

DH04: Framework for Integrating Omics Data with Other Clinical and Non-Clinical Data Sources

Adrian Czaban–Novo Nordisk
Mehdi Harek–Bristol Myers Squibb

PHUSE
EU CONNECT 2025



Introduction

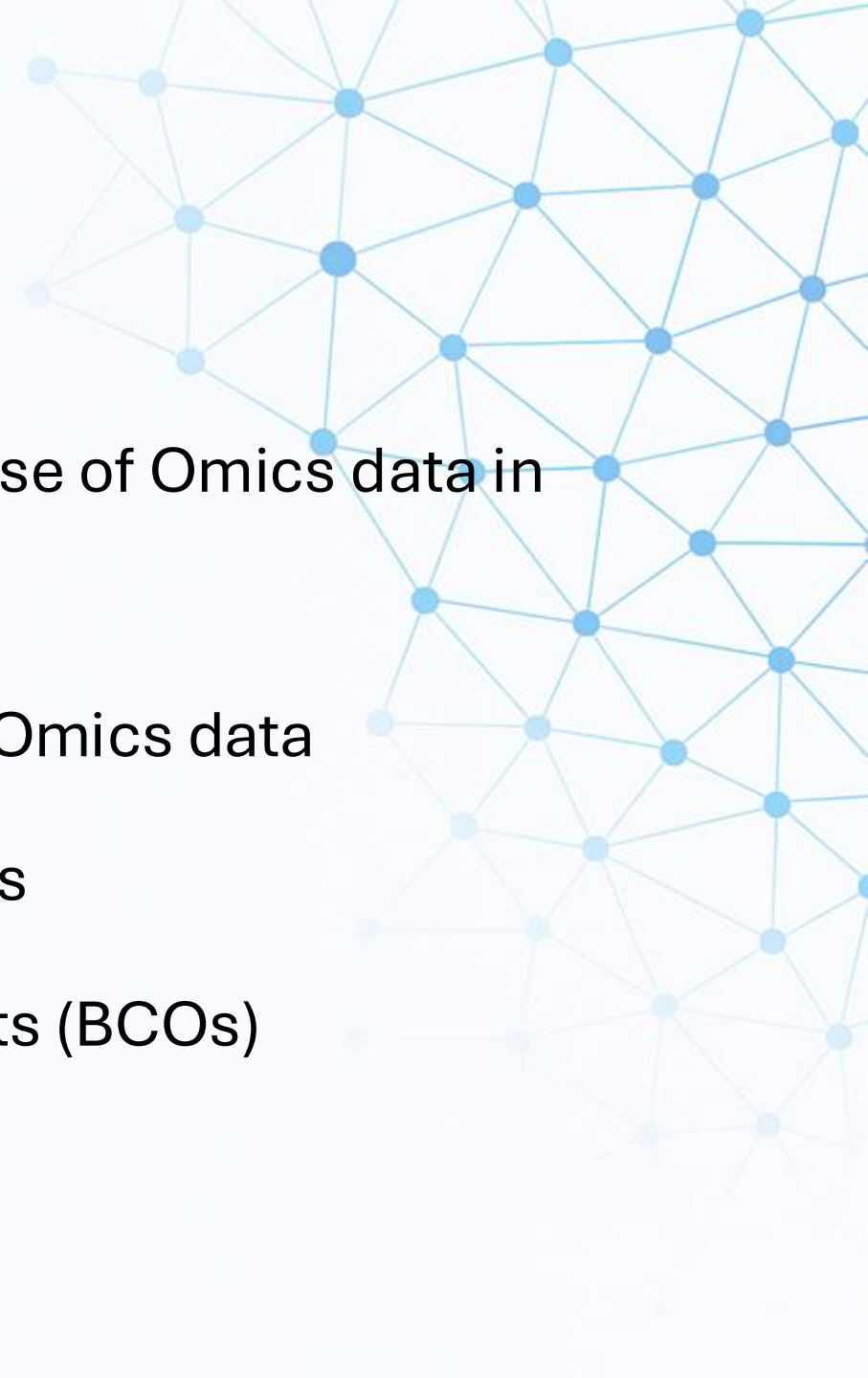
- Phuse WG: *“Integration of Omics data into Clinical Drug Development”*
- Cross pharma and research
 - Novo Nordisk
 - Bristol Myer Squibb
 - George Washington University
 - Independent Global and Genomics Standards Expert



Introduction

Goals:

- Initiating and driving the discussion regarding the use of Omics data in Clinical Development
- Establishing the best practices when working with Omics data
- Guide on creation and validation of Omics pipelines
- Submission considerations for BioCompute Objects (BCOs)
- Sharing of BCOs across the industry



Framework for Integrating Omics Data with Other Clinical and Non-Clinical Data Sources

A decorative background graphic on the right side of the slide, consisting of a network of blue dots connected by thin blue lines, resembling a molecular or data network.

More omics analysis within protocols

- Submission including Omics data
- Perform personalized medicine

Current standards do not fully address our needs

- Promote Interoperability
- Facilitate Technical Integration
- Share Best Practices

Framework for Integrating Omics Data with Other Clinical and Non-Clinical Data Sources



Omics data

Genomics (DNA)
Transcriptomics (RNA)
Proteomics
Metabolomics
Epigenomics



Clinical Trial Data

Structured data collected under controlled conditions



Clinical Data

Health-related information collected during routine medical care (diagnoses, treatments)

Recorded in electronic health records (EHRs)



Non-Clinical Data

Data *not* collected as part of conventional clinical trial CRFs or EHRs

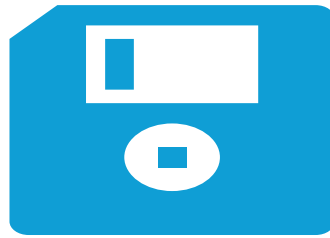
Problem Statement



Heterogeneity
of omics data
formats



Data quality
and
reproducibility
across
platforms



Large-scale
data storage



Complexity of
bioinformatics
analysis



Secondary
use of
genomics /
omics

Heterogeneity of omics data formats



- Complicated, often bespoke data processing solutions
- Can include correlation, covariance-based techniques, matrix decomposition, probabilistic, and/or Bayesian strategies

Data quality and reproducibility across platforms



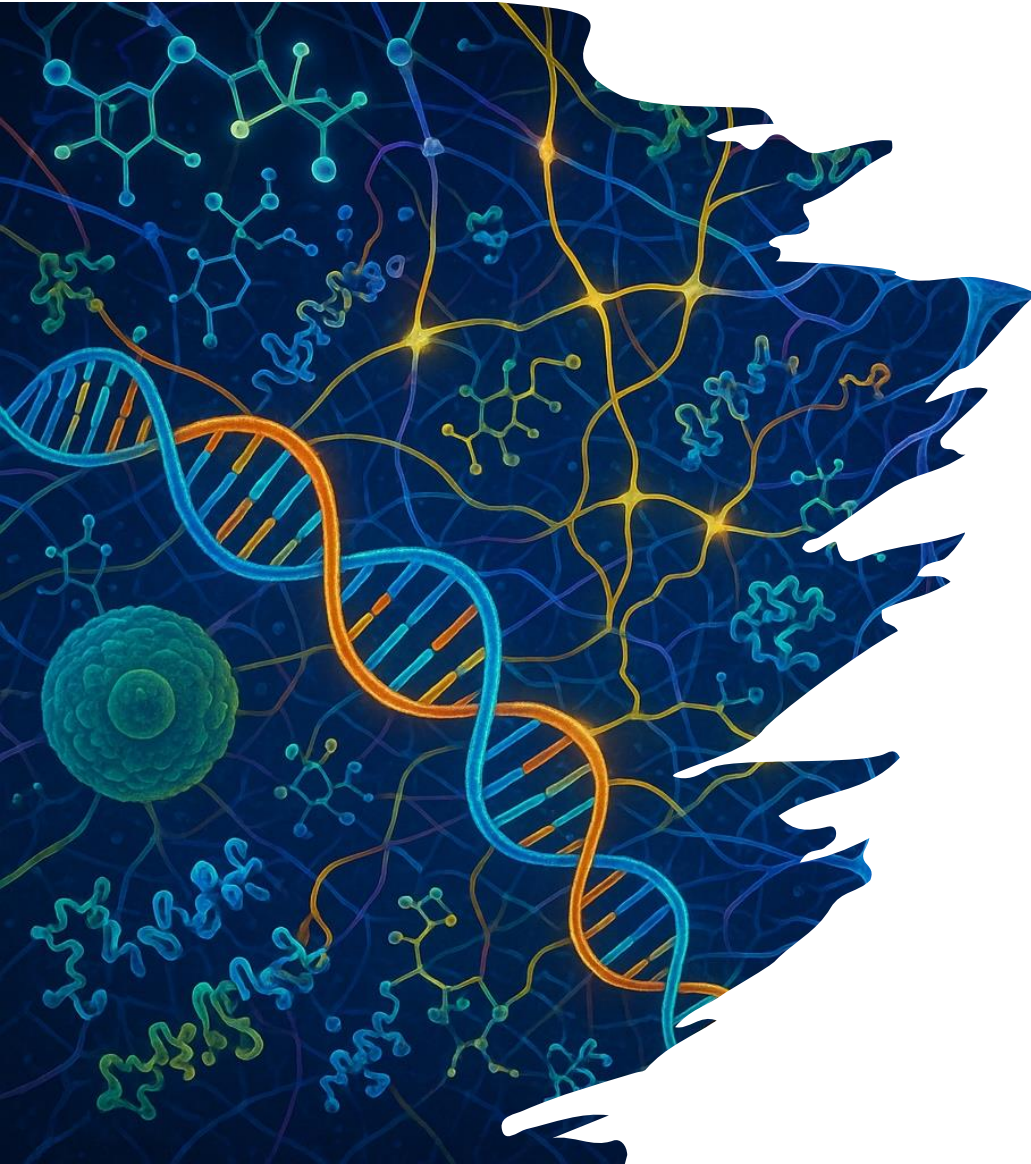
- Reflection of one's confidence that a particular piece of data represents reality. The way in which that confidence is represented can vary by data type.
- Data models for assessing quality are highly divergent by resource. There is no single, unifying confidence metric that can be applied across all data types

Large-scale data storage

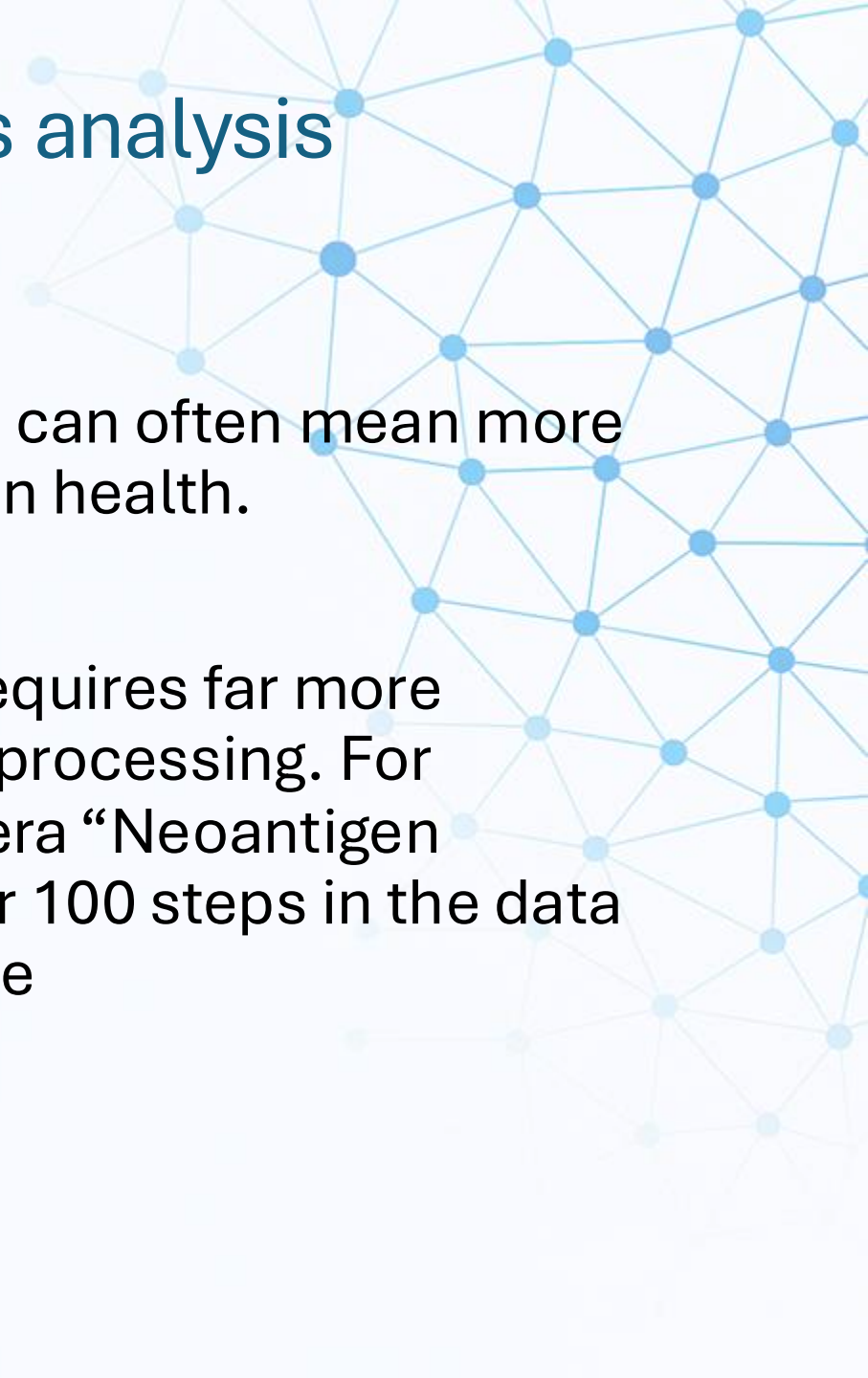


- Modern Next Generation Sequencing techniques easily generate raw data in the size of tens of terabytes
- Even if the traditional systems used by Pharma companies or CROs can store those amounts of data, they cannot process it in a reasonable way

Complexity of bioinformatics analysis



- More kinds of data can often mean more insights into human health.
- However, it also requires far more complicated data processing. For example, the Velsera “Neoantigen workflow” has over 100 steps in the data processing pipeline



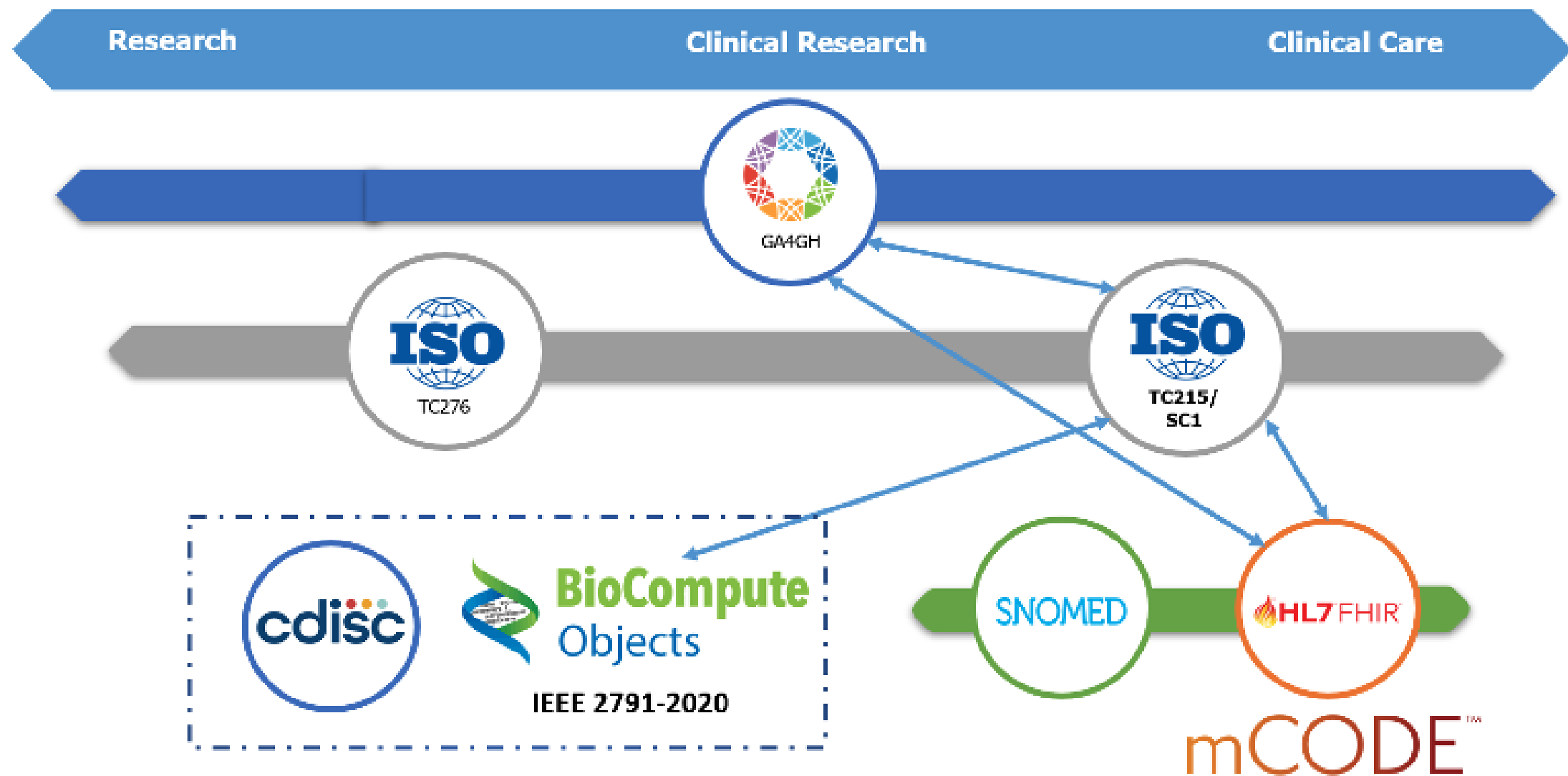
Secondary use of genomics / omics

SECONDARY USE OF CLINICAL DATA

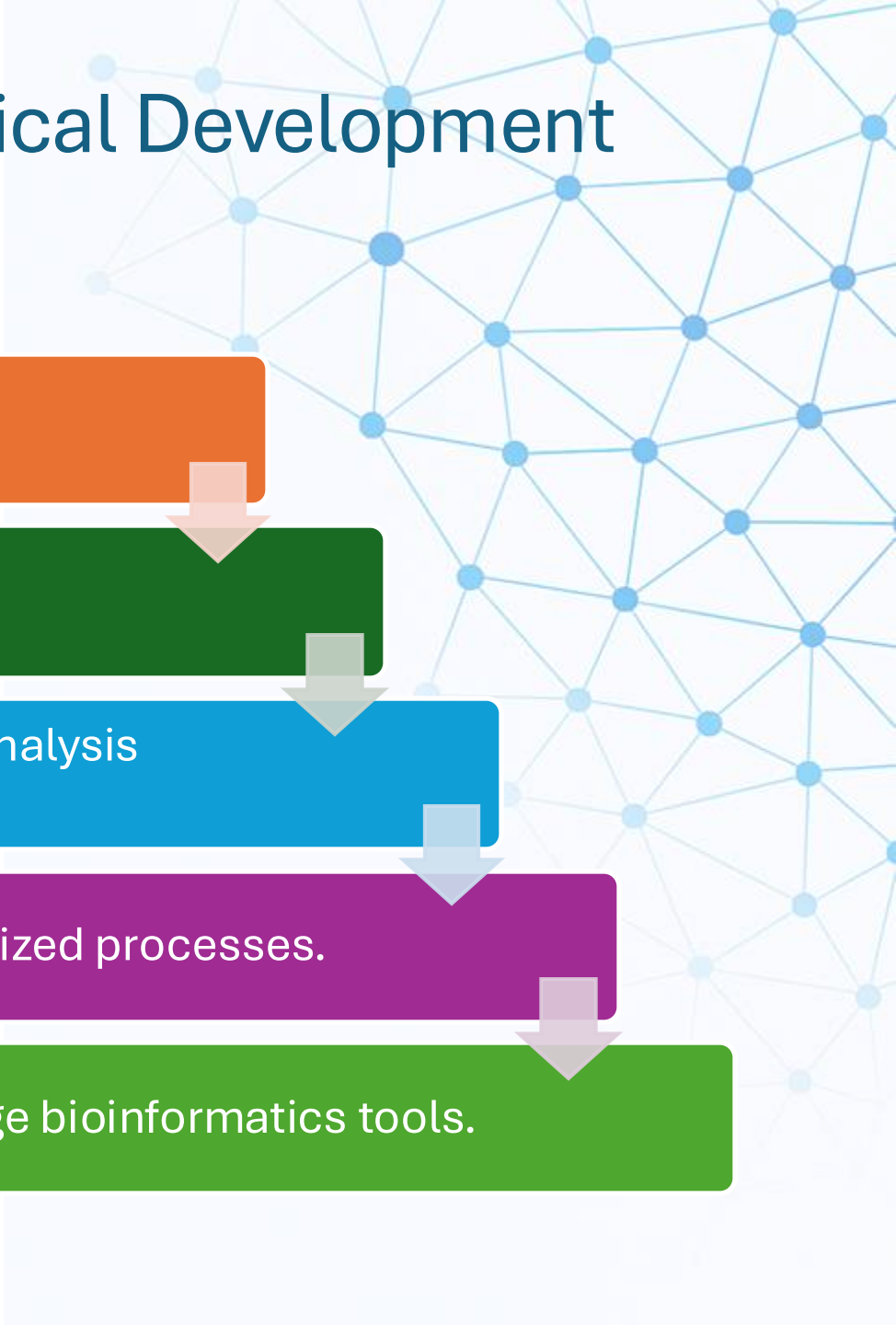


- Secondary use identified as a priority
- Different standards and frameworks make re-use difficult.
- Legal considerations remain an important barrier

Evolving Data Standards Landscape



Needs of new approaches in Clinical Development



Adopting innovative workflows to enhance omics data integration.

Rapid implementation of external standards to improve compatibility.

Fostering agility in clinical and non-clinical data analysis methods.

Encouraging collaboration through standardized processes.

Driving modernization with cutting-edge bioinformatics tools.

Results and Solutions - BioCompute

<pre>"created": "2020-08-01T16:29:55-04:00", "modified": "2020-08-01T16:29:55-04:00", "contributors": [{ "name": "Josiah Carberry",</pre>	Provenance
<pre>] "Identify baseline single nucleotide polymorphisms (SNPs)[SO:0000694], (insertions)[SO:0000667], and (deletions)[SO:0000045] that correlate with reduced antiviral drug efficacy in (Hepatitis C virus subtype 1)[taxonomy:31646]. Identify treatment emergent amino acid (substitutions) [SO:1000002] that correlate with</pre>	Usability
<pre>"pipeline_steps": [{ "step_number": 1, "name": "HIVE-Hexagon", "version": "2.4.3",</pre>	Description
<pre>"input_subdomain": [{ "uri": { "filename": "Hepatitis C virus genotype 1", "uri": "https://hive2.biochemistry.gwu.edu/dna.cgi?cmd=objFile&ids=16130"</pre>	IO
<pre>{ "script_driver": "HIVE", "software_prerequisites": [{</pre>	Execution
<pre>{ "param": "_type", "value": "svc-align-hexagon", "step": "1"</pre>	Parametric
<pre>"algorithmic_error": { "AthenaFRSCORE_threshold": 0.5, "AthenaQUALITY": 25, "AthenaCOVERAGE": 5000</pre>	Error
<pre>{ "extension_schema": "https://www.w3id.org/biocompute/extension_domain/1.1.0/dataset/dataset_extension.json", "dataset_extension": { "additional_license": {</pre>	Extension

Differences between BioCompute and CDISC



No requirement on formats



No requirements on Tools

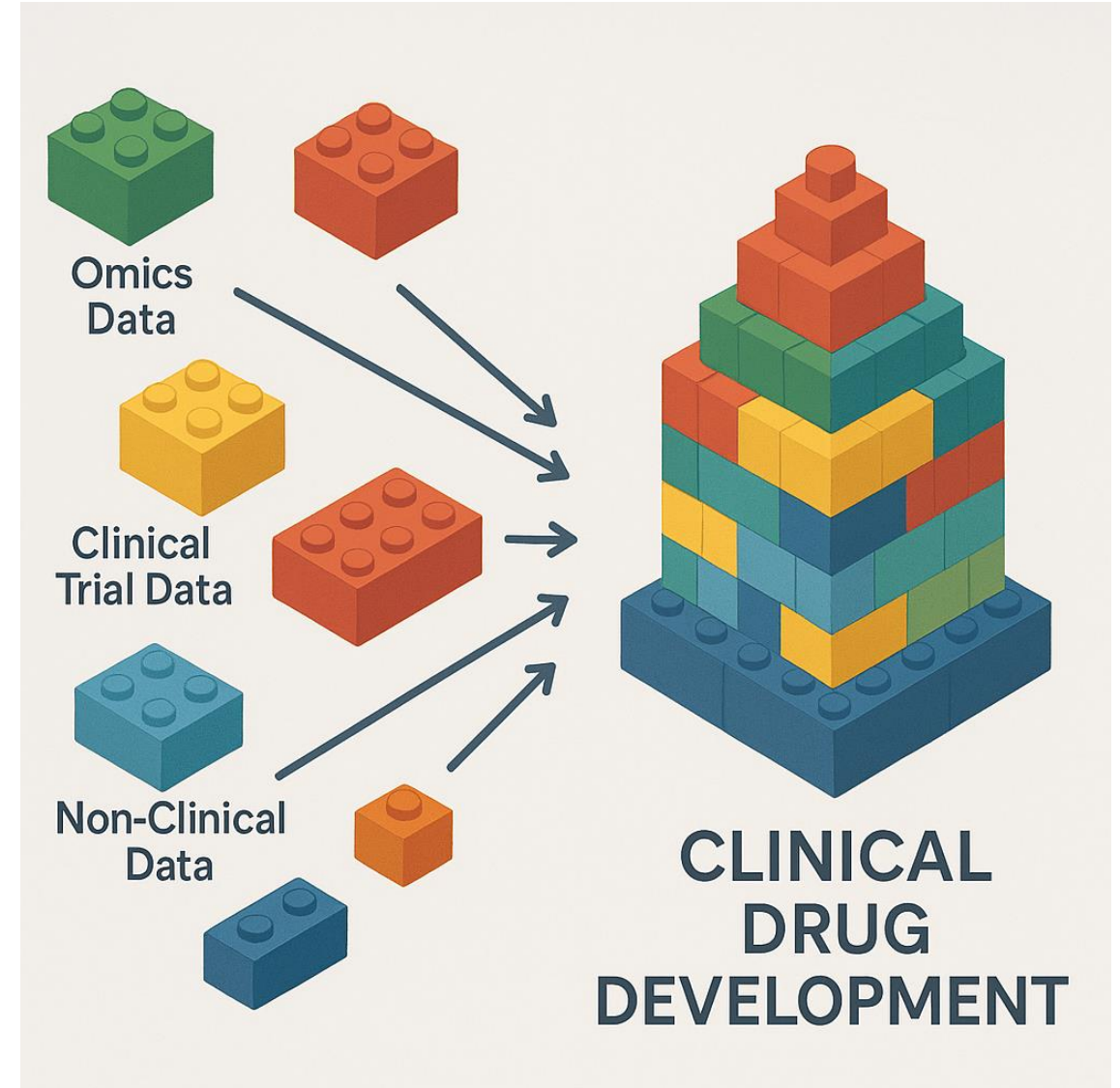
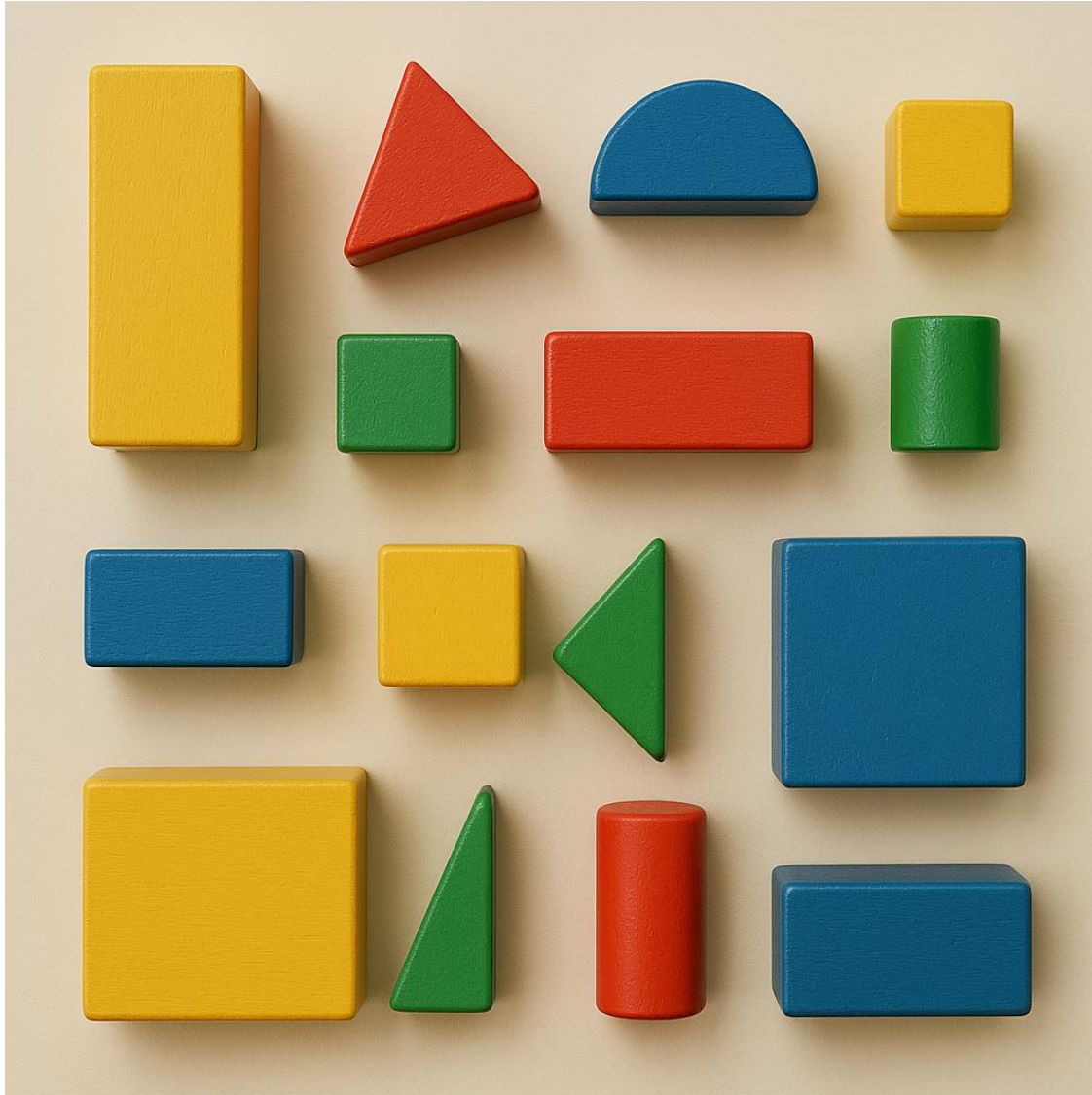


Language Agnostic

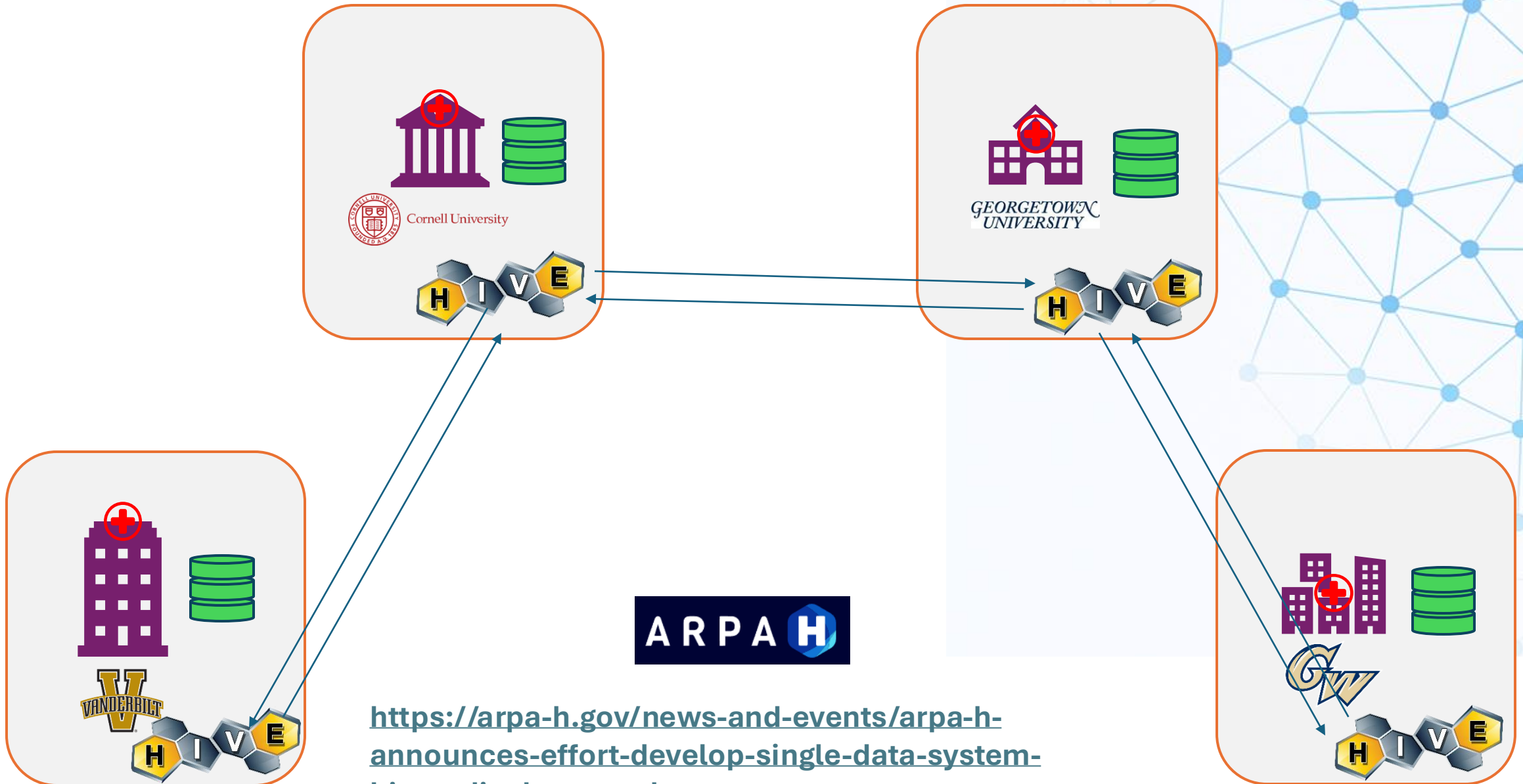


Can be used beyond Omics

Taking different building blocks to a coherent structure

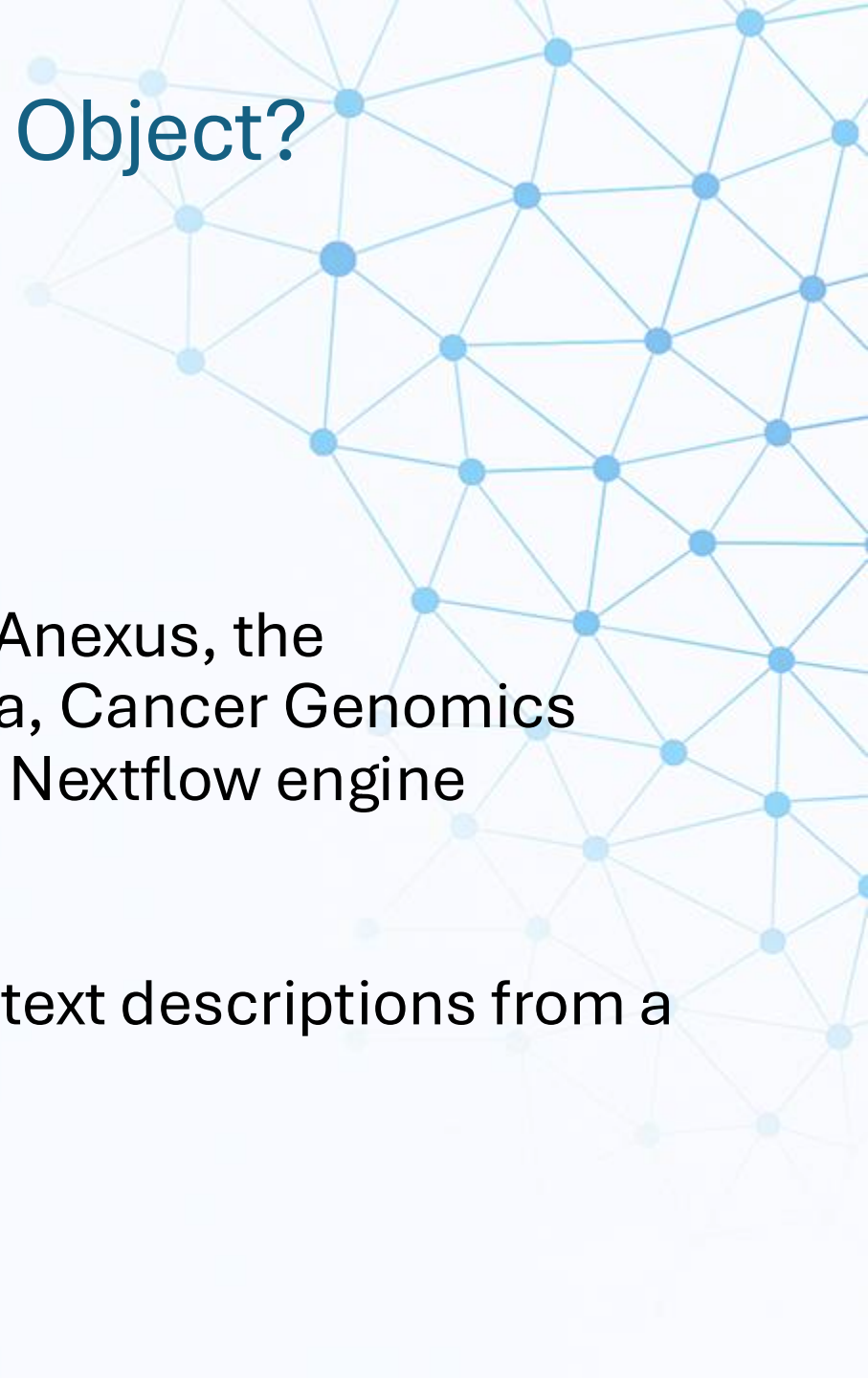


BCO Security Implementation



How to build a BioCompute Object?

- Whirl R package
- R Shiny app
- Platform-specific tools like the BCOnexus for DNAnexus, the BioCompute tool on Velsera's platforms (Cavatica, Cancer Genomics Cloud, Seven Bridges, and BioData Catalyst), the Nextflow engine
- BioCompute export on Galaxy
- LLM- based tool to help with generating the plain text descriptions from a workflow

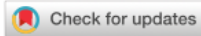


Results and Solutions - MAE

FOCUS ON COMPUTER RESOURCES | NOVEMBER 01 2017

Software for the Integration of Multiomics Experiments in Bioconductor

Marcel Ramos; Lucas Schiffer; Angela Re; Rimsha Azhar; Azfar Basunia; Carmen Rodriguez; Tiffany Chan; Phil Chapman; Sean R. Davis; David Gomez-Cabrero; Aedin C. Culhane; Benjamin Haibe-Kains; Kasper D. Hansen; Hanish Kodali; Marie S. Louis; Arvind S. Mer; Markus Riester; Martin Morgan; Vince Carey; Levi Waldron 



[+ Author & Article Information](#)

Cancer Res (2017) 77 (21): e39–e42.

<https://doi.org/10.1158/0008-5472.CAN-17-0344> [Article history](#) 

 Split-Screen  Views  PDF  Share  Tools  Versions

Abstract

Multiomics experiments are increasingly commonplace in biomedical research and add layers of complexity to experimental design, data integration, and analysis. R and Bioconductor provide a generic framework for statistical analysis and visualization, as well as specialized data classes for a variety of high-throughput data types, but methods are lacking for integrative analysis of multiomics experiments. The MultiAssayExperiment software package, implemented in R and leveraging Bioconductor software and design principles, provides for the coordinated representation of, storage of, and operation on multiple diverse genomics data. We provide the unrestricted multiple 'omics data for each cancer tissue in The Cancer Genome Atlas as ready-to-analyze MultiAssayExperiment objects and demonstrate in these and other datasets how the software simplifies data representation, statistical analysis, and visualization. The MultiAssayExperiment Bioconductor package reduces major obstacles to efficient, scalable, and reproducible statistical analysis of multiomics data and enhances data science applications of multiple omics datasets. *Cancer Res*; 77(21); e39–42. ©2017 AACR.

- MAE = Multi Assay Experiment
- R package which utilize Bioconductor software
- Represent, integrate, manage, analyze and visualize multi omics data
- Overcome the challenges of:
 - Heterogeneity
 - Large scale data storage
 - Bioinformatic Analysis
 - Reproducibility

Results and Solutions - MAE

colData

	patientID <chr>	years_to_birth <int>	vital_status <int>	days_to_death <int>
TCGA-OR-A5J1	TCGA-OR-A5J1	58	1	1355
TCGA-OR-A5J2	TCGA-OR-A5J2	44	1	1677
TCGA-OR-A5J3	TCGA-OR-A5J3	23	0	NA
TCGA-OR-A5J4	TCGA-OR-A5J4	23	1	423
TCGA-OR-A5J5	TCGA-OR-A5J5	30	1	365

RNASeq2GeneNorm assay from ExperimentList

	TCGA-OR-A5J1-01A-11R-A29S-07 <dbl>	TCGA-OR-A5J2-01A-11R-A29S-07 <dbl>	TCGA-OR-A5J3-01A-11R-A29S-07 <dbl>
DIRAS3	1487.0317	9.6631	18.9602
MAPK14	778.5783	2823.6469	1061.7686
YAP1	1009.6061	2305.0590	1561.2502
CDKN1B	2101.3449	4248.9584	1348.5410
ERBB2	651.2968	246.4098	90.0607

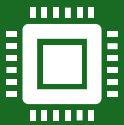
assay <fct>	primary <chr>	colname <chr>
RNASeq2GeneNorm	TCGA-OR-A5J1	TCGA-OR-A5J1-01A-11R-A29S-07
RNASeq2GeneNorm	TCGA-OR-A5J2	TCGA-OR-A5J2-01A-11R-A29S-07
RNASeq2GeneNorm	TCGA-OR-A5J3	TCGA-OR-A5J3-01A-11R-A29S-07
RNASeq2GeneNorm	TCGA-OR-A5J5	TCGA-OR-A5J5-01A-11R-A29S-07
RNASeq2GeneNorm	TCGA-OR-A5J6	TCGA-OR-A5J6-01A-31R-A29S-07

sampleMap

Advantages of MAE



Standardisation: MAE enforces a consistent structure for multi-omics datasets, reducing complexity and improving reproducibility.



Interoperability: Integrates seamlessly with other Bioconductor packages for downstream analysis, including survival modelling and machine learning.

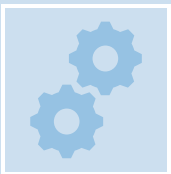


Scalability: Handles large datasets across multiple platforms, making it suitable for clinical development pipelines.

Combination of MAE and Biocompute

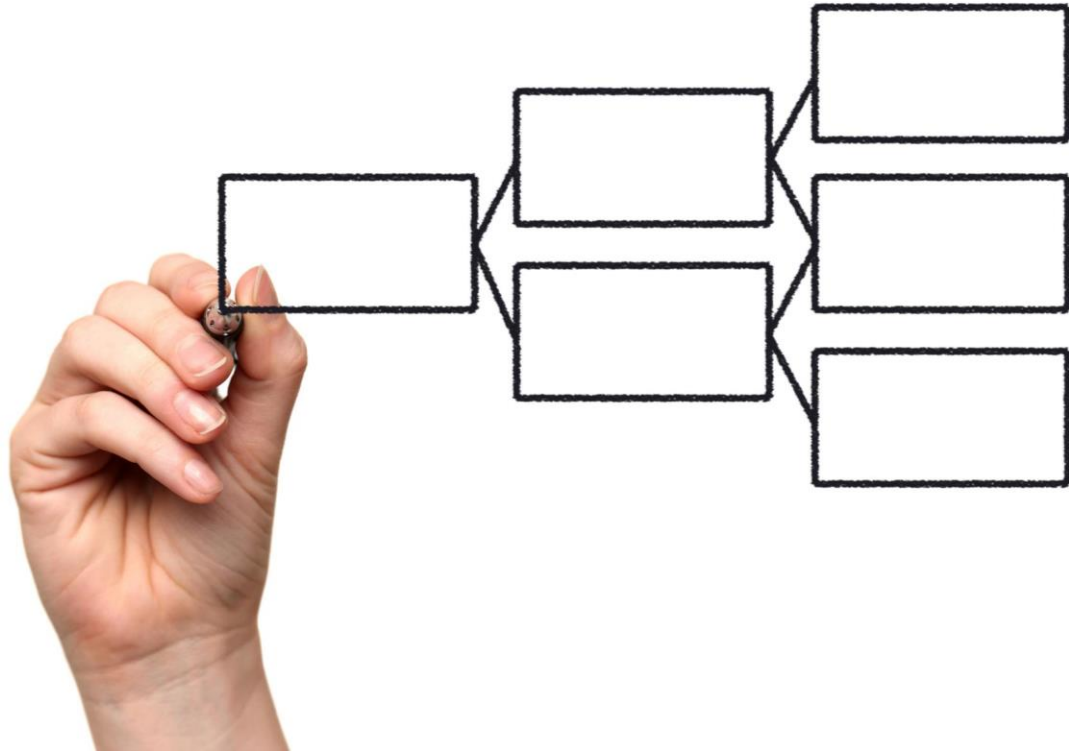


MAE is used for formal analysis



BCO is used to document the workflow, data provenance, and harmonization of data sources.

Conclusion



- High-dimensional omics data suits hierarchical JSON models over flat tables.
- BioCompute (IEEE-2791) standard captures workflows, data provenance, and harmonization.
- MAE is recommended for formal analysis combined with BioCompute documentation.
- CDISC SDTM.GF can be leveraged for structured communication with regulators.
- Early engagement with regulatory authorities improves study clarity and acceptance.
- Open-source software adoption introduces validation, maintenance, and security challenges.
- Pistoia alliance project presentation

References



Phuse Working Group

workinggroups@phuse.global



PHUSE Paper:

*DH04: Framework for Integrating
Omics Data with Other Clinical
and Non-Clinical Data Sources*



www.biocomputeobject.org

Acknowledgement

Co-author of the PHUSE Paper:

- Jonathon Keeney (George Washington University)
- Bron Kisler (Independent Global and Genomics Standards Expert)
- Ashwini Yermal Shanbhogue (Independent)

Contributor of the paper:

- Jeffrey Long (GSK)
- Sireesha Krosuri (Ephicity)

PHUSE Organization



Any questions ?

