Data Management Plans for Health Sciences Research

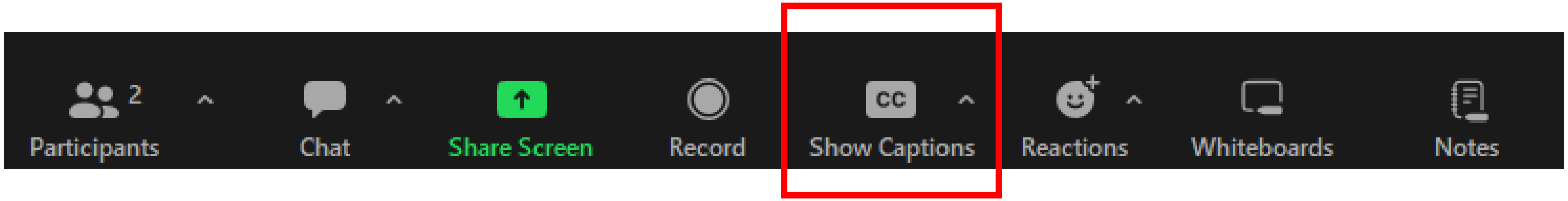
October 10, 2024

UNIVERSITY OF TORONTO

Centre for Research
& Innovation Support

Turning ‘live captions’ on and off

- On your meeting controls, click on “**Show Captions**”



Land Acknowledgement

We wish to acknowledge this land on which the University of Toronto operates.

For thousands of years it has been the traditional land of the Huron-Wendat, the Seneca, and the Mississaugas of the Credit.

Today, this meeting place is still the home to many Indigenous people from across Turtle Island and we are grateful to have the opportunity to work on this land.

Housekeeping

- This webinar is being recorded and transcribed
- A link to the recording and presenter slides will be sent to all participants after the session
- Please put questions into the chat, we will hold all questions until the end of the presentations

Purpose & Agenda

Bring together the University of Toronto tri-campus and TAHSN health sciences research community for facilitated conversations about research data management.

Learning Objectives:

- DMP benefits and challenges
- Strategies and tools
- Strengthen DMP practices

10:10 – 10:15 am	**DMP Basics**
10:15 – 10:40 am	**Presentations** • ***Dr. Victoria Hodgkinson****, Executive Director, Canadian Neuromuscular Disease Registry* • ***Dr. Trevor Pugh****, Professor, Department of Medical Biophysics & Senior Scientist, Princess Margaret Cancer Centre* • ***Dr. Denise Mak****, Director of Data Science & Innovation, GEMINI*
10:40 – 11:25 am	**Panel Discussion and Q&A**

Data Management Plan (DMP) – The Basics

What

- Covers practices, processes, and strategies for data management
- A living document that should be updated

“DMPs guide researchers in articulating their plans for managing data; they do not necessarily compel researchers to manage data differently.”
Tri-Agency Policy

Why

- Identify opportunities and challenges early
- Adapt to unanticipated obstacles
- Engage partners and collaborators
- Improve research design and efficiency
- **Meet funder requirements**

How

Components

- Data collection
- Data security, storage, & backup
- Data preservation & sharing
- Roles & responsibilities
- Ethical, legal, commercial constrains
- Other

Tools & templates

- DMP Assistant
- McMaster Data Management Plan Database
- Funder-specific requirements

Tri-Agency Research Data Management Policy

Panelists

Dr. Victoria Hodgkinson
Executive Director, Canadian Neuromuscular Disease Registry

Dr. Trevor Pugh
Professor, Department of Medical Biophysics & Senior Scientist, Princess Margaret Cancer Centre

Dr. Denise Mak
Director of Data Science & Innovation, GEMINI

Data Management Plans

Victoria Hodgkinson, PhD,

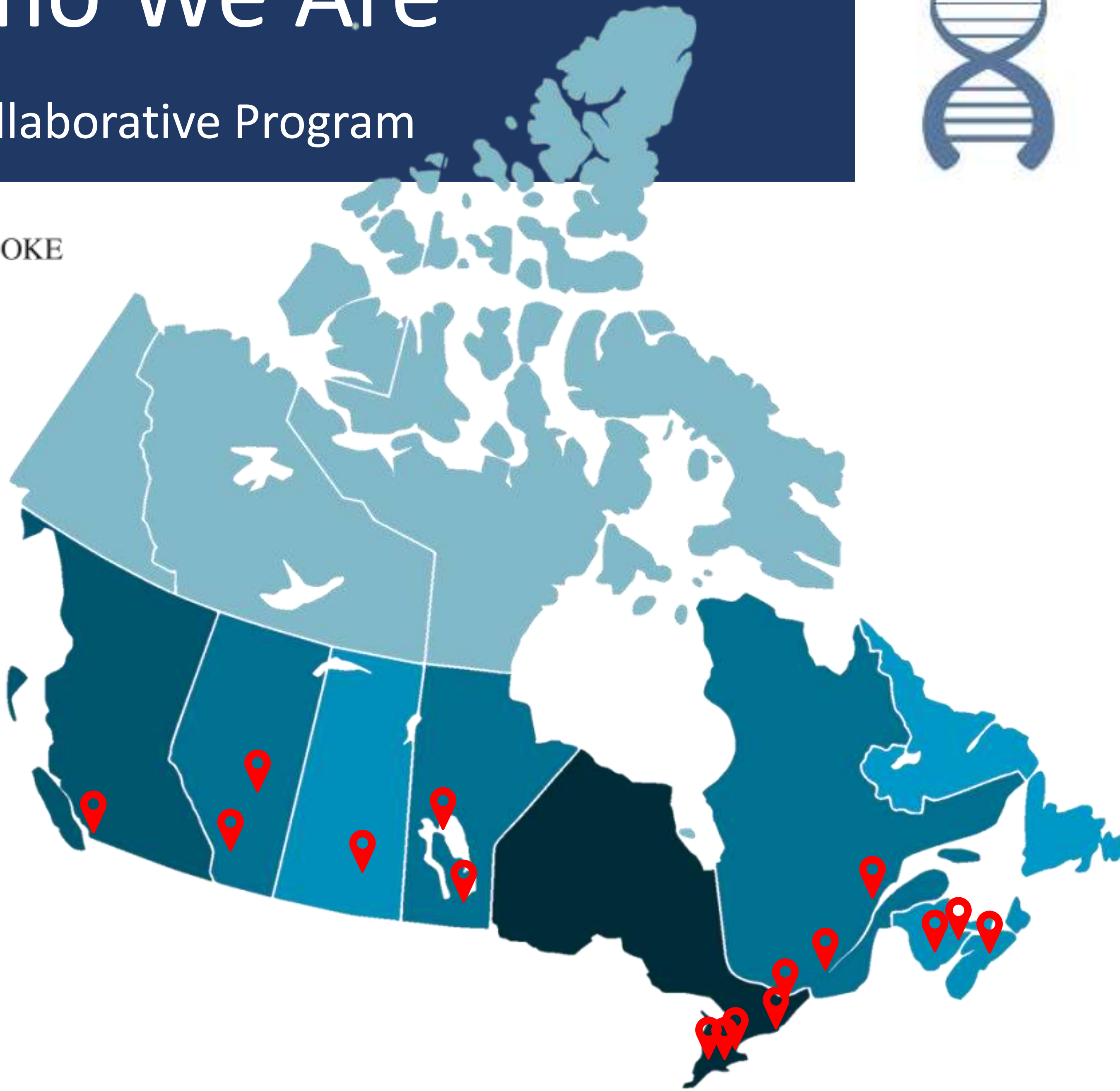
Canadian Neuromuscular Disease Registry (CNDR)

vhodgkin@ucalgary.ca

The CNDR: Who We Are

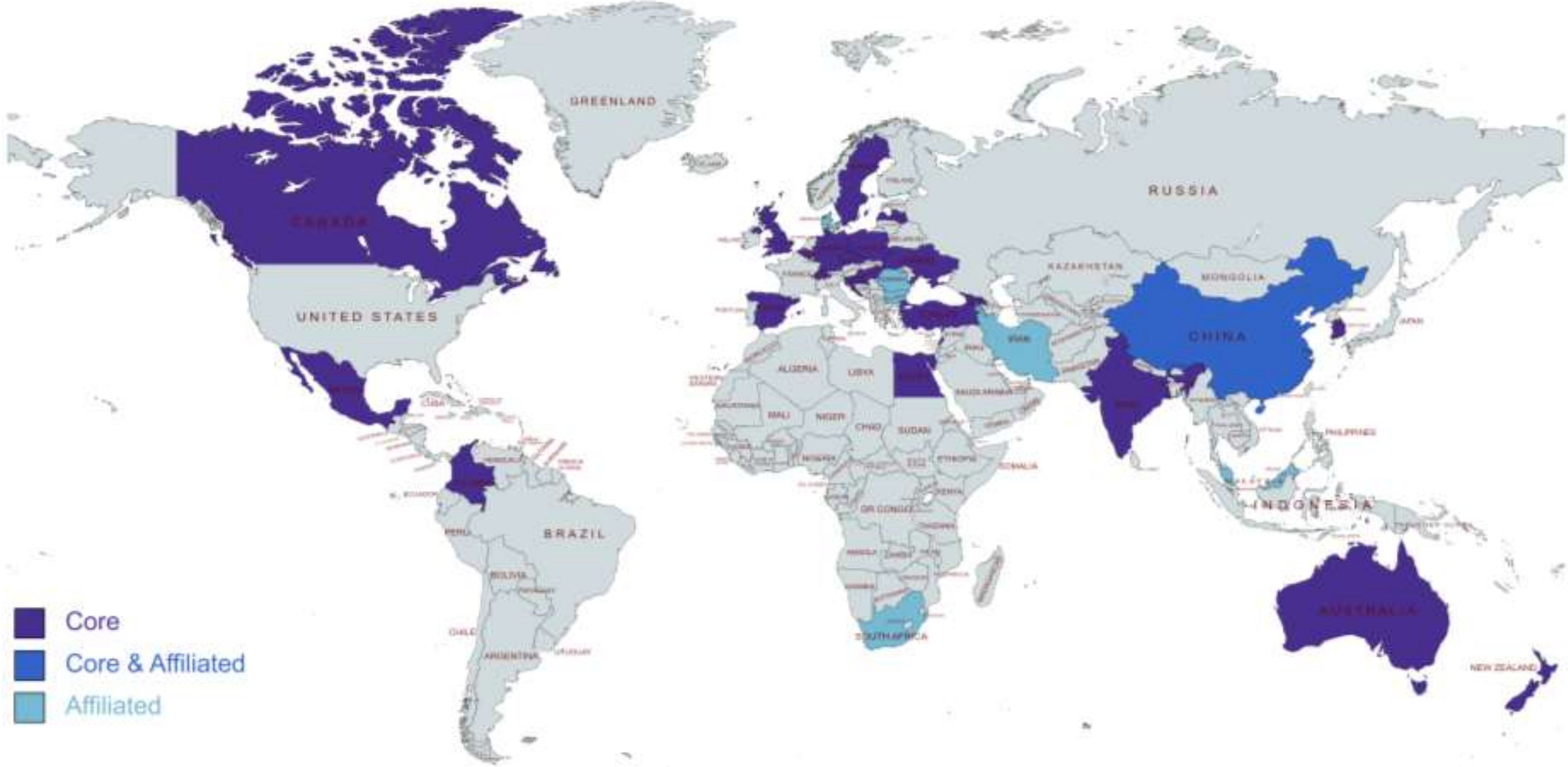
A Multi-centre, National Collaborative Program

University of Calgary · The University of British Columbia · University of Manitoba · Université de Sherbrooke · University of Alberta · University of Saskatchewan · Queens University · Dalhousie University · University of Toronto · Western University · Canada · uOttawa · McMaster University · McGill University · Université de Montréal · Université Laval



- Launched in 2011
- 38 neuromuscular clinics (pediatric & adult)
- > 136 clinician investigators in network (48 core)
- Consent-based
- Clinical data abstraction (SMA, DMD, ALS, DMD, FSHD, …)
- Currently over 6000 patients nationwide from all provinces and territories
- Broad data use by academic, not-for-profit, industry, and regulators, in Canada and internationally

TREAT-NMD Global Registry Network: SMA

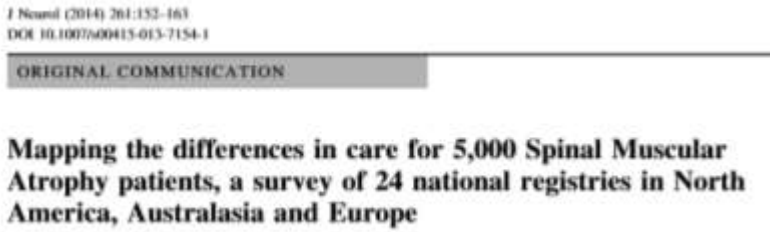


37-58* Members

9,726 Patients

7,735 Genetically-confirmed Patients

Victoria Hodgkinson
TREAT-NMD SMA Chair
Canada



J Neurol (2014) 261:152–163
DOI 10.1007/s00415-013-7154-1

ORIGINAL COMMUNICATION

Mapping the differences in care for 5,000 Spinal Muscular Atrophy patients, a survey of 24 national registries in North America, Australasia and Europe

Global Collaborative Data Collection for Real World Evidence in Spinal Muscular Atrophy

Hodgkinson V., et al.

inflamed brain alliance

bring health to mind.

CHAER

Canadian Hereditary Angioedema Registry

RGCAOH

Registre Canadien d'angioedéme hèréditaire

McCAIG INSTITUTE FOR BONE AND JOINT HEALTH

Canadian Registry For Amyloidosis Research

mitocanada

AcroNEXT
Global Acromegaly Registry

BetterLife FSHD

HEADS REGISTRY

HEADACHE, EAR, AUDITORY, DIZZINESS, SINUS REGISTRY

Canadian Acromegaly Registry

Real-World Evidence for Canadian Neuromuscular Disease: Establishing a Framework for National Integration of Patient Reported Outcomes, Clinical Registry Data, Healthcare Utilization and Healthcare Associated Costs

CASE STUDY

CIHR Rare Disease Team Grant

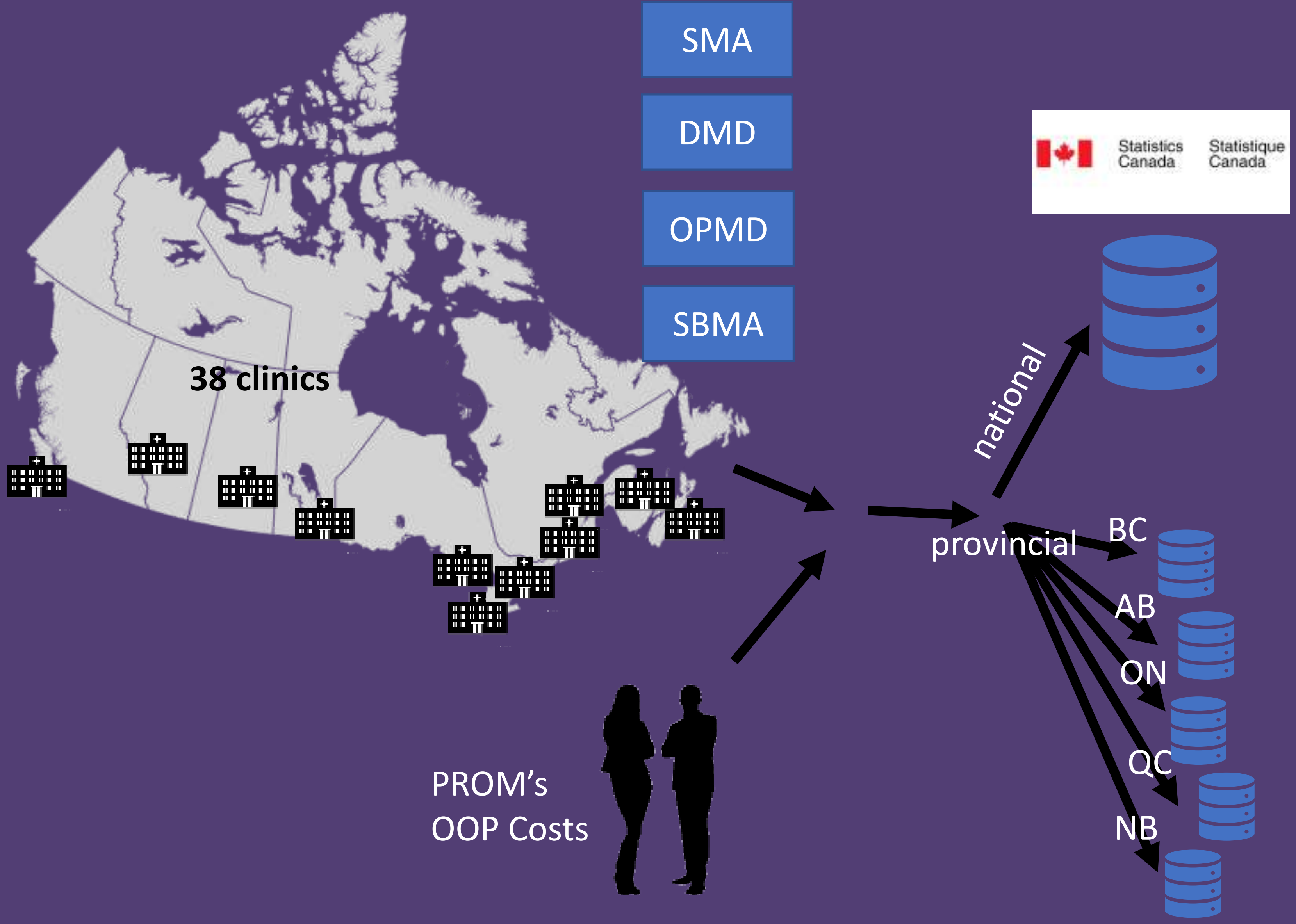
NPI: Dr. Reshma Amin (Sick Kids)

CIHR IRSC

MUSCULAR DYSTROPHY CANADA
DYSTROPHIE MUSCULAIRE CANADA

DEFEAT DUCHENNE
CANADA

cure
SMA



SMA

DMD

OPMD

SBMA

38 clinics

PROM’s
OOP Costs

Statistics Canada Statistique Canada

national

provincial

BC

AB

ON

QC

NB

Acute Healthcare Service Utilization

Mortality

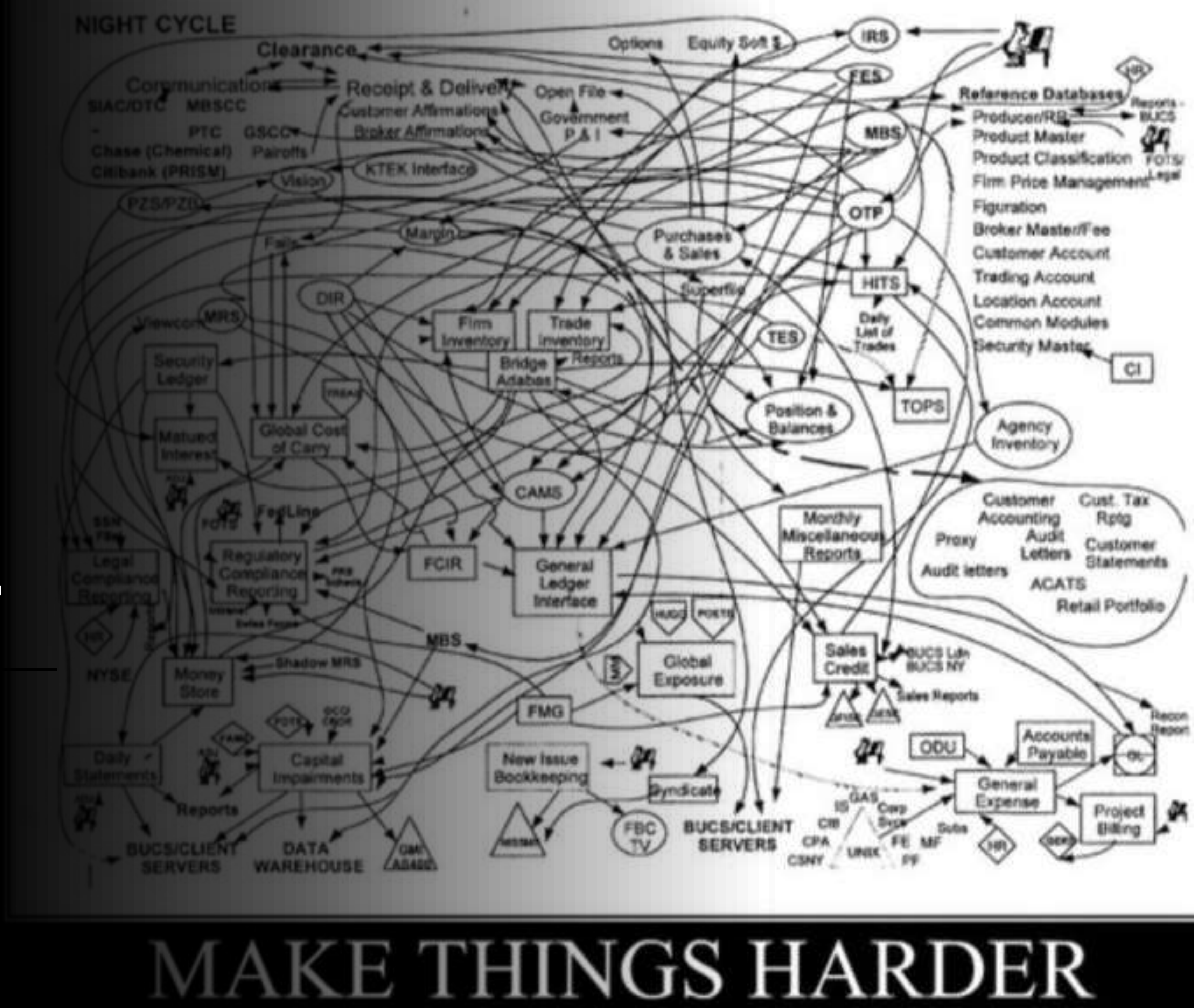
Health-related Quality of Life

Out of pocket costs

Total Healthcare Service Utilization

Publicly Funded Healthcare Costs

DMP PROCESS


NIGHT CYCLE
Clearance
Communications
SIAC/DTC
MBSCC
PTC
GSCC
Chase (Chemical)
Citibank (PRISM)
Payoffs
Receipt & Delivery
Customer Affirmations
Broker Affirmations
Open File
Government P & I
Options
Equity Soft $
IRS
FES
MBS
OTP
Reference Databases
Producer/RR
Product Master
Product Classification
Firm Price Management
Figuration
Broker Master/Fee
Customer Account
Trading Account
Location Account
Common Modules
Security Master
Reports - BUCS
FOTS/ Legal
CI
Vision
KTEK Interface
PZS/PZB
Margin
Fails
Purchases & Sales
Superfile
HITS
Daily List of Trades
TES
TOPS
DIR
MRS
Viewcon
Firm Inventory
Trade Inventory
Bridge Adabas
Reports
Position & Balances
Agency Inventory
Security Ledger
Matured Interest
Global Cost of Carry
CAMS
Customer Accounting
Cust. Tax Rptg
Proxy
Audit Letters
Customer Statements
Audit letters
ACATS
Retail Portfolio
Monthly Miscellaneous Reports
FedLine
Legal Compliance Reporting
Regulatory Compliance Reporting
FCIR
General Ledger Interface
Global Exposure
Sales Credit
BUCS Ldn
BUCS NY
Sales Reports
MBS
Shadow MRS
NYSE
Money Store
FMG
Daily Statements
Capital Impairments
New Issue Bookkeeping
Syndicate
ODU
Accounts Payable
GL
Recon Report
General Expense
Project Billing
Reports
BUCS/CLIENT SERVERS
DATA WAREHOUSE
FBC TV
BUCS/CLIENT SERVERS
GAS
Corp
CPA
CSNY
UNIX
FE
MF
PF
Subs
MAKE THINGS HARDER

Where to Start?

72 team members

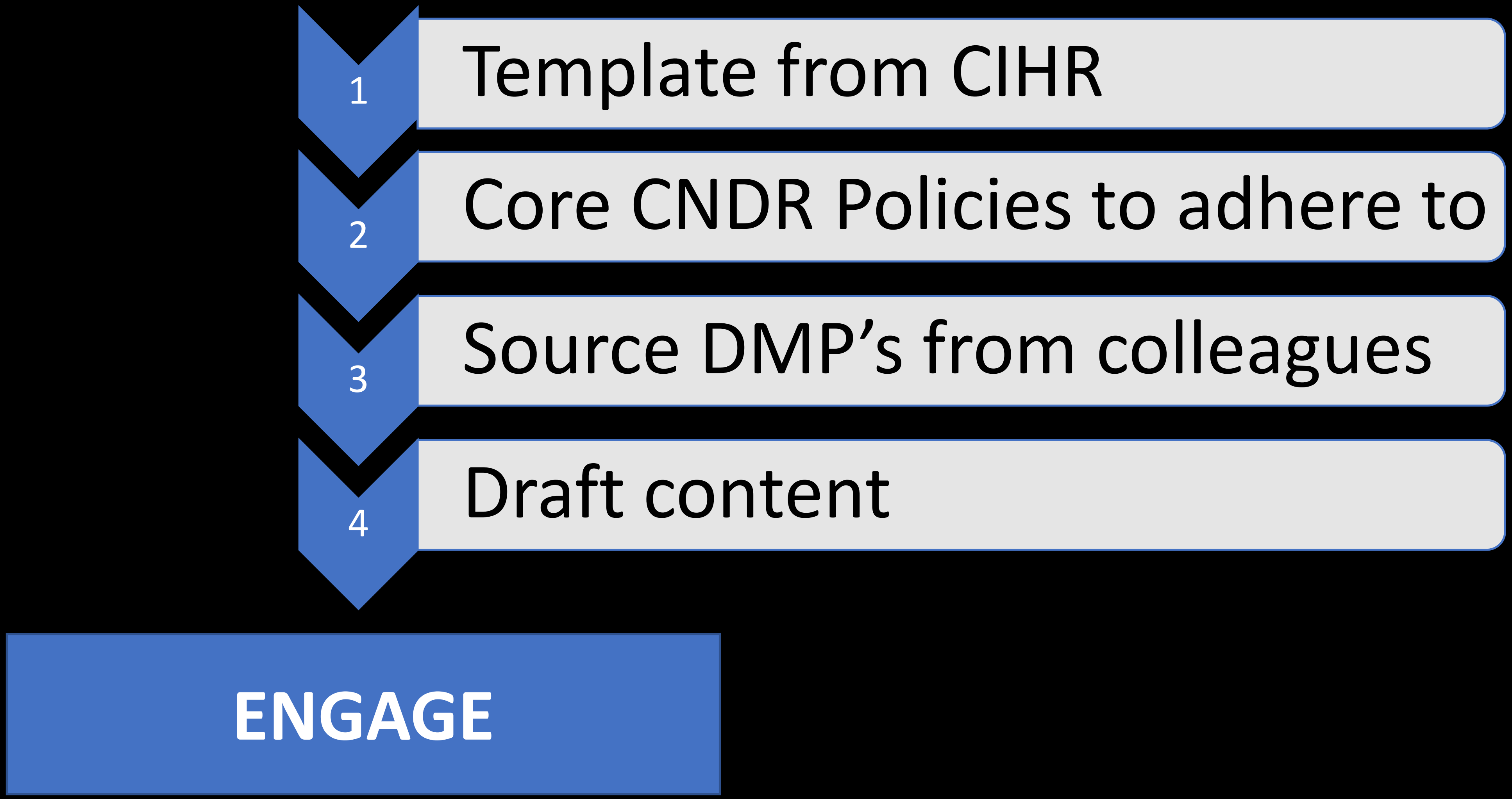
Clinicians, patient organizations, patient partners, community guiding circle

Statistics Canada, CIHI, Provincial health authorities

Non-member collaborators

CDA, other Canadian registries, RWE4decisions, TREAT-NMD

Where to Start? Process



1. Template from CIHR
2. Core CNDR Policies to adhere to
3. Source DMP’s from colleagues
4. Draft content

ENGAGE

DMP

1. Data collection and storage
    1. Collect
    2. Generate
    3. Link (SDLE, income, claims data)
    4. Data protection
    5. Data management
        a) Context of Indigenous people
2. Ethics and legal compliance, access
3. Data storage technical and preservation
4. Data sharing
        a) Context of Indigenous people
        b) Sensitive data
5. Data standards and international alignment

Thanks!

Data Management Plans for Clinical Genomics Research

Trevor Pugh, PhD, FACMG

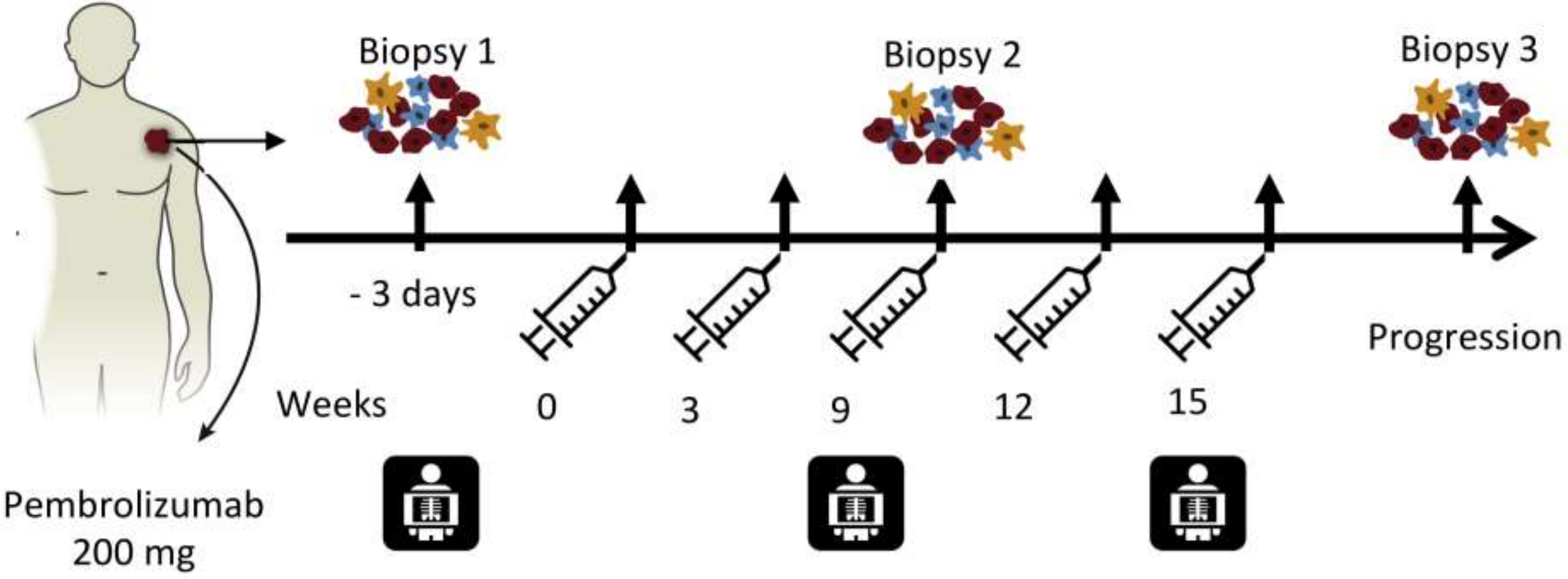
Canada Research Chair in Translational Genomics
Senior Scientist, Princess Margaret Cancer Centre
Director, Genomics, Ontario Institute for Cancer Research
Professor, Dept. of Medical Biophysics, University of Toronto
trevor.pugh@utoronto.ca

My Perspective

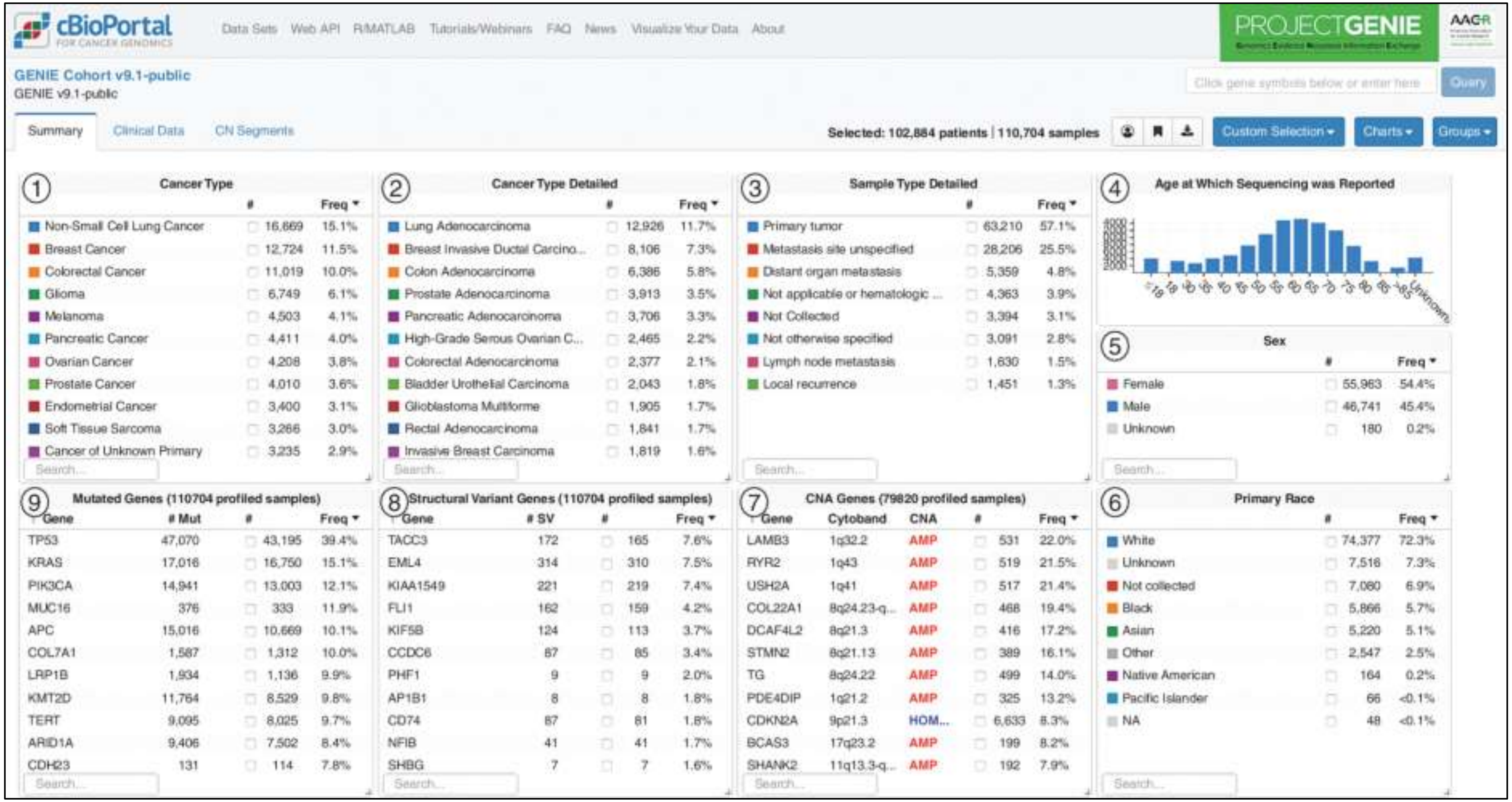
Scientific Director of the Princess Margaret Genomics Centre (www.pmgenomics.ca) → basic research core

Medical Director of the OICR Genomics Program (https://genomics.oicr.on.ca) → translational/clinical research core

Principal investigator using genomics to understand which cancer patients respond to treatment, when, & why (www.pughlab.org)



Our Goal: To enable reuse and integration of clinical and genomic data to answer new scientific questions



GENIE Cohort v9.1-public
GENIE v9.1-public

Selected: 102,884 patients | 110,704 samples

Cancer Type

	#	Freq
Non-Small Cell Lung Cancer	□ 16,669	15.1%
Breast Cancer	□ 12,724	11.5%
Colorectal Cancer	□ 11,019	10.0%
Glioma	□ 6,749	6.1%
Melanoma	□ 4,503	4.1%
Pancreatic Cancer	□ 4,411	4.0%
Ovarian Cancer	□ 4,208	3.8%
Prostate Cancer	□ 4,010	3.6%
Endometrial Cancer	□ 3,400	3.1%
Soft Tissue Sarcoma	□ 3,266	3.0%
Cancer of Unknown Primary	□ 3,235	2.9%

Cancer Type Detailed

	#	Freq
Lung Adenocarcinoma	□ 12,926	11.7%
Breast Invasive Ductal Carcino...	□ 8,106	7.3%
Colon Adenocarcinoma	□ 6,386	5.8%
Prostate Adenocarcinoma	□ 3,913	3.5%
Pancreatic Adenocarcinoma	□ 3,706	3.3%
High-Grade Serous Ovarian C...	□ 2,465	2.2%
Colorectal Adenocarcinoma	□ 2,377	2.1%
Bladder Urothelial Carcinoma	□ 2,043	1.8%
Glioblastoma Multiforme	□ 1,905	1.7%
Rectal Adenocarcinoma	□ 1,841	1.7%
Invasive Breast Carcinoma	□ 1,819	1.6%

Sample Type Detailed

	#	Freq
Primary tumor	□ 63,210	57.1%
Metastasis site unspecified	□ 28,206	25.5%
Distant organ metastasis	□ 5,359	4.8%
Not applicable or hematologic ...	□ 4,363	3.9%
Not Collected	□ 3,394	3.1%
Not otherwise specified	□ 3,091	2.8%
Lymph node metastasis	□ 1,630	1.5%
Local recurrence	□ 1,451	1.3%

Sex

	#	Freq
Female	□ 55,963	54.4%
Male	□ 46,741	45.4%
Unknown	□ 180	0.2%

Mutated Genes (110704 profiled samples)

Gene	# Mut	#	Freq
TP53	47,070	□ 43,195	39.4%
KRAS	17,016	□ 16,750	15.1%
PIK3CA	14,941	□ 13,003	12.1%
MUC16	376	□ 333	11.9%
APC	15,016	□ 10,669	10.1%
COL7A1	1,587	□ 1,312	10.0%
LRP1B	1,934	□ 1,136	9.9%
KMT2D	11,764	□ 8,529	9.8%
TERT	9,095	□ 8,025	9.7%
ARID1A	9,406	□ 7,502	8.4%
CDH23	131	□ 114	7.8%

Structural Variant Genes (110704 profiled samples)

Gene	# SV	#	Freq
TACC3	172	□ 165	7.6%
EML4	314	□ 310	7.5%
KIAA1549	221	□ 219	7.4%
FLI1	162	□ 159	4.2%
KIF5B	124	□ 113	3.7%
CCDC6	87	□ 85	3.4%
PHF1	9	□ 9	2.0%
AP1B1	8	□ 8	1.8%
CD74	87	□ 81	1.8%
NFIB	41	□ 41	1.7%
SHBG	7	□ 7	1.6%

CNA Genes (79820 profiled samples)

Gene	Cytoband	CNA	#	Freq
LAMB3	1q32.2	AMP	□ 531	22.0%
RYR2	1q43	AMP	□ 519	21.5%
USH2A	1q41	AMP	□ 517	21.4%
COL22A1	8q24.23-q...	AMP	□ 468	19.4%
DCAF4L2	8q21.3	AMP	□ 416	17.2%
STMN2	8q21.13	AMP	□ 389	16.1%
TG	8q24.22	AMP	□ 499	14.0%
PDE4DIP	1q21.2	AMP	□ 325	13.2%
CDKN2A	9p21.3	HOM...	□ 6,633	8.3%
BCAS3	17q23.2	AMP	□ 199	8.2%
SHANK2	11q13.3-q...	AMP	□ 192	7.9%

Primary Race

	#	Freq
White	□ 74,377	72.3%
Unknown	□ 7,516	7.3%
Not collected	□ 7,080	6.9%
Black	□ 5,866	5.7%
Asian	□ 5,220	5.1%
Other	□ 2,547	2.5%
Native American	□ 164	0.2%
Pacific Islander	□ 66	<0.1%
NA	□ 48	<0.1%

Data from >110,000 tumors from >100,000 people treated at 19 cancer centers worldwide

Predicted enrollment on genome-guided clinical trials

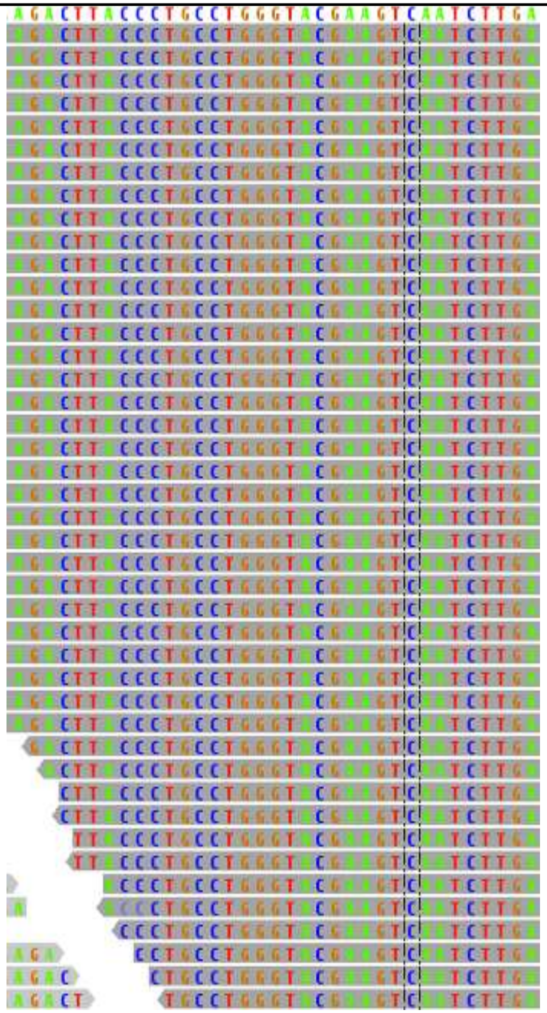
Discovered driver alterations in rare tumors

Identified cancer types without actionable mutations that could benefit from whole genome sequencing

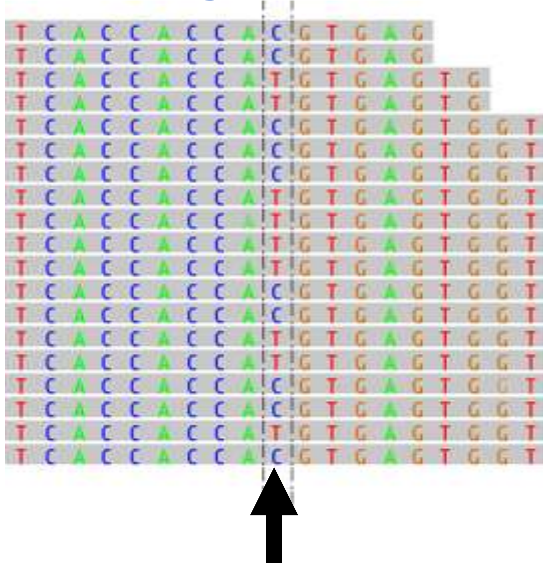
“AACR Project GENIE: 100,000 Cases and Beyond”. AACR Project GENIE Consortium, Genomics and Analysis Working Group. *Cancer Discovery* (2022) 12 (9): 2044–2057.

Primary, Secondary, and Tertiary data require different systems that reflect identifiability of underlying data

Generation

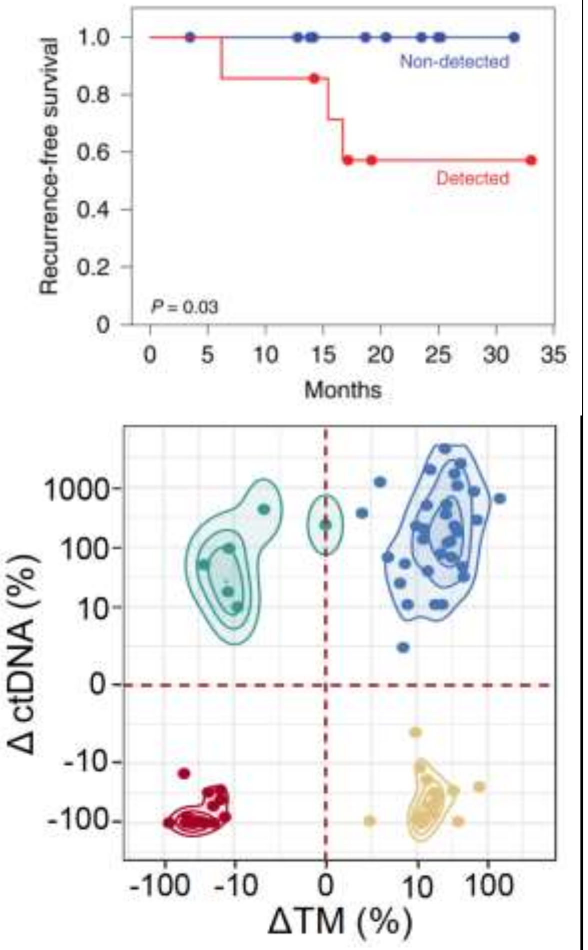


Primary



Heterozygous variant
Total coverage = 19
C = 10, T = 9
Allele balance = 0.47
Gene: *LMNA*
Protein: H566H
Effect: Splice region
Clinical significance: Benign

Secondary



Tertiary

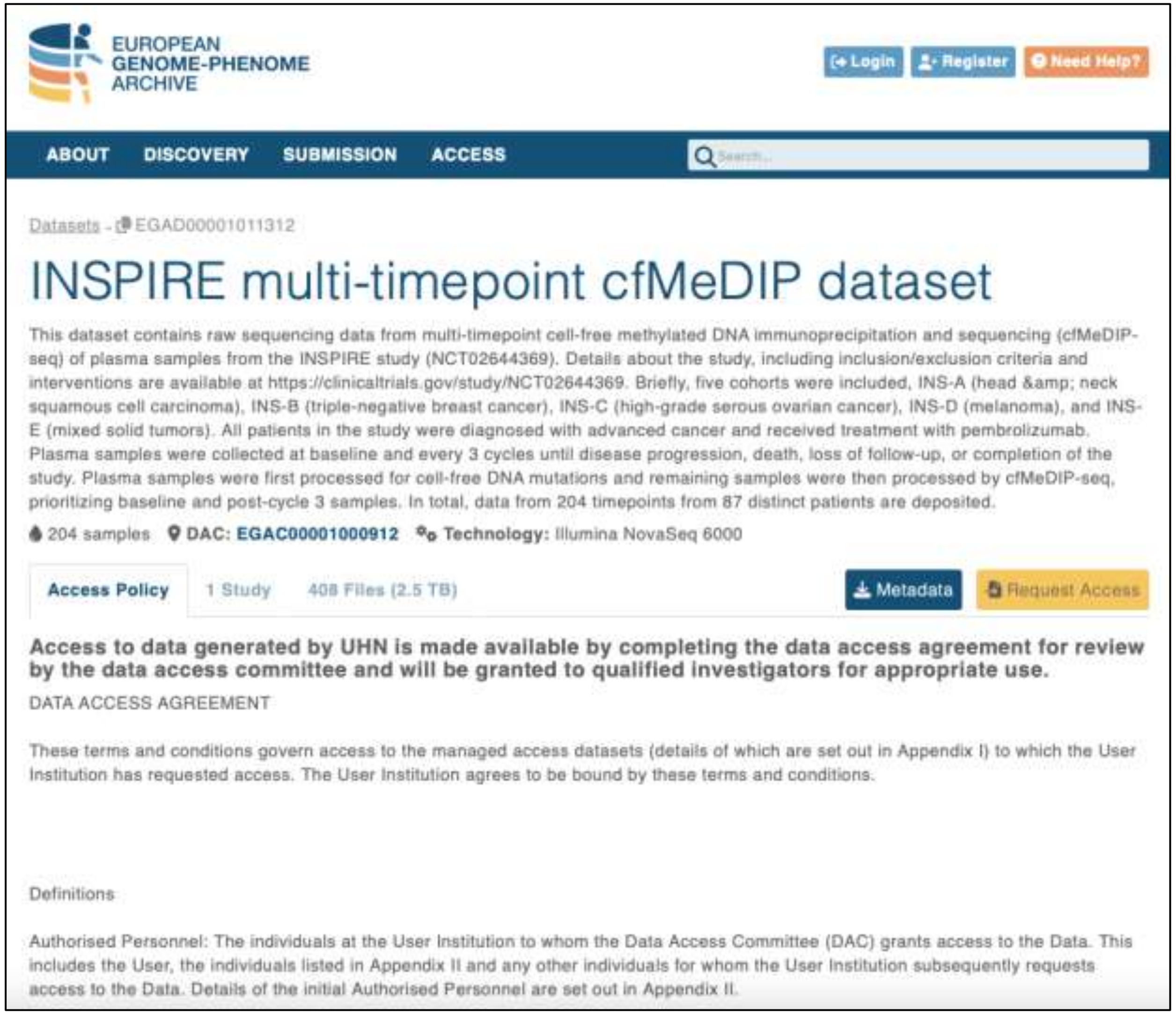
Primary: Controlled access databases for original genome sequencing reads (EGA and dbGAP)

https://ega-archive.org/datasets/EGAD00001011312

Description of the study, cancer types, timing of collection, and genomics technology (cfMeDIP-seq)

Access policy, any over-arching studies or related data sets, number and size of files for download

Governance details including data access agreement and information on the Data Access Committee who will evaluate the request



INSPIRE multi-timepoint cfMeDIP dataset

This dataset contains raw sequencing data from multi-timepoint cell-free methylated DNA immunoprecipitation and sequencing (cfMeDIP-seq) of plasma samples from the INSPIRE study (NCT02644369). Details about the study, including inclusion/exclusion criteria and interventions are available at https://clinicaltrials.gov/study/NCT02644369. Briefly, five cohorts were included, INS-A (head &amp; neck squamous cell carcinoma), INS-B (triple-negative breast cancer), INS-C (high-grade serous ovarian cancer), INS-D (melanoma), and INS-E (mixed solid tumors). All patients in the study were diagnosed with advanced cancer and received treatment with pembrolizumab. Plasma samples were collected at baseline and every 3 cycles until disease progression, death, loss of follow-up, or completion of the study. Plasma samples were first processed for cell-free DNA mutations and remaining samples were then processed by cfMeDIP-seq, prioritizing baseline and post-cycle 3 samples. In total, data from 204 timepoints from 87 distinct patients are deposited.

204 samples **DAC:** **EGAC00001000912** **Technology:** Illumina NovaSeq 6000

Access to data generated by UHN is made available by completing the data access agreement for review by the data access committee and will be granted to qualified investigators for appropriate use.

DATA ACCESS AGREEMENT

These terms and conditions govern access to the managed access datasets (details of which are set out in Appendix I) to which the User Institution has requested access. The User Institution agrees to be bound by these terms and conditions.

Definitions

Authorised Personnel: The individuals at the User Institution to whom the Data Access Committee (DAC) grants access to the Data. This includes the User, the individuals listed in Appendix II and any other individuals for whom the User Institution subsequently requests access to the Data. Details of the initial Authorised Personnel are set out in Appendix II.

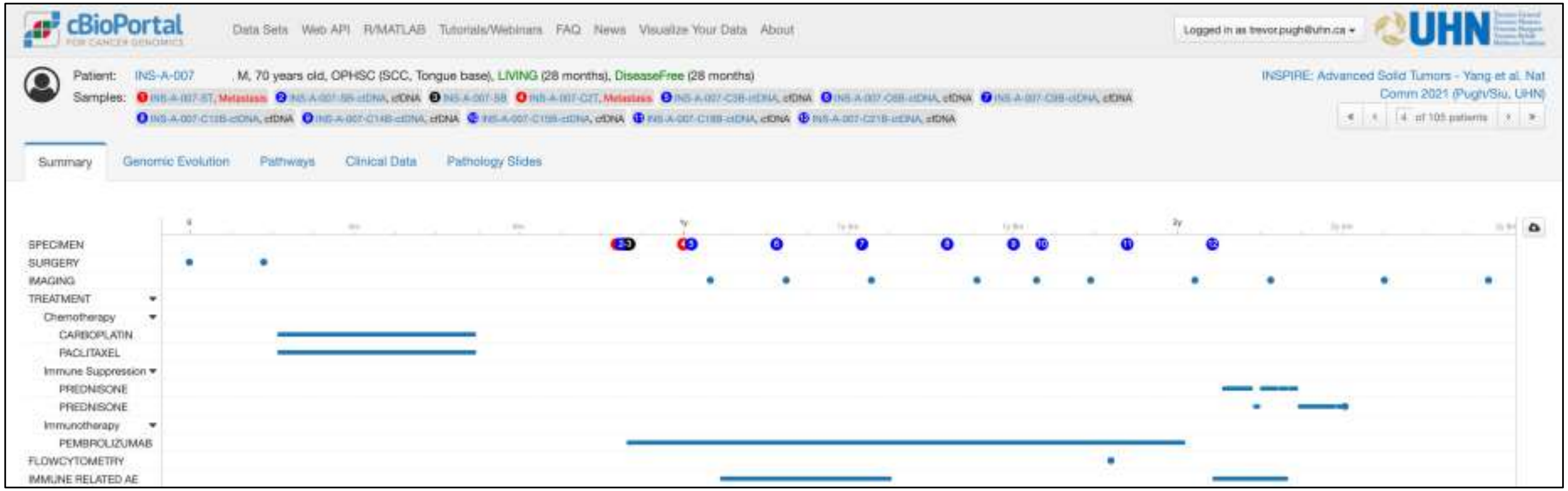
Similar information for NIH-funded data sets at https://dbgap.ncbi.nlm.nih.gov

Secondary: Open access databases for open sharing and searching (e.g. cBioPortal Patient View)

Demographics & cancer type

Link-outs to other data systems

Time of surgery, treatments, imaging, and adverse events



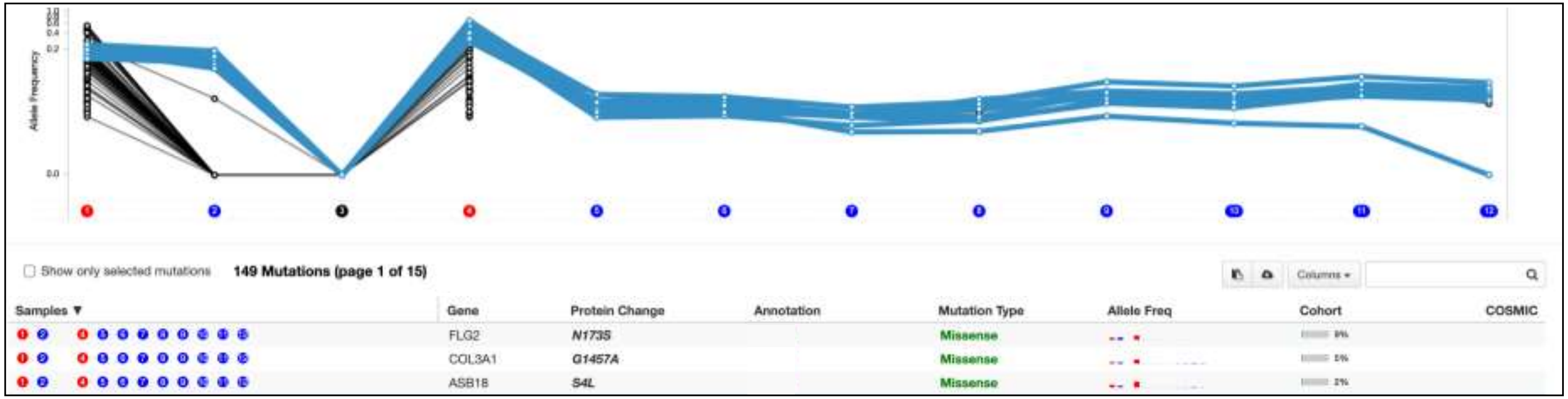
Patient: INS-A-007 , M, 70 years old, OPHSC (SCC, Tongue base), LIVING (28 months), DiseaseFree (28 months)

Samples: 1 INS-A-007-ST, Metastasis 2 INS-A-007-SB-ctDNA, cfDNA 3 INS-A-007-SB 4 INS-A-007-C2T, Metastasis 5 INS-A-007-C3B-ctDNA, cfDNA 6 INS-A-007-C6B-ctDNA, cfDNA 7 INS-A-007-C9B-ctDNA, cfDNA 8 INS-A-007-C12B-ctDNA, cfDNA 9 INS-A-007-C14B-ctDNA, cfDNA 10 INS-A-007-C15B-ctDNA, cfDNA 11 INS-A-007-C18B-ctDNA, cfDNA 12 INS-A-007-C21B-ctDNA, cfDNA

INSPIRE: Advanced Solid Tumors - Yang et al. Nat Comm 2021 (Pugh/Siu, UHN)

Circulating tumour DNA measurements

Genomic variant calls from tumours and cell-free DNA

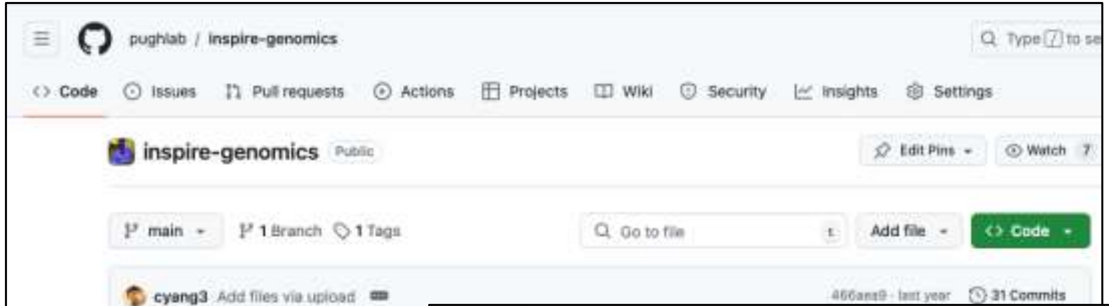


□ Show only selected mutations **149 Mutations (page 1 of 15)**

Samples ▼	Gene	Protein Change	Annotation	Mutation Type	Allele Freq	Cohort	COSMIC
1 2 4 5 6 7 8 9 10 11 12	FLG2	*N173S*		Missense		9%	
1 2 4 5 6 7 8 9 10 11 12	COL3A1	*G1457A*		Missense		5%	
1 2 4 5 6 7 8 9 10 11 12	ASB18	*S4L*		Missense		2%	

Data from Yang et al. *Nat Commun. 2021 Aug 26;12(1):5137.* cBioPortal accessible at www.cbioportal.org

Tertiary: Code repositories for reproduction of correlative analysis and figures (github and CodeOcean)



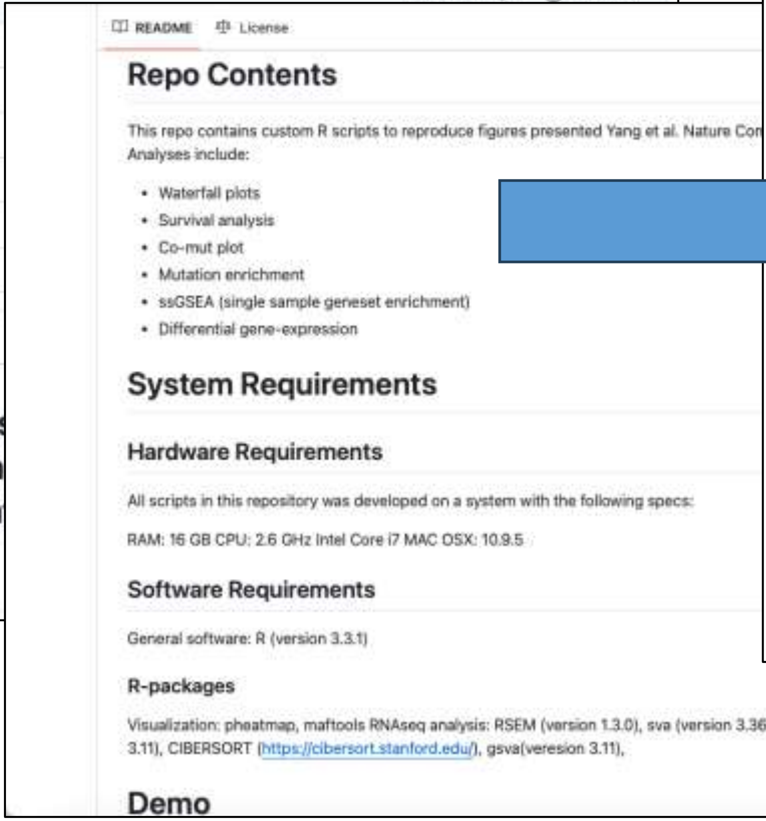
cyang3 Add files via upload 466aea9 · last year 31 Commits

- R
- VIGEX
- data
- main
- LICENSE
- README.md

README License

Pan-cancer analys
reveals genomic a
dynamics associa

Contents



Repo Contents

This repo contains custom R scripts to reproduce figures presented Yang et al. Nature Con
Analyses include:

- Waterfall plots
- Survival analysis
- Co-mut plot
- Mutation enrichment
- ssGSEA (single sample geneset enrichment)
- Differential gene-expression

System Requirements

Hardware Requirements

All scripts in this repository was developed on a system with the following specs:

RAM: 16 GB CPU: 2.6 GHz Intel Core i7 MAC OSX: 10.9.5

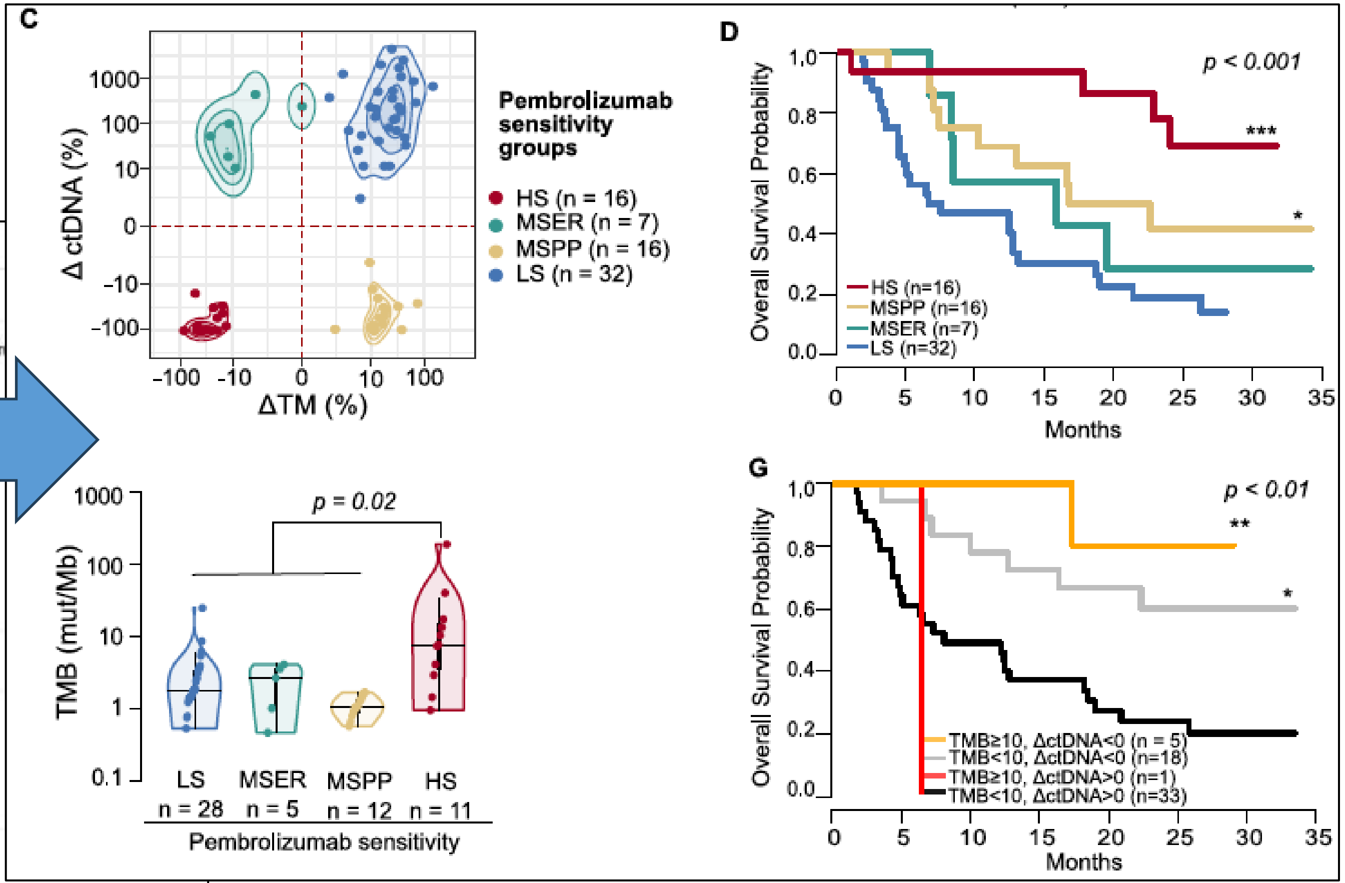
Software Requirements

General software: R (version 3.3.1)

R-packages

Visualization: pheatmap, maftools RNAseq analysis: RSEM (version 1.3.0), sva (version 3.36), DESeq2 (version 3.11), CIBERSORT (https://cibersort.stanford.edu/), gsva(veresion 3.11),

Demo



https://github.com/pughlab/inspire-genomics

Anatomy of a brief data management plan for genomics research

This project will generate comprehensive genomic and methylation profiles from a substantial number of tumour tissues, blood cells, and blood cell-free DNA collecting from consented participants on the clinical trial who have consented to data sharing. These data will be processed to produce primary (alignments), secondary (variant calls), and tertiary (interpretative analysis) data that will be shared through multiple data sharing platforms within the bounds of patient consent for use of data. Clinical data will follow the clinical data standard of the Marathon of Hope Cancer Centres Network (www.marathonofhopecancercentres.ca). ***Primary sequence alignments*** (i.e. bam files) for genomic and methylation tests of specimens from patients who have consented to genomic data sharing will be uploaded to the European Genome-Phenome Archive (EGA), a controlled access database. The Data Access Committee (DAC) for these data will be the UHN DAC[19] (EGAC00001000912) to ensure that data requests are compliance with patient consent use of the samples. EGA repository records will become publicly viewable and open for requests when a manuscript is accepted for publication and representing a cohesive, quality-controlled data set. ***Secondary variant calls*** (mutations, copy number variants, structural variants, and methylation calls) will be shared through three avenues: 1) EGA alongside the primary sequence alignments, 2) Supplementary data tables of published manuscripts, and 3) cBioPortal.org, a publicly-accessible web-based system for searching clinical and genomic data. These call files will follow standardized formats defined by the Global Alliance for Genomics and Health (GA4GH). Clinical data and copies of the genomic variant calls will also be made available as cBioPortal upload files (https://docs.cbioportal.org/file-formats/) to facilitate reuse by other studies. ***Tertiary interpretive results*** will take the form of published manuscripts, machine learning models, and a white paper for uptake by decision makers outlining potential clinical use of these assays in the context of liver transplantation for hepatocellular carcinoma. Software and code to reproduce our work will be released through open-source repositories on the Pugh Lab github site (https://github.com/pughlab). Results of our research will be reported in peer-reviewed publications in open access journals.

Introduce types and sources of data and the Primary, Secondary, Tertiary concept

Primary management with specific data standard formats and governance for gaining controlled access

Secondary management with specific repositories, URLs, and data standards

Tertiary data types including data systems to enable reproducible research

Commitment to open access

UNITY HEALTH
TORONTO

GEMINI

Canada’s largest hospital data and analytics network

Denise Mak, PhD

Director, Data Science & Innovation

GEMINI

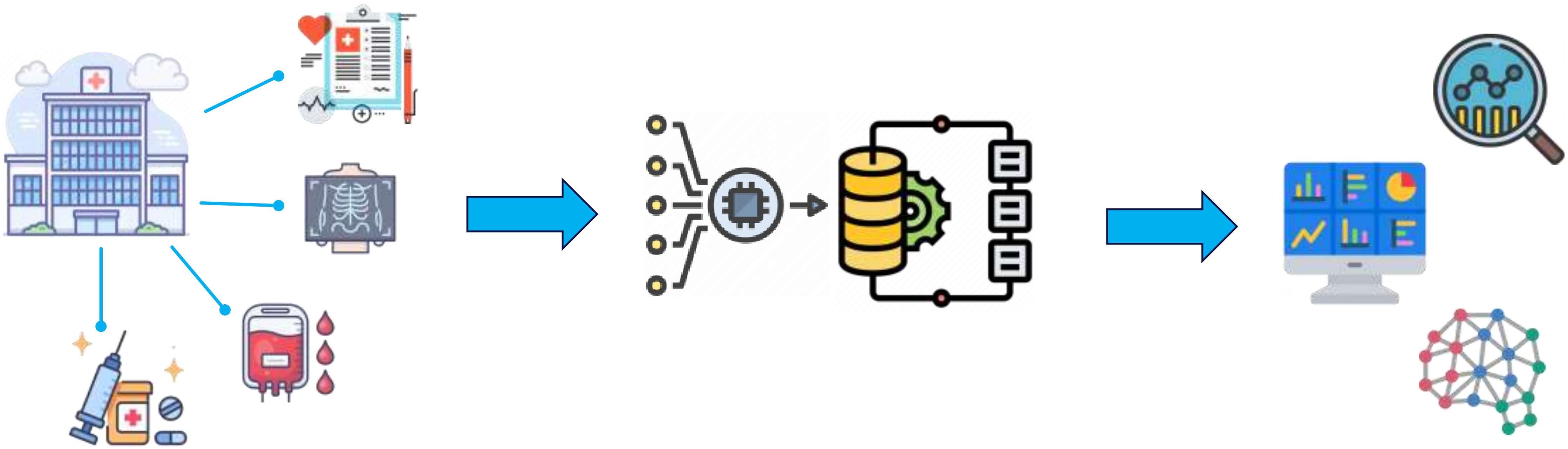
About GEMINI

- Established in 2015 to collect routinely-generated clinical and administrative data from seven (7) University of Toronto-affiliated hospitals for research and quality improvement

- Today, GEMINI currently contains data on **>2 million** hospitalizations from **>30** Ontario hospitals, representing **~60%** of all medical and ICU beds in the province

- Used by **>200 scientists & students** and **>100 active projects** to study patient care, outcomes, resource utilization and more

GEMINI Data Platform



- Electronic in-hospital patient data
- Secure data transfer to Unity Health Toronto

- Data processing pipeline (deidentification, integration, standardized, etc)
- High performance computing environment at Unity Health Toronto

- Deidentified research-ready data
- High performance computing environment at HPC4Health (Sick Kids)

GEMINI’s Data Management Practices

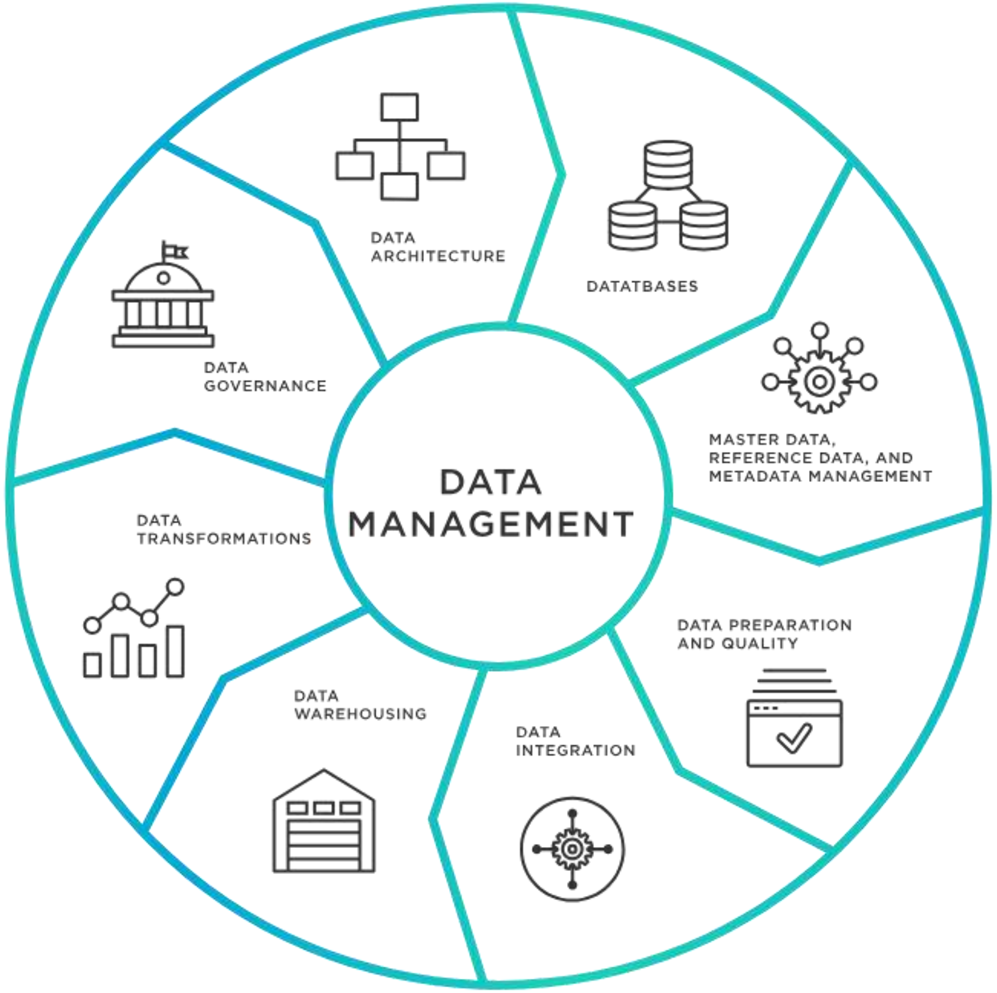
- REB study protocols
- Data sharing agreements
- Data Governance Policy
- Privacy Impact Assessment
- Security Risk Assessment
- Data Dictionary for Research Use
- Internal SOPs (Data Processing, Information Security, etc)

Alignment with DMP Components

DMP Components (based on Alliance's Template)

GEMINI	Data Collection	Documentation and Metadata	Storage and Backup	Preservation	Sharing and Reuse	Responsibilities and Resources	Ethics and Legal Compliance
REB study protocols	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Data sharing agreements	🟦		🟦		🟦	🟦	🟦
Data Governance Policy	🟦		🟦		🟦		🟦
Privacy Impact Assessment	🟦						🟦
Security Risk Assessment						🟦	
Data Dictionary for Research Use		🟦			🟦		
Internal SOPs	🟦	🟦	🟦	🟦	🟦	🟦	

Data Management Practices



https://www.tibco.com/glossary/what-is-data-management

Benefits and Challenges with DMPs

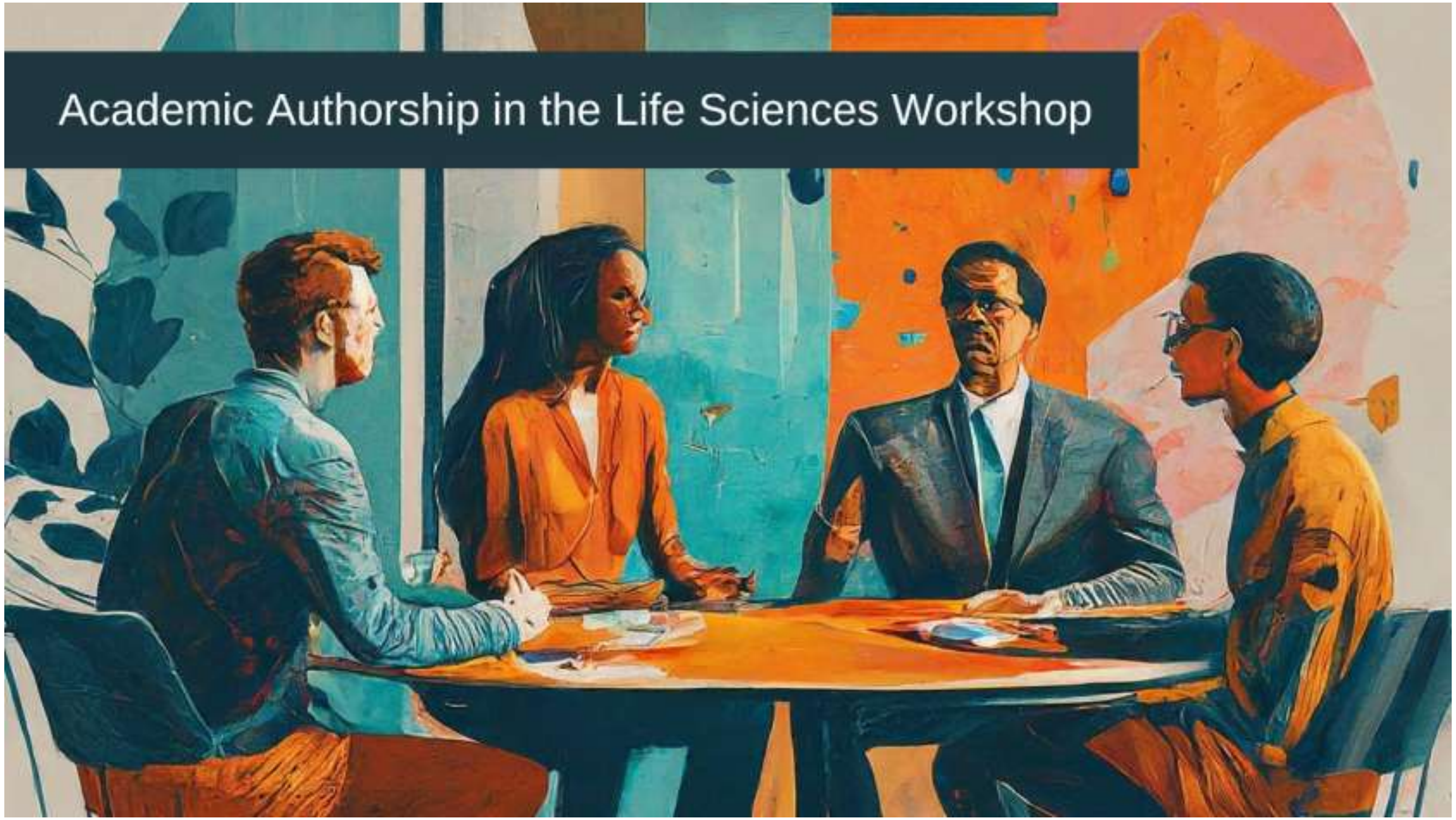
- ✅ Be proactive about Data Management
- ✅ Align and unify existing approvals, policies, and other documents
- ✅ Improve understanding and appreciation of data workflow

- 🚩 Be too prescriptive with templates
- 🚩 Find appropriate and relevant technical resources
- 🚩 Initiate DMPs for new projects

Panel Discussion

Please use the chat

Upcoming Event



October 16th @ 2:30 pm – 4 pm

October 23rd @ 2:30 pm – 4 pm

Thank you

- A link to the recording, presenter slides, and feedback form will be sent out after the session
- Follow-up questions can be addressed to cris@utoronto.ca

Dr. Victoria Hodgkinson

Dr. Trevor Pugh

Dr. Denise Mak