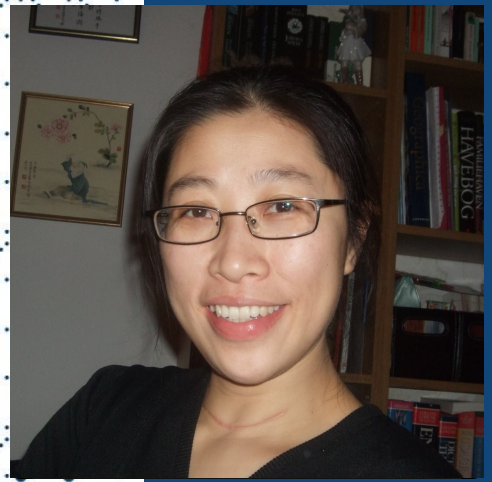
Fitting Multi-Omics Data into SDTM

Presented by Adrian Czaban, Principal Clinical Data Scientist, Novo Nordisk
Vicky Poulsen, Standards Director, Novo Nordisk



Meet the Speakers

Adrian Czaban

Title: Principal Clinical Data Scientist

Organization: Novo Nordisk A/S

Adrian Czaban is a Clinical Data Scientist with close to 10 years of experience. He has been involved in multiple submission projects in his career and is very interested in different new data types. His curiosity has led him to work with Omics data, and he is currently leading a Phuse Omics Project focused on technical aspects of handling those data in a Clinical Study setting.

Vicky Poulsen

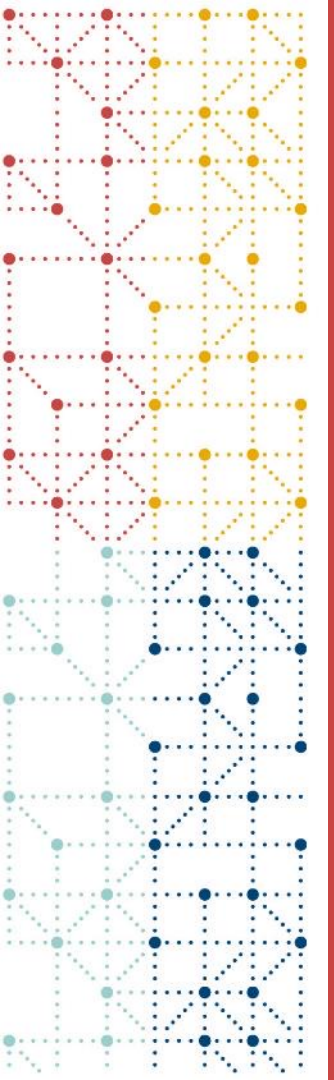
Title: Standards Director

Organization: Novo Nordisk A/S

Vicky Poulsen joined Novo Nordisk A/S in 2015 after being a SAS consultant for a decade in the public health sector specialising in End-to-End BI solutions. In her present role as strategic expert, besides being responsible for the operational part of the SDTM strategy, Vicky also acts as business solution architect across systems involved in the clinical data flow and helps define the strategic direction of related initiatives.

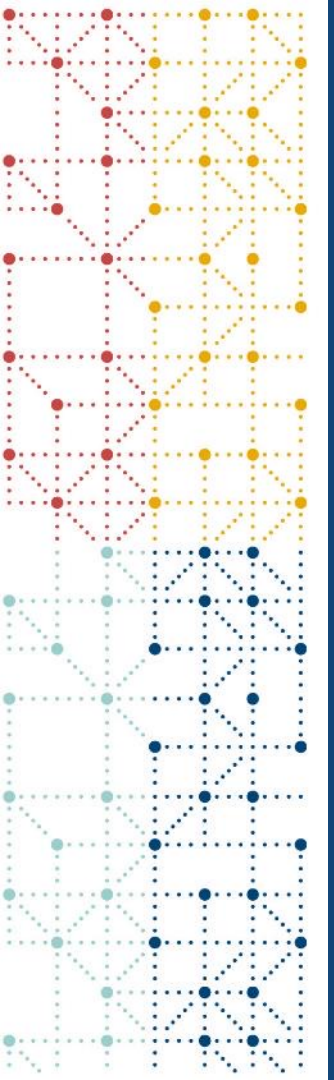
Disclaimer and Disclosures

- *The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of CDISC and Novo Nordisk A/S.*

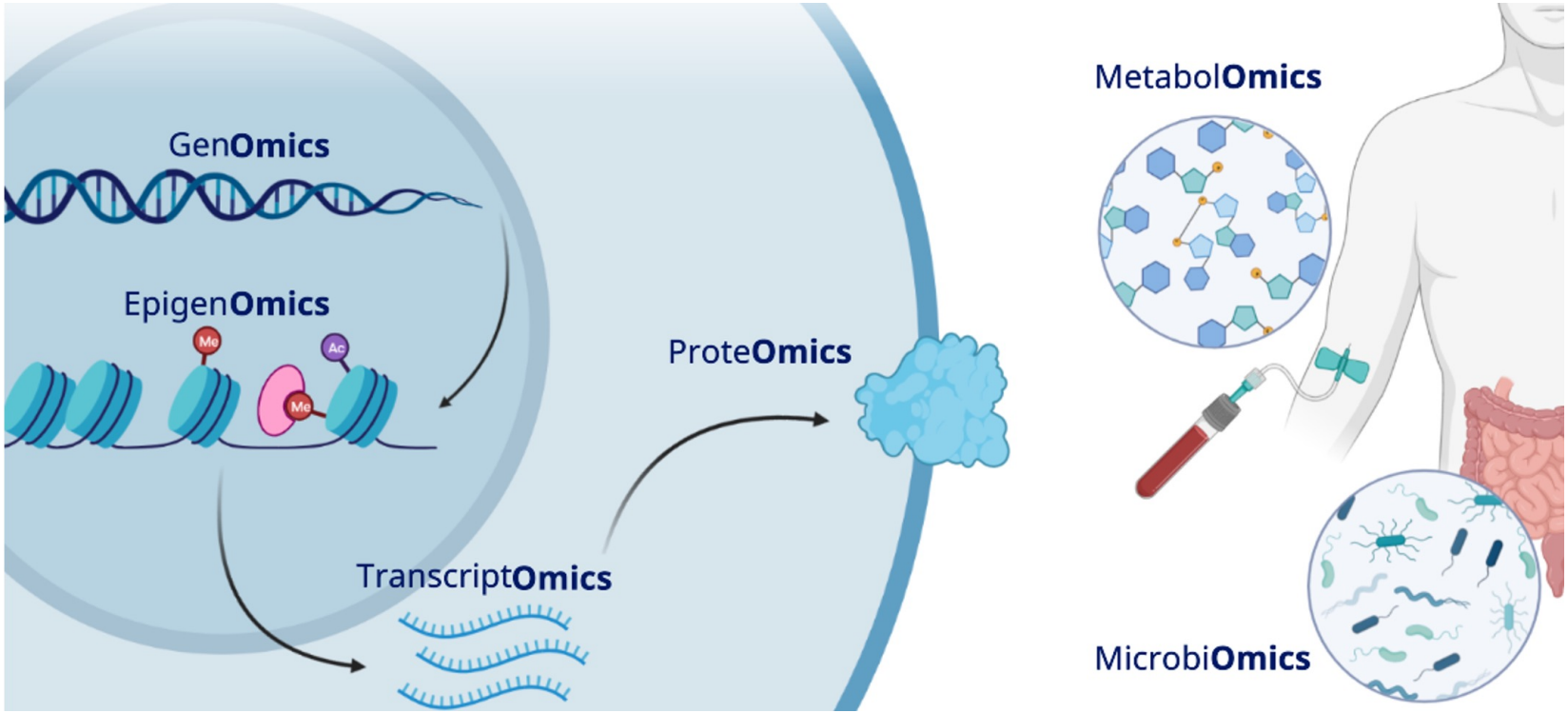


Agenda

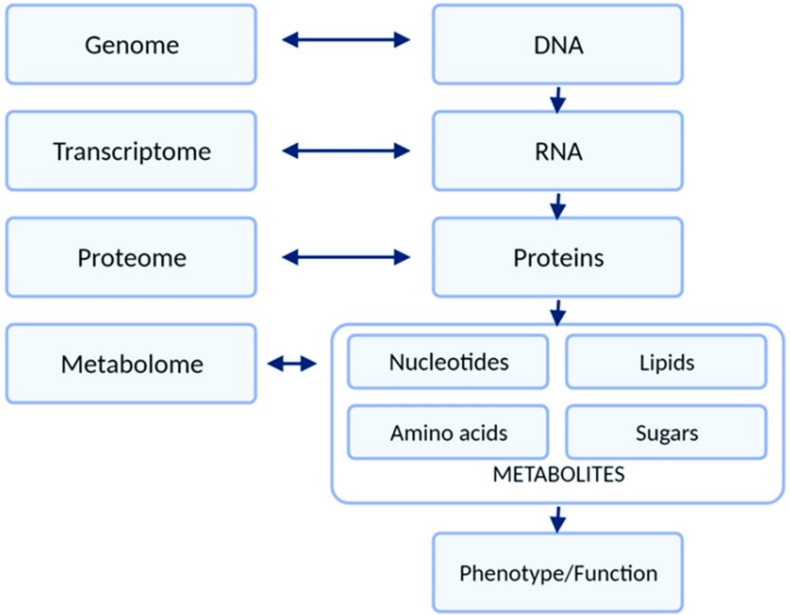
1. What is Omics Data?
2. Current Landscape
3. Fitting Multi-Omics Data into SDTM
4. Opportunities going forward



What is Omics Data?


GenOmics
EpigenOmics
Me
Ac
Me
ProteOmics
TranscriptOmics
MetabolOmics
MicrobiOmics

Multi - Omics



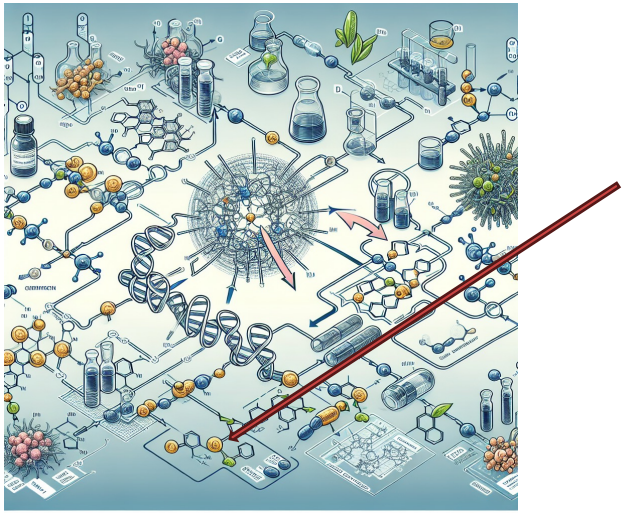
Pietkiewicz, Dagmara & Klupczynska, Agnieszka & Plewa, Szymon & Misiura, Magdalena & Horała, Agnieszka & Miltyk, Wojciech & Nowak-Markwitz, Ewa & Kokot, Zenon & Matysiak, Jan. (2021). Free Amino Acid Alterations in Patients with Gynecological and Breast Cancer: A Review. Pharmaceuticals. 14. 731. 10.3390/ph14080731.

What opportnities does it bring?

First Letter ↓	Second Letter U	Second Letter C	Second Letter A	Second Letter G	Third Letter ↓
U	phenylalanine	serine	tyrosine	cysteine	U
	phenylalanine	serine	tyrosine	cysteine	C
	leucine	serine	STOP	STOP	A
	leucine	serine	STOP	tryptophan	G
C	leucine	proline	histidine	arginine	U
	leucine	proline	histidine	arginine	C
	leucine	proline	glutamine	arginine	A
	leucine	proline	glutamine	arginine	G
A	isoleucine	threonine	asparagine	serine	U
	isoleucine	threonine	asparagine	serine	C
	isoleucine	threonine	lysine	arginine	A
	methionine & START	threonine	lysine	arginine	G
G	valine	alanine	aspartate	glycine	U
	valine	alanine	aspartate	glycine	C
	valine	alanine	glutamate	glycine	A
	valine	alanine	glutamate	glycine	G

Precision Medicine

Targeted drug discovery



Both Primary and Secondary use

Size

Varies according to modality. Proteomics and metabolomics datasets are big, but manageable (GB sizes). Genetics and transcriptomics can easily reach multiple terabytes

Complexity

The amount of information available in source data is quite astounding – we are not measuring selected parameters, we are measuring all that there is to measure

The data formats require deep understanding in order to perform QC and work

FASTA

```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATTTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT
```

FASTQ

```
@ERR000589.41 EAS139_45:5:1:2:111/1
CTTTCCTCCCTGCTTTCCTGGCCCCACCATTTCCAGGGAACATCTTGTCAT
+
3IIIIIIIIIIIII>1IIIFF9BG08E00I%IG+&?(4)%00646.C1#&(
@ERR000589.42 EAS139_45:5:1:2:1293/1
AGTTGTTAAAATCCAAGCCAATTAAGATAGTCTTATCTTTTTAAAAGAAAT
+
IIIIIGII.AIIII=?I9G-/II=+I=4?761BA2C9I+5A711+&>1$/I
```

SAM

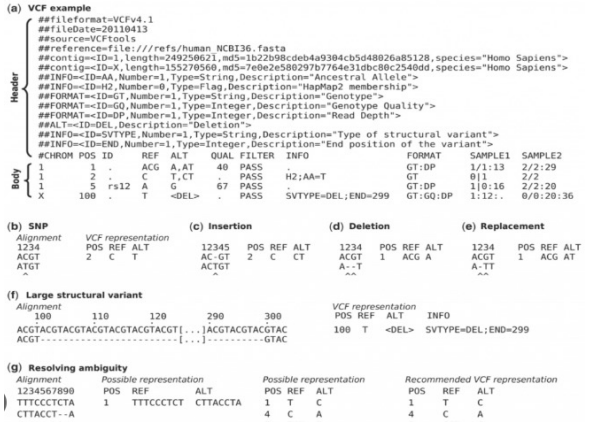
```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
        CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C       BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDCCDCDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
        AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961        16      chr20   193953  50      100M    *       0       0
        TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G       DCDDDDEDDDDDDDCDDDDDDDCCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFCCC
```

BED

##reference=file:///group/difazio/populus/popv3/Populus_trichocarpa_with_chloroplast_and_mitochondria.fa
##source=SelectVariants

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	201782_400	201782_400	201782_400	201782_40(
Chr01	67	.	C	A	7082.01	VQSRTranch	AC=92;AF=0.	GT:AD:DP:GC	./.:1,0:1	0/0:6,0:6:3:	0/0:16,0:16:	0/0:6,0:6:0:
Chr01	92	.	A	T	12084.37	VQSRTranch	AC=92;AF=1.	GT:AD:DP:GC	./.:2,0:2	./.:0,0:0	./.:0,0:0	./.:1,0:1
Chr01	95	.	C	T	12341.46	VQSRTranch	AC=92;AF=0.	GT:AD:DP:GC	./.:2,0:2	./.:0,0:0	./.:0,0:0	./.:1,0:1
Chr01	102	.	C	A	12939.33	VQSRTranch	AC=92;AF=1.	GT:AD:DP:GC	./.:3,0:3	./.:0,0:0	./.:0,0:0	./.:0,0:0
Chr01	103	.	C	A	12984.32	VQSRTranch	AC=92;AF=1.	GT:AD:DP:GC	./.:3,0:3	./.:0,0:0	./.:0,0:0	./.:0,0:0
Chr01	109	.	C	A	13524.27	VQSRTranch	AC=92;AF=1.	GT:AD:DP:GC	./.:4,0:4	./.:0,0:0	./.:0,0:0	./.:0,0:0
Chr01	117	.	C	A	14514.24	VQSRTranch	AC=92;AF=1.	GT:AD:DP:GC	./.:7,0:7	./.:0,0:0	./.:0,0:0	./.:0,0:0
Chr01	132	.	C	G	247.44	VQSRTranch	AC=2;AF=1.5	GT:AD:DP:GC	0/0:8,0:8:24:	./.:0,0:0	./.:0,0:0	./.:0,0:0

VCF

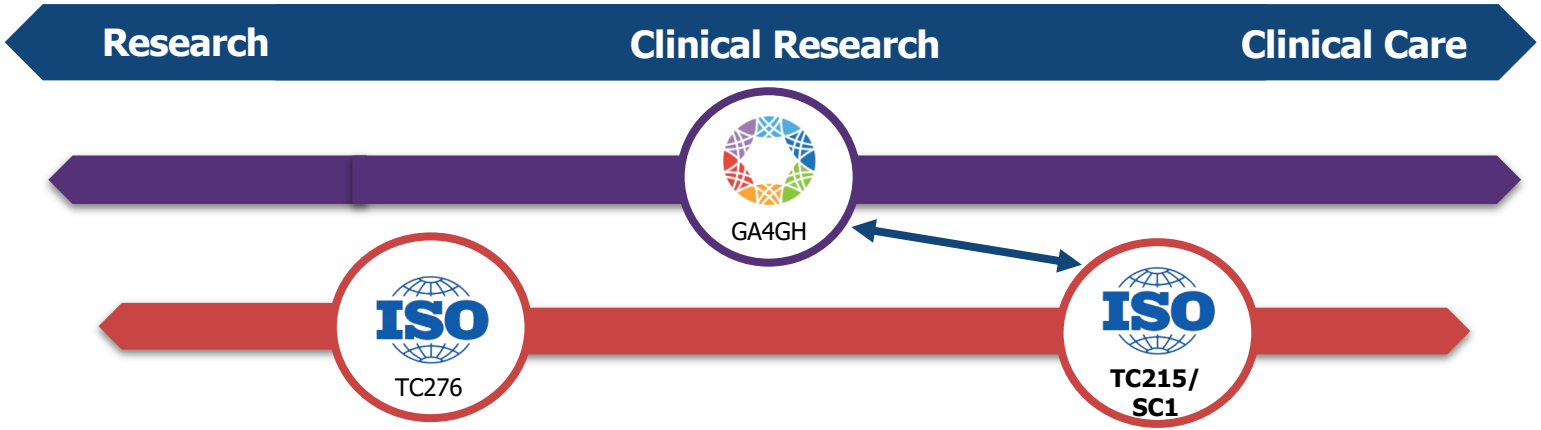


ISO TC/215 SC1 Genomics Informatics

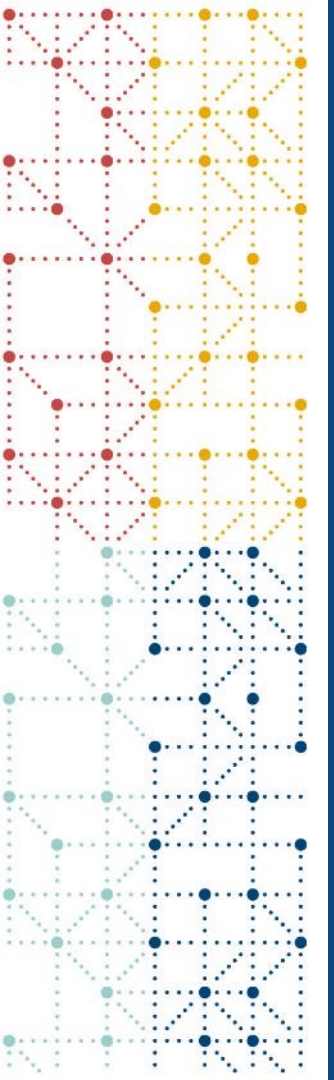
Standardization of computable data, information, and knowledge, including the representation and metadata for the application of omics – including but not limited to genomics, phenomics and transcriptomics to support human health and clinical research

Note: - SC1 moving to ”Omics” May 2025; 17 countries participating
- CDISC has been involved in ISO/TC 215 since 2016

Cross-SDO Coordination



ISO SC1 + GA4GH Joint Development
- Phenotypes (published)
- Pedigree & Family History (new project)
- Verifiable Credentials (future)
- Quality Control Metrics for DNA Sequencing (future)



Fitting Multi-Omics Data into SDTM

A Case Study

Trial objective:
Investigate drug-induced changes in immune cell signatures in cerebrospinal fluid (CSF) and blood in patients with early AD

Endpoints
- Primary: Change in gene expression assessed by single-cell sequencing (cells in CSF and blood)
- Exploratory: Change in T cell clonal landscape assessed by single-cell T-cell receptor sequencing (cells in CSF and blood)
- Exploratory: Changes in proteome (cells in CSF and plasma)

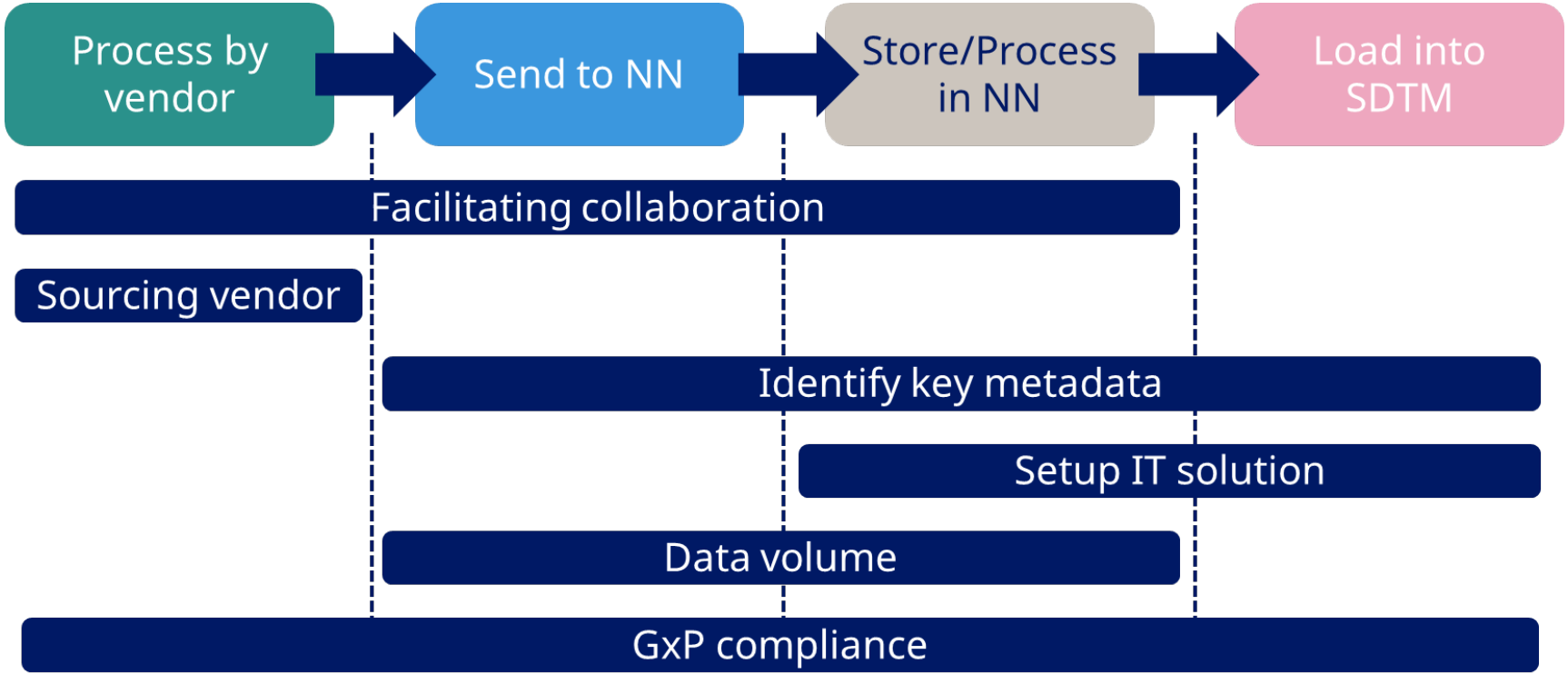
Cerebrospinal fluid samples

- Single-cell transcriptomics
- Proteomics

Blood samples

- Single-cell transcriptomics
- Proteomics

Data journey



Fitting into to SDTM

What to be included?

Which domain to use?

How to handle large data volume?

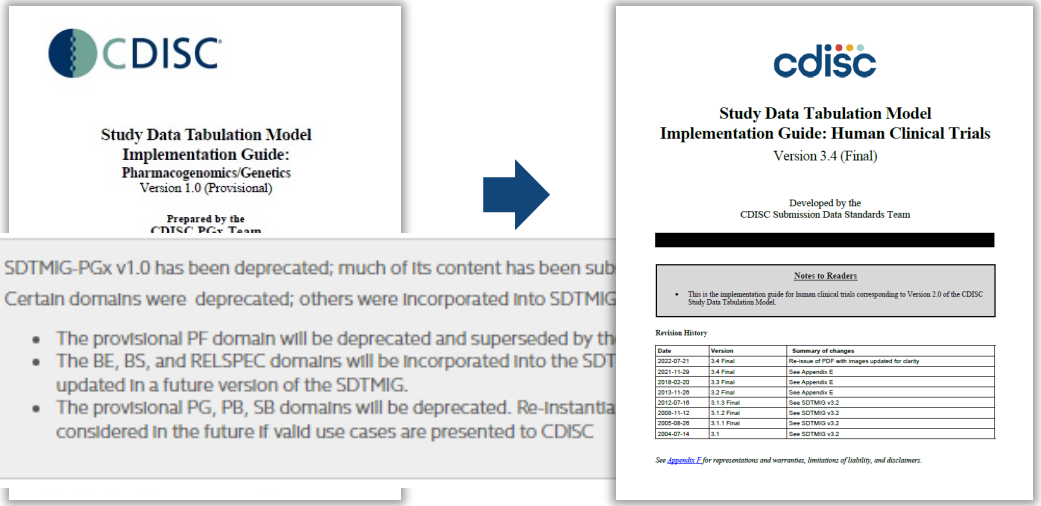
What to be included?

Case by case evaluation

- Omics type
- Sample type

- Omics findings
  - Granularity
  - Scope

Which domain to use?



Omics Findings (XO)

- NN Custom domain
- One record per finding per specimen per subject
- Model after MI domain

GF – Assumptions

1. The Genomics Findings domain is used to represent findings related to the structure, function, evolution, mapping, and editing of subject and non-host organism genomic material of interest. This domain includes but is not limited to assessments and results for genetic variation and transcription, and summary measures derived from these assessments. The GF domain is used for findings from characteristics assessed from nucleic acids and may include subsequent inferences and/or predictions about related proteins/amino acids. However, direct assessments of proteins (e.g., assessments of amino acids) are out of scope for this domain.

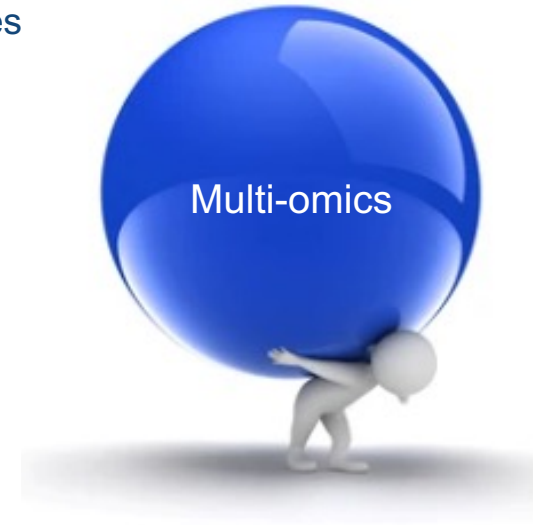
How to handle large data volume?

OMICS data could potential become voluminous

- There are about 20 000 human protein-coding genes that produce 75 000 to 100 000 proteins

Loading large volume of OMICS findings into SDTM poses some challenges

- Long processing time and storage
  - Burdens CDMS and SCE
  - Delayed availability of updated SDTM data
- Hampers usage when placed with other data within the same domain
  - Computing resources required for data filtering
  - Worsens when supplemental qualifiers are involved



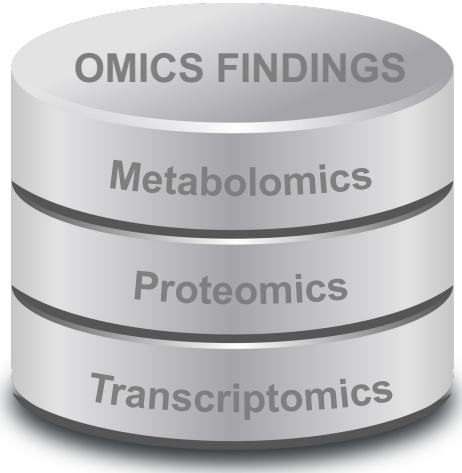
NN's Approach – *Stack* Dataset Concept

- A general observation domain can be split into separate datasets by --CAT
- All datasets will have the same structure of the domain to ensure data can be "stacked" back together, when necessary
- **Not the same as the split requirement (when dataset is above 5 GB) in FDA's TCG**

4.1.7 Splitting Domains

Sponsors may choose to split a domain of topically related information into physically separate datasets.

- A domain based on a general observation class may be split according to values in --CAT. When a domain is split on --CAT, --CAT must not be null.
- The Findings About (FA) domain (Section 6.4.4, Findings About) may alternatively be split based on the domain of the value in --OBJ. For example, FACM would store Findings About CM records. See Section 6.4.2, Naming Findings About Domains, for more details.



- *Stack* dataset concept can be expanded when need arises
  - XOP: Proteomics Findings
  - XOT: Transcriptomics Findings
- Governed centrally by NN's Global Standard Team

Opportunities going forward

Need for new/more domains

- Proteomics, Metabolomics, Epigenomics, Microbiomics
- Multi – Omics domains?

Interplay with other standards, such as BioCompute

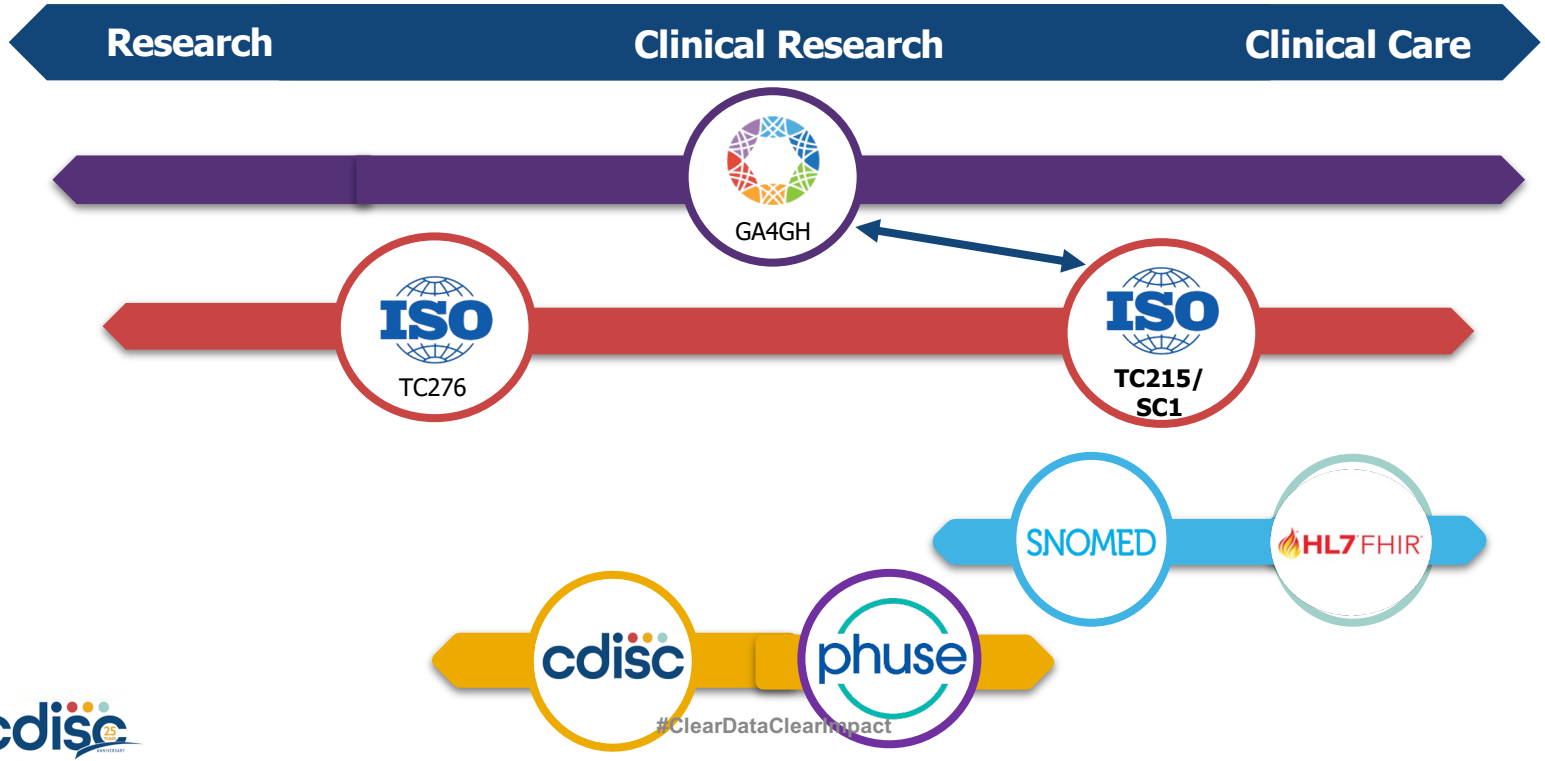
Phuse Omics Project initiative

WORKING GROUPS / **Integration of Omics Data into Clinical Drug Development**

Integration of Omics Data into Clinical Drug Development

NN LM Owned by Nicola Newton •••
Last updated: Mar 11, 2025 • Legacy editor

Cross-SDO Coordination



Acknowledgements

- Bron Kisler, chair of ISO/TC 215/SC 1 Genomics Informatics for consultations and help with the slides

Thank You!

Adrian Czaban (adcz@novonordisk.com)

Vicky Poulsen (vicp@novonordisk.com)