single cell TCR sequencing (scTCRseq); and proteomics. The Novo Nordisk strategy is to build internal competencies within omics technologies, which means that we handled the complete data journey – from raw sequencing data, until outputs for the clinical study report. The main focus of this work is on the scRNAseq, since this was documented in the protocol as the primary endpoint of the study[2]. scRNAseq is a technique where a sample containing cells is sequenced, with the resulting data containing the gene sequences for each individual cell. This usually means at least 20000 genes in about 3-4000 cells, and several hundreds of gigabites of data per processed sample.

Due to the cutting-edge nature of scientific studies of transcriptomics, we had to delicately balance high demands for reproducibility, traceability and compliance required for a clinical study with using recently published data analysis codebases. Suffice to say, the solutions were not covered by the existing operating procedures within our organization, and how to handle the process was worked out in parallel to the code development. We adopted a risk based approach inspired by the white paper published by the R® Consortium validation hub[3]. This means that we looked at individual parts of open-source software and evaluated it for both risk and impact, and then did internal review according to the conclusions. We also took guidance from the FAIR (Findable, Accessible, Interoperable and Reusable)[4], and worked on how to combine those principles with the clinical data standards that are used across Novo Nordisk clinical studies. More details about the good clinical practice aspect of our solution was reported by Adrian Czaban[5], so that will not be covered here in detail. This paper will focus on the technical implementations, considerations and outcomes that supported the successful inclusions of the omics data endpoints.

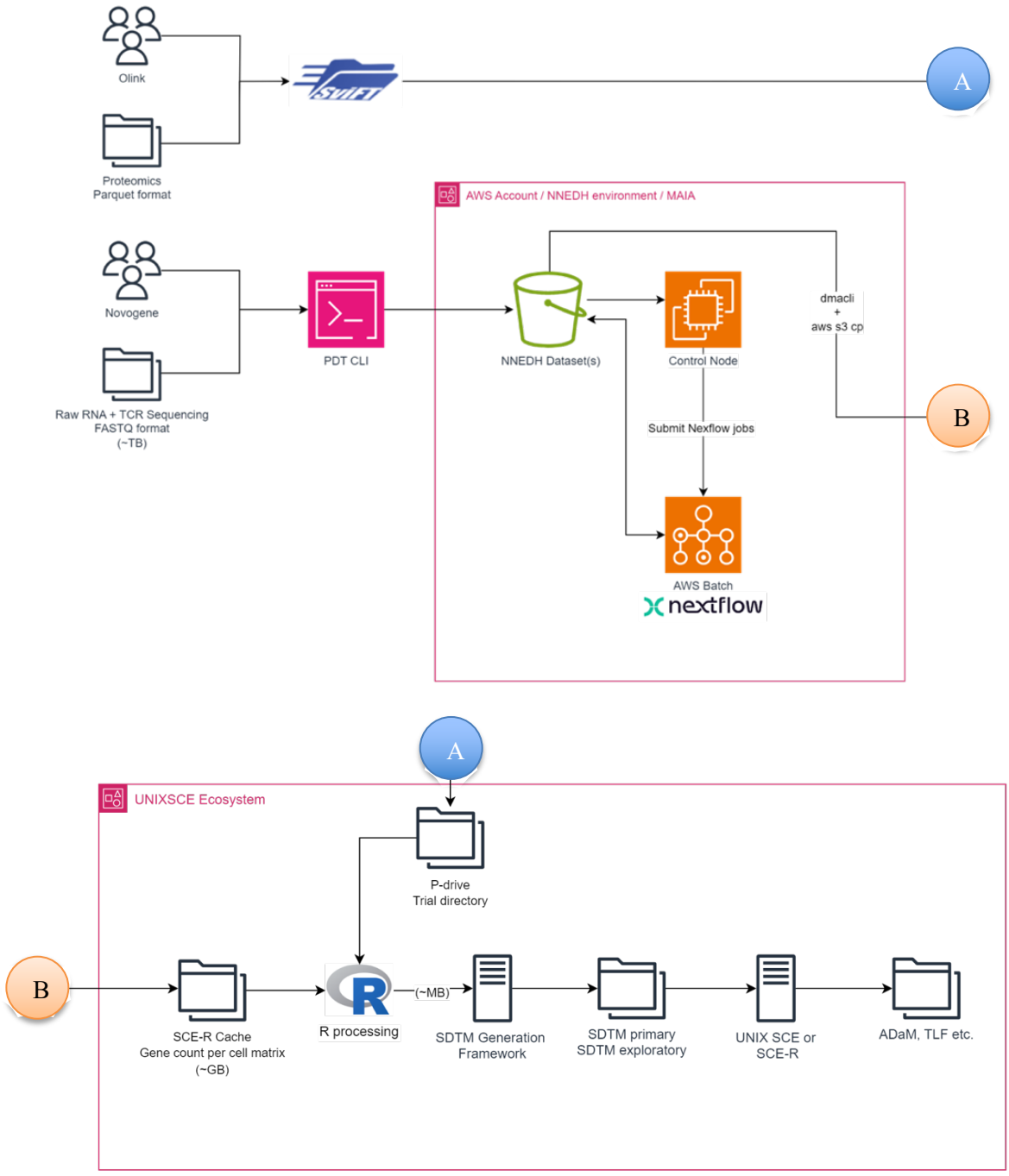

Figure 1 Map of the data infrastructure – beginning with delivery from two different vendors, proteomics going straight into SDTM, and transcriptomics being processed first in the high computing environment MAIA, then transferred to the SCE, UNIX-SCE.

IT INFRASTRUCTURE

The statistical computation environment (SCE) that is used for analysis and reporting of clinical data in Novo Nordisk could not handle the volume of data that we needed to analyze for this study. Therefore we used a high computing environment (MAIA) in combination with the SCE. MAIA is an IT solution consisting of data storage in Datahub combined with data processing running on Amazon Web Services® (AWS). Both the transfer and storage of the raw sequencing data, and preprocessing was done within this system. AWS has advanced functions for data transfer built in, which we used to develop a transfer client for the vendor to use when transferring the data to our systems. Figure 1 shows an overview of the two infrastructures, as well as the various components within.

PREPROCESSING IN AMAZON WEB SERVICES

The part of our data processing that was run in the AWS infrastructure, is performed using the NextFlow module core/scrnaseq, which is created specifically for the type of data that we are working on here. BioCompute objects is a data standard which is specifically developed to be used for the complex data structures that are used for bio-informatics pipelines. It therefore lends itself perfectly to be used for documenting the data processing performed within the MAIA infrastructure, and we decided to use it for documentation.

QUALITY CONTROL PIPELINE

Quality control, the next step in the data processing, was performed in the SCE, but the data volume exceeded the old school UNIX SCE infrastructure, so we decided to use a cache feature in the SCE-R infrastructure. Fig. 2 shows an overview that includes the quality control steps. This way we took advantage of the established procedures for SCE-R at the earliest opportunity in the data processing pipeline. Using built in functionalities of the AWS infrastructure, we created a transfer script which uses checksums to keep track of the data transfer, and outputs logs that can be used to check that the transfer has completed successfully.

The pipeline we developed for quality control of the scRNAseq data followed generally accepted best practice within the transcriptomics research field, using a combination of the R packages Seurat[7] and scCustomize[8]. We remove empty droplets using emptyDropsCellranger() and ambient RNA with soupX[9]. Then we filter out cells with abnormally high or low gene counts, removing any cells with less than 300 features. We also remove ones with high mitochondrial gene content. We use various sanity checks along the way to make sure the data looks right - Figure 3 (A) shows examples of these plots, these only give a general idea as they are generated using publicly available data, and artificially cut to be for two subjects for two visits. After filtering, we do dimension reduction using Principal Component Analysis (PCA) – first we identify the principal components exhibiting cumulative percent greater than 90% and where the % variation associated with the PC is less than 5%. Then we determine the difference between the variation each PC and subsequent PC and include the PCs where the change of % of variation is more than 0.1%. Fig. 3(B) shows an elbow plot of the PCs. We use scDblFinder[10] to remove suspected doubles – where two cells were analyzed simultaneously. Finally, the data is clean enough to be sorted into clusters. In this step, each measured cell is compared to each other and those with similarities are clustered together. We use Harmony[11] for this step, and the final result is saved as a Seurat object (*.rds). The filtering and harmonization reduce the data volume to less than 100 GB, which makes it possible to process using the existing SCE infrastructure.

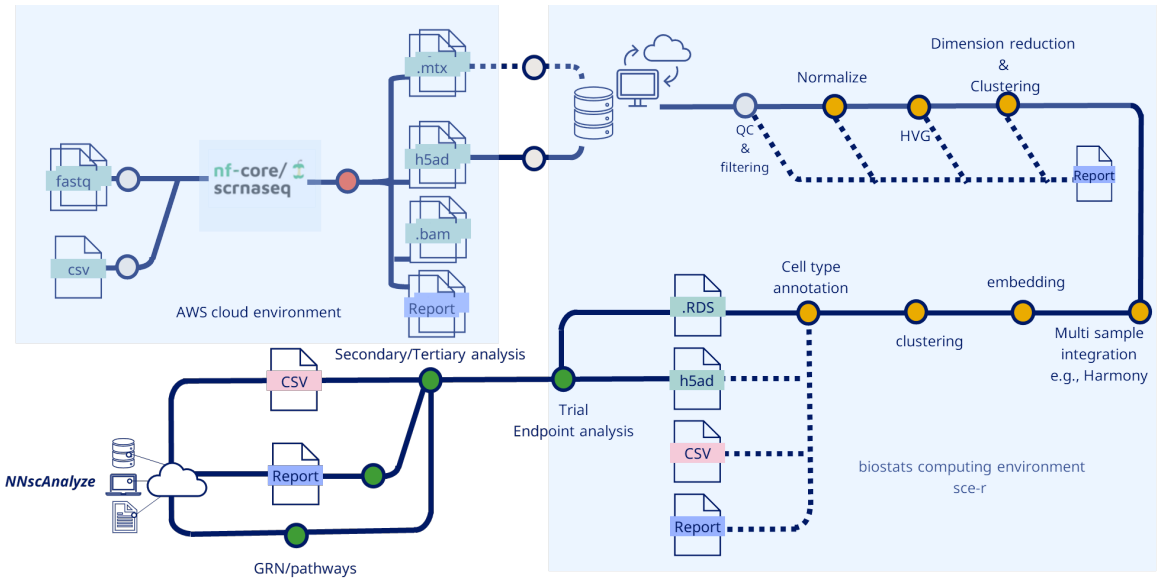

Figure 2: Metromap of the whole dataflow – from raw fastq to input to SDTM generation. AWS is shown top left, with the incoming data types moving through the Nextflow scRNAseq pipeline, and the preprocessed datafiles outputted that are then copied into the SCE on the right. There the various steps of the quality control pipeline are shown – Normalization, dimension reduction etc, until the "Trial Endpoint analysis" which produces the data that the biostatistician uses for the tables, figures and listings for the CSR.

The result of the pipeline was summarized as a count of the number of genes that were differentially expressed for each participant before and after treatment. This count is the primary endpoint described in the protocol for the study and should therefore be published in the clinical study report. Therefore, we created custom Study Data Tabulation Model (SDTM) domains to contain the data and to be used as input in the further statistical analysis. The resulting new SDTM domains are presented by Vicky Poulsen in the Standards Implementation stream of PHUSE EU Connect 2024[12] and will not be described in further detail in this paper.

DATA GOVERNANCE AND REPRODUCIBILITY

We wrote detailed data transfer specifications (DTS) together with the stakeholders responsible for SDTM generation and the data manager of the study. These documents are used as a mutual contract between data deliverer and data receiver and are standard practice for Novo Nordisk clinical studies. They are living documents, that are continuously updated until all the details of data delivery are clear and have been decided, and they are worked on collaboratively by both data receiver and data deliverer. The main content are basic details such as naming conventions, data format, data content descriptions, and responsible people on both sides. But they are an essential task to complete, to avoid any surprises upon data delivery. They are also important for staying GCP compliant, as they describe the data that is expected and are a measuring stick to compare the transfers to.

Taking inspiration from both the FAIR[13] principles as well as generally accepted documentation and traceability measures, we used GIT when developing the code we used, thereby documenting all decision made during development. Early in the project we considered which programming language to use, and chose R® and the Seurat package, due to its wide use within the transcriptomics community. To keep track of the package versions used in the project, we used the r-package renv, as well as a thorough readme describing the whole procedure for running the scripts in the pipeline. We are currently working on reordering the code from our pipeline together with a few visualizations as a releasable r-package, which we are hoping to be able to publish sometime next year.

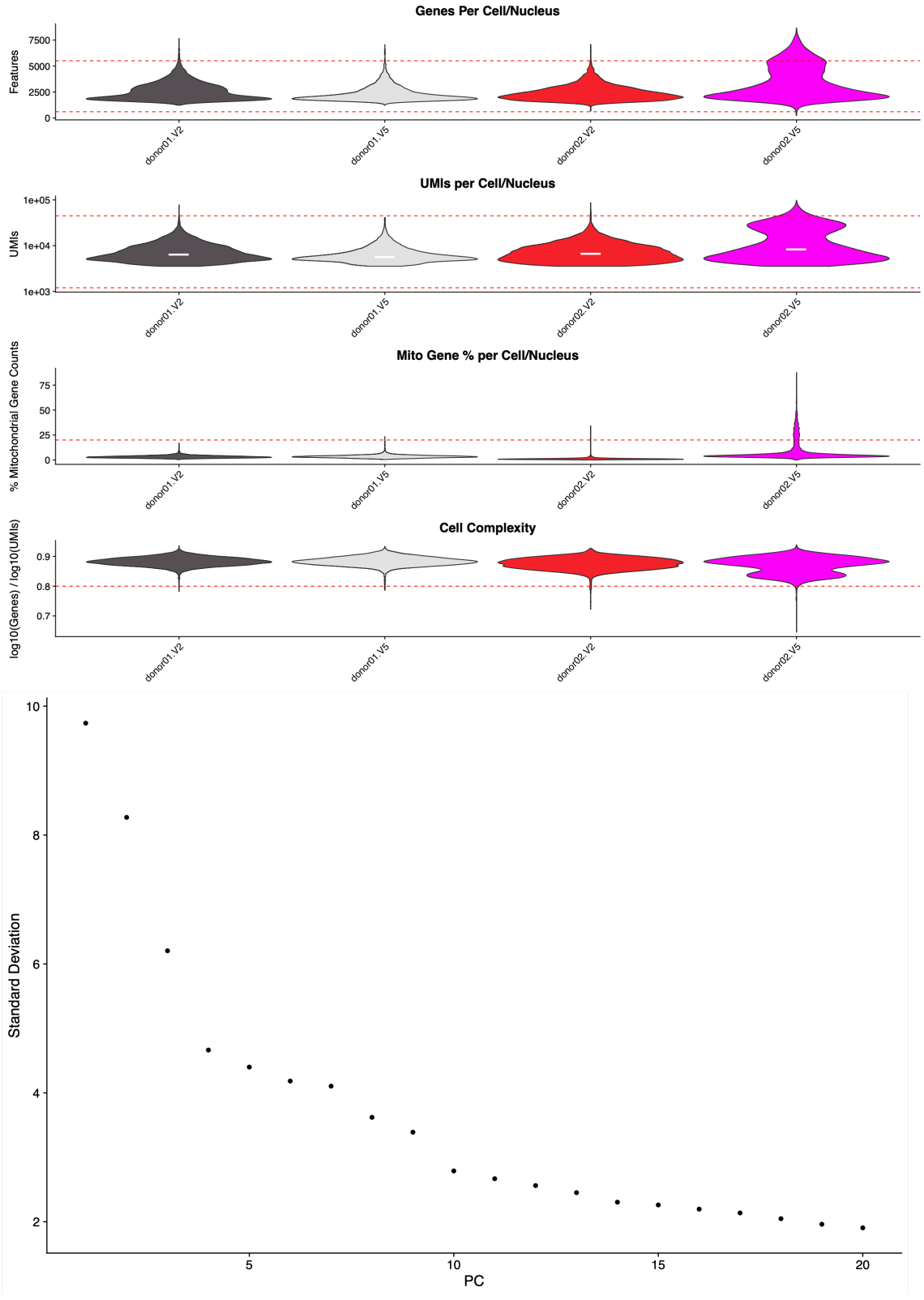

Figure 3: Example outputs from the quality control pipeline. (A – top left): Violin plot of the genes pr cell, UMIs pr cell, Mitochondrial genes pr cell, and cell complexity. (B – top right): The principal components vs the amount of the standard deviation associated with each. Disclaimer: The data shown here is publicly available data from 10Xgenomics® (https://www.10xgenomics.com/datasets/pbm-cs-from-a-healthy-donor-whole-transcriptome-analysis-3-1-standard-4-0-0).

CONCLUSION

As many resources are invested into the conduct of clinical studies, it is a priority to be following scientific state of the art when doing the data analysis and interpretation. Unfortunately, this has not been feasible within the current SCE, which is based on older IT infrastructure which is incompatible with most modern analysis pipelines. Therefore, one of the main conclusions from the effort going into the studies that are currently being conducted within Novo Nordisk that include these novel data types, is that investments into modernizing the IT compute and storage facilities for clinical data analysis is crucial.

In conclusion, we succeeded in the inclusion of two transcriptomics endpoints in the clinical study report. This was achieved by combining a novel IT infrastructure (MAIA) with the already published SCE-R. The success relied on a combination of skills from across several areas, data scientists, transcriptomics experts, data managers, data standards representatives and quality assurance. We created a quality control pipeline using the Seurat package, and designed two new SDTM domains to contain the endpoint data. All of these are steps towards omics data – and other complex data types – being included in submissions to authorities, enabling the use of these novel data types to benefit patients within many therapeutic areas.

REFERENCES

[1] Drouard, G., et al. (2023). "Longitudinal multi-omics study reveals common etiology underlying association between plasma proteome and BMI trajectories in adolescent and young adult twins." BMC Med 21(1): 508.
[2] https://www.clinicaltrialsregister.eu/ctr-search/trial/2022-003384-24/SE#A
[3] Andy Nicholls, “A Risk-based Approach for Assessing R package Accuracy within a Validated Infrastructure”,2023,link: https://www.pharmar.org/white-paper/
[4] Barker et al, “Introducing the FaIR Principles for research software”, 2022, https://doi.org/10.1038/s41597-022-01710-x
[5] Adrian Czaban, “TT02: Working with Omics Data – A Statistical Programmer’s Perspective”, PHUSE EU Connect Conference proceedings 2024
[6] Jonatan Selsing et al, “How Novo Nordisk built a modern data architecture on AWS”, link: https://aws.amazon.com/blogs/big-data/how-novo-nordisk-built-a-modern-data-architecture-on-aws/
[7] Hao, Y., Stuart, T., Kowalski, M.H. et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat Biotechnol 42, 293–304 (2024). https://doi.org/10.1038/s41587-023-01767-y
[8] Marsh SE (2021). scCustomize: Custom Visualizations & Functions for Streamlined Analyses of Single Cell Sequencing. https://doi.org/10.5281/zenodo.5706430. RRID:SCR_024675.
[9] Matthew D Young, Sam Behjati, SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data, GigaScience, Volume 9, Issue 12, December 2020, giaa151, https://doi.org/10.1093/gigascience/giaa151
[10] Germain P, Lun A, Garcia Meixide C, Macnair W, Robinson M (2022). “Doublet identification in single-cell sequencing data using scDblFinder.” f1000research. doi:10.12688/f1000research.73600.2.
[11] Korsunsky, I., Millard, N., Fan, J. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods 16, 1289–1296 (2019). https://doi.org/10.1038/s41592-019-0619-0
[12] Vicky Poulsen, “SI08: Shoehorning Multi-Omics Data into SDTM: Navigating Challenges and Solutions”, PHUSE EU Connect Conference proceedings 2024
[13] Barker, M., Chue Hong, N.P., Katz, D.S. et al. Introducing the FAIR Principles for research software. Sci Data 9, 622 (2022). https://doi.org/10.1038/s41597-022-01710-x

ACKNOWLEDGMENTS

The author would like to acknowledge that this work was a huge team effort, combining skills from several distinct skill areas. The TrialOmics team could not have done this without the support from data managers and SDTM experts, as well as IT infrastructure achitects.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Lea Yoshida Hildebrandt
Company: Novo Nordisk
Address: Vandtårnsvej 108, 2860 Søborg, Denmark
Work Phone: +4530753008
Email: lrnd@novonordisk.com
Website: novonordisk.com