Breakout Session 4: Track B

Generating AI / ML-Ready Data for Type 1 Diabetes

Dr. Bobbie-Jo Webb-Robertson
Division Director, Biological Sciences, Pacific Northwest National Laboratory

Pacific Northwest
NATIONAL LABORATORY

Generating AI/ML-Ready Data for Type 1 Diabetes

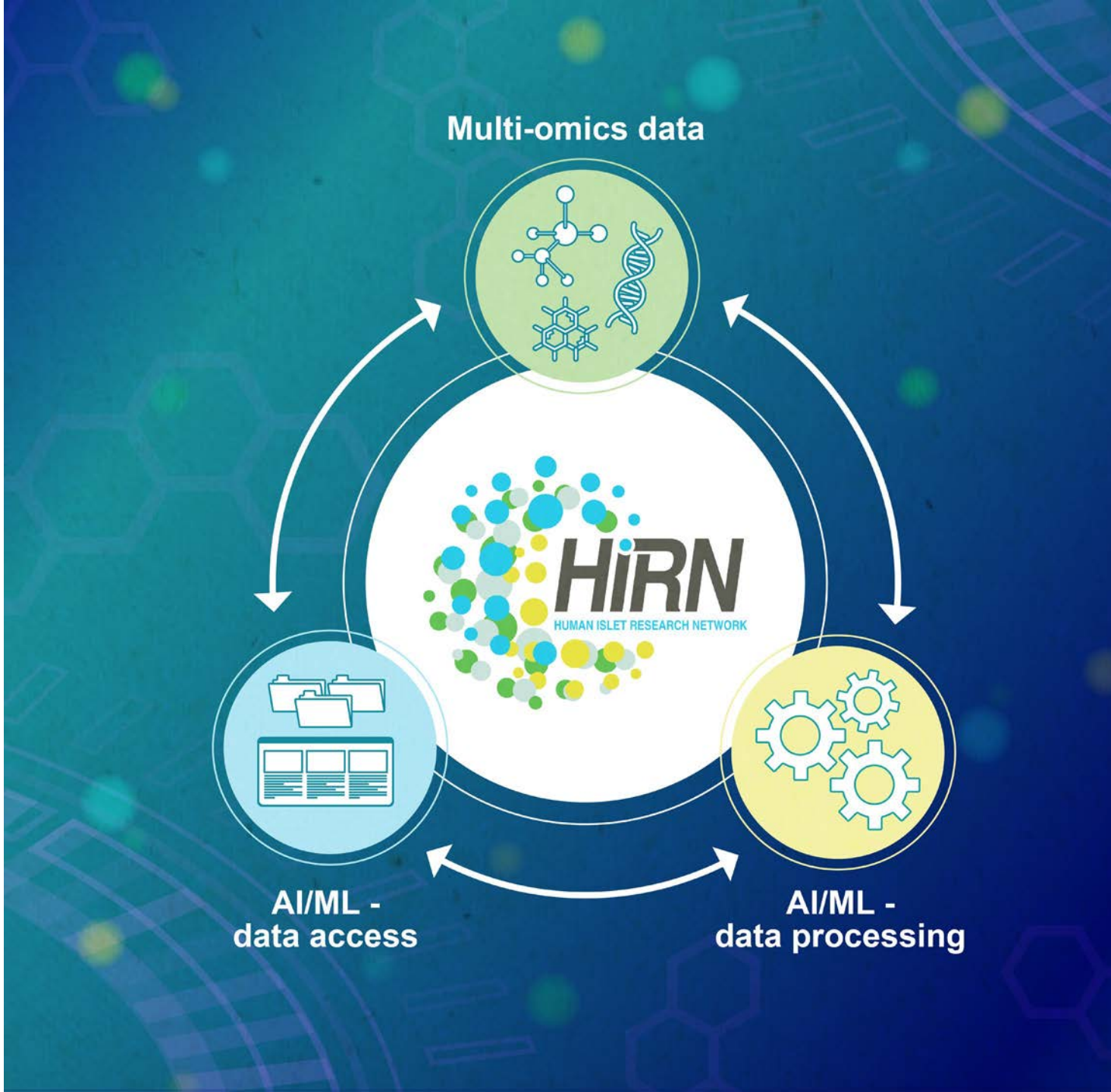
June 17, 2020

March 27, 2024

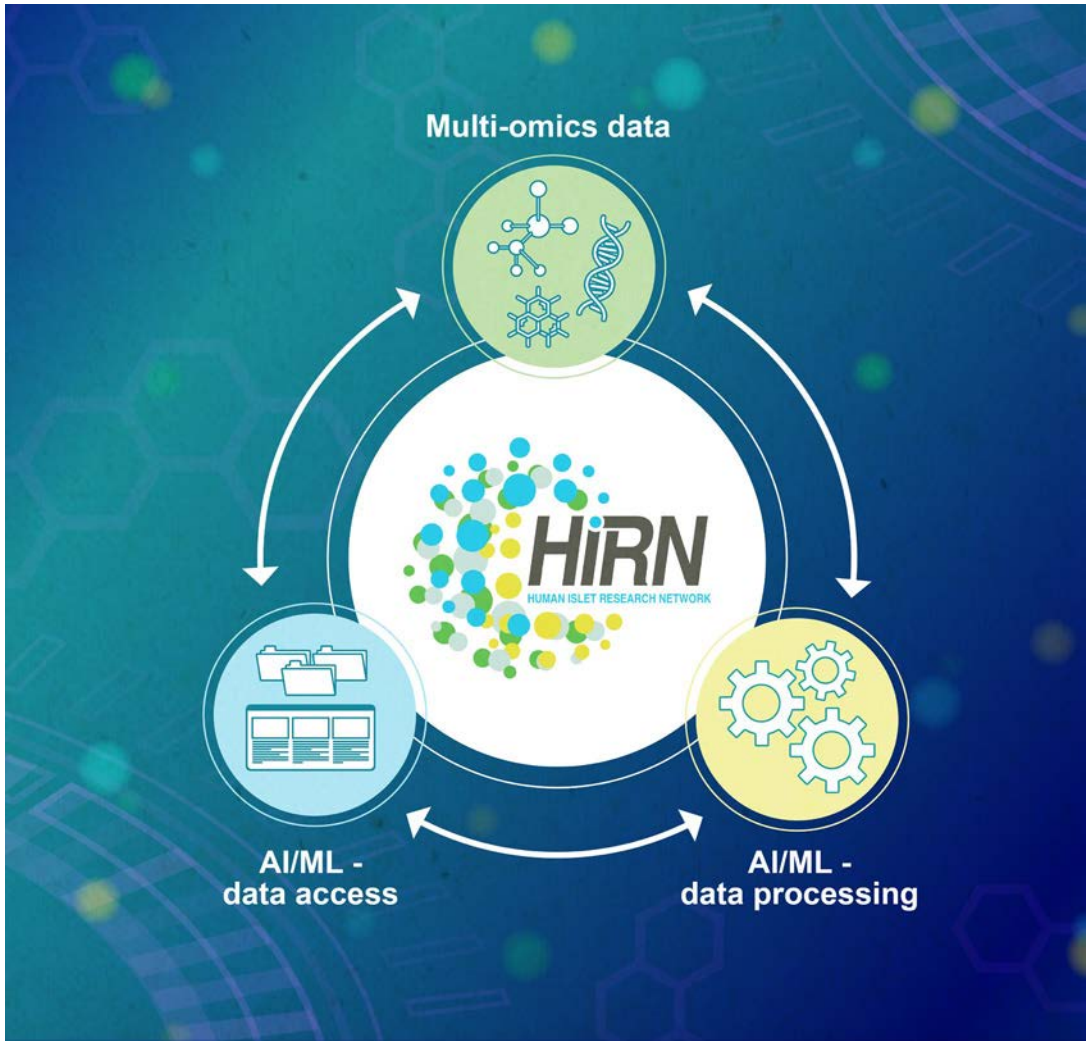
Bobbie-Jo Webb-Robertson, PhD
Raghu Mirmira, MD, PhD

U.S. DEPARTMENT OF ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy



Focus on Type 1 Diabetes Data



- Generation of AI/ML ready omics data with appropriate meta-data to improve pre-processing of omics data
- Generation of multi-omics AI/ML ready data with appropriate clinical and immunologic meta-data for testing new methods in biomarker discovery and validation.

Pacific Northwest NATIONAL LABORATORY

Properties of AI/ML-Ready Data

Cleaned and processed data that is in a usable format that can be applied to an AI/ML application

Quality	Documentation	Access	Preparation
• Data is consistently formatted from a one-time step or data file to the next	• There is support and context associated with the data or domain	• Data is available in a variety of formats and delivery options	• Data has gone through preprocessing steps to support AI/ML tools/software

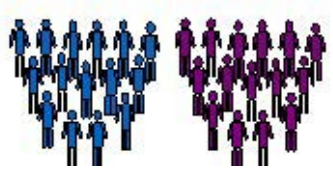
Much of the text above is directly quoted from : https://bipartisanpolicy.org/explainer/ai-ready-open-data/

Alternative Splicing as a T1D biomarker

HiRN
HUMAN ISLET RESEARCH NETWORK

Training Set

Alternative Splicing Events in Human Blood
(180M Reads)



Healthy Control (n=12) New Onset 1D (n=12

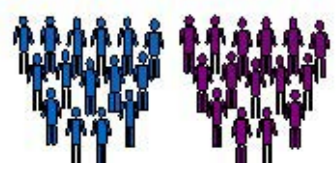
Healthy Controls
8 Male, 4 Female
Age: 12±4
BMI: 19.5±2.5

New Onset T1D
8 Male, 4 Female
Age: 12±4
BMI: 19.2±5.1

Boot-strap Random Forest Model

Validation Set

Alternative Splicing Events in Human Blood
(150M Reads)



Healthy Control (n=12) New Onset 1D (n=12)

Healthy Controls
8 Male, 4 Female
Age: 11±5
BMI: ?

New Onset T1D
8 Male, 4 Female
Age: 11±5
BMI: ?

Genetic Risk

Starting Point

15x increased risk of T1D in those with relatives of disease

Immune Activation

Immune Activation

Immune Response

Seroconversion

STAGE 1

Normoglycemia
≥ 2 autoantibodies

STAGE 2

Dysglycemia
≥ 2 autoantibodies

STAGE 3

Clinical Diagnosis
≥ 2 autoantibodies

STAGE 4

Established/ Long-standing T1D

Type 1 Diabetes TrialNet

Model Card Example

Human Islet Research Network (HIRN): Alternative Splicing Events, Random Forest Model Card

Javier E. Flores

2023-01-26

Data

Inclusion levels of alternative splicing (AS) events of five different varieties (i.e. skipped exon (SE), retained intron (RI), alternative 5' splice site (A5SS), alternative 3' splice site (A3SS), and mutually exclusive exons (MXE)) were measured in human blood samples from two separate cohorts of patients.

Cohort 1 (Training Cohort):

- 12 healthy controls; 12 new onset type 1 diabetic (T1D) cases
- cases and controls matched on biological sex, age, and body mass index (BMI)
- 180 million reads

Cohort 2 (Testing Cohort):

- 12 healthy controls; 12 new onset type 1 diabetic (T1D) cases
- cases and controls matched on biological sex and age. BMI not recorded.
- 150 million reads

Event	Total Events (Cohort 1)	Total Events (Cohort 2)	Total Events (Shared)
Skipped exon (SE)	104590	69597	56530
Retained intron (RI)	4768	4158	4088
Alternative 5' splice site (A5SS)	5544	4169	3919
Alternative 3' splice site (A3SS)	8521	6374	6001
Mutually exclusive exon (MXE)	20666	12064	8332

Approach

Model: Random Forest

- Implemented in R using the tidymodels and ranger packages.

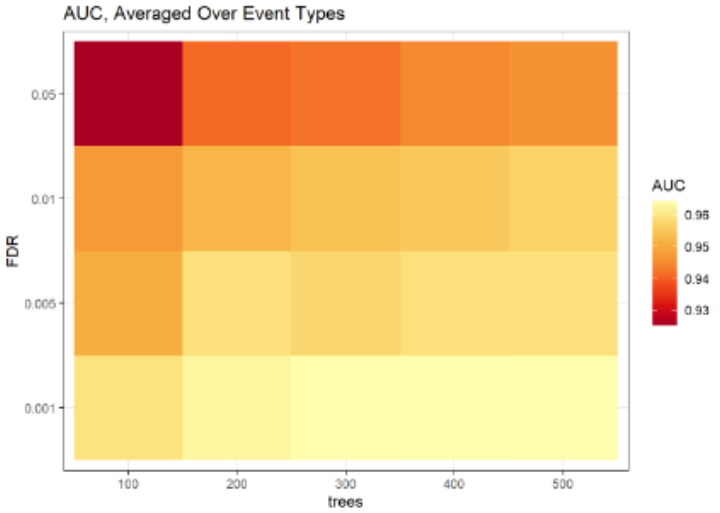
Preprocessing: Event data in Cohort 1 that are missing in Cohort 2 are imputed based on the means of the Cohort 1 data.

Tuning: Grid-search

- Repeated 3-fold cross-validation with 25 repeats
- Tuned over the number of trees (100, 200, 300, 400, 500) and false discovery rate (FDR) threshold (0.05, 0.01, 0.005, 0.001)
- Other model hyperparameters (i.e. the number of randomly selected predictors and the minimal node size) were kept at software defaults
- Area-under-the-curve (AUC) was used as the selection metric

Final Model:

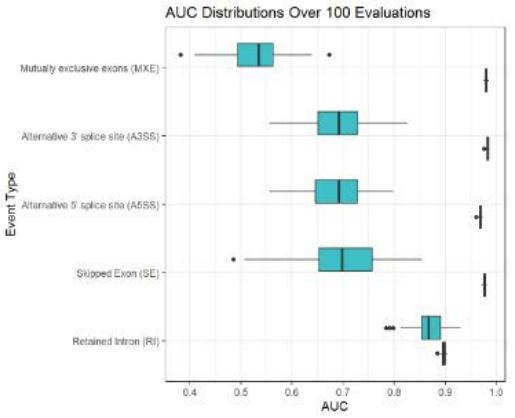
- 300 trees; FDR threshold of 0.001
- Evaluated on training data through repeated 3-fold cross-validation with 100 repeats
- Evaluated on (mean-imputed) testing data
- Evaluations on training and test data repeated 100 times
- AUC used as the evaluation metric

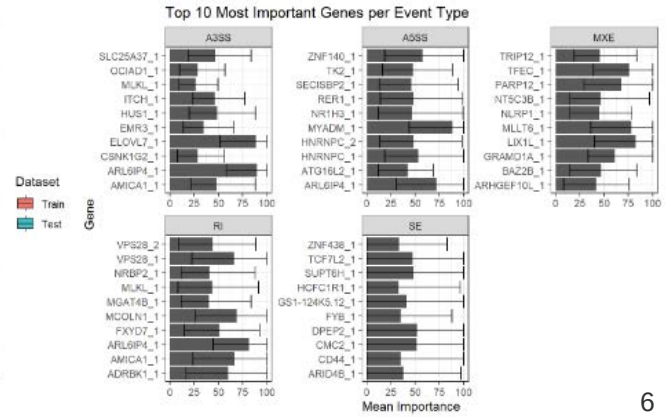


Data are accessible on DataHub. Code for data processing, model tuning, and final model fitting/evaluation is available on GitHub.

Results

Event	Event Count	AUC, Training (95% CI)	AUC, Test (95% CI)
Retained Intron (RI)	370	0.897 (0.889, 0.904)	0.869 (0.799, 0.913)
Skipped Exon (SE)	1872	0.977 (0.972, 0.981)	0.695 (0.524, 0.837)
Alternative 5' splice site (A5SS)	179	0.969 (0.964, 0.973)	0.69 (0.583, 0.781)
Alternative 3' splice site (A3SS)	273	0.983 (0.979, 0.986)	0.688 (0.569, 0.778)
Mutually exclusive exons (MXE)	251	0.981 (0.977, 0.985)	0.53 (0.427, 0.612)





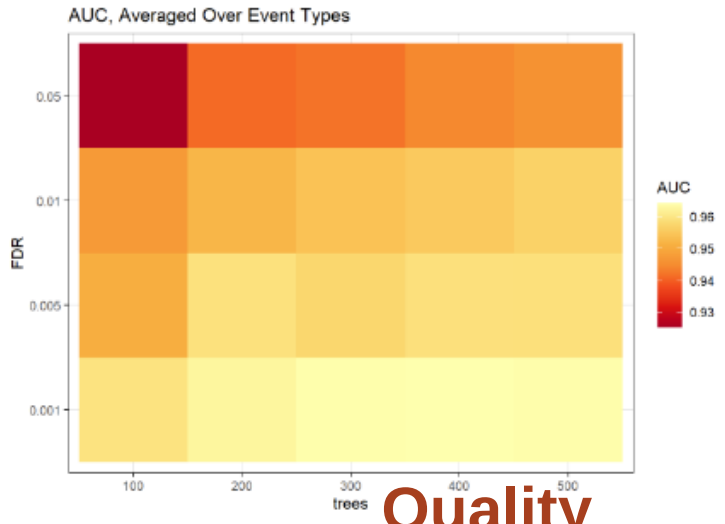
Model Card Example

Approach

Model: Random Forest

- Implemented in R using the tidymodels and ranger packages.

Preprocessing: Event data in Cohort 1 that are missing in Cohort 2 are imputed based on the means of the Cohort 1 data.



Human Isle… Splicing Ev…

Javier E. Flores

2023-01-26

Data

Inclusion levels of alternative s… (A5SS), alternative 3' splice sit… of patients.

Cohort 1 (Training Cohort):

- 12 healthy controls; 12 r…
- cases and controls matc… mass index (BMI)
- 180 million reads

Cohort 2 (Testing Cohort):

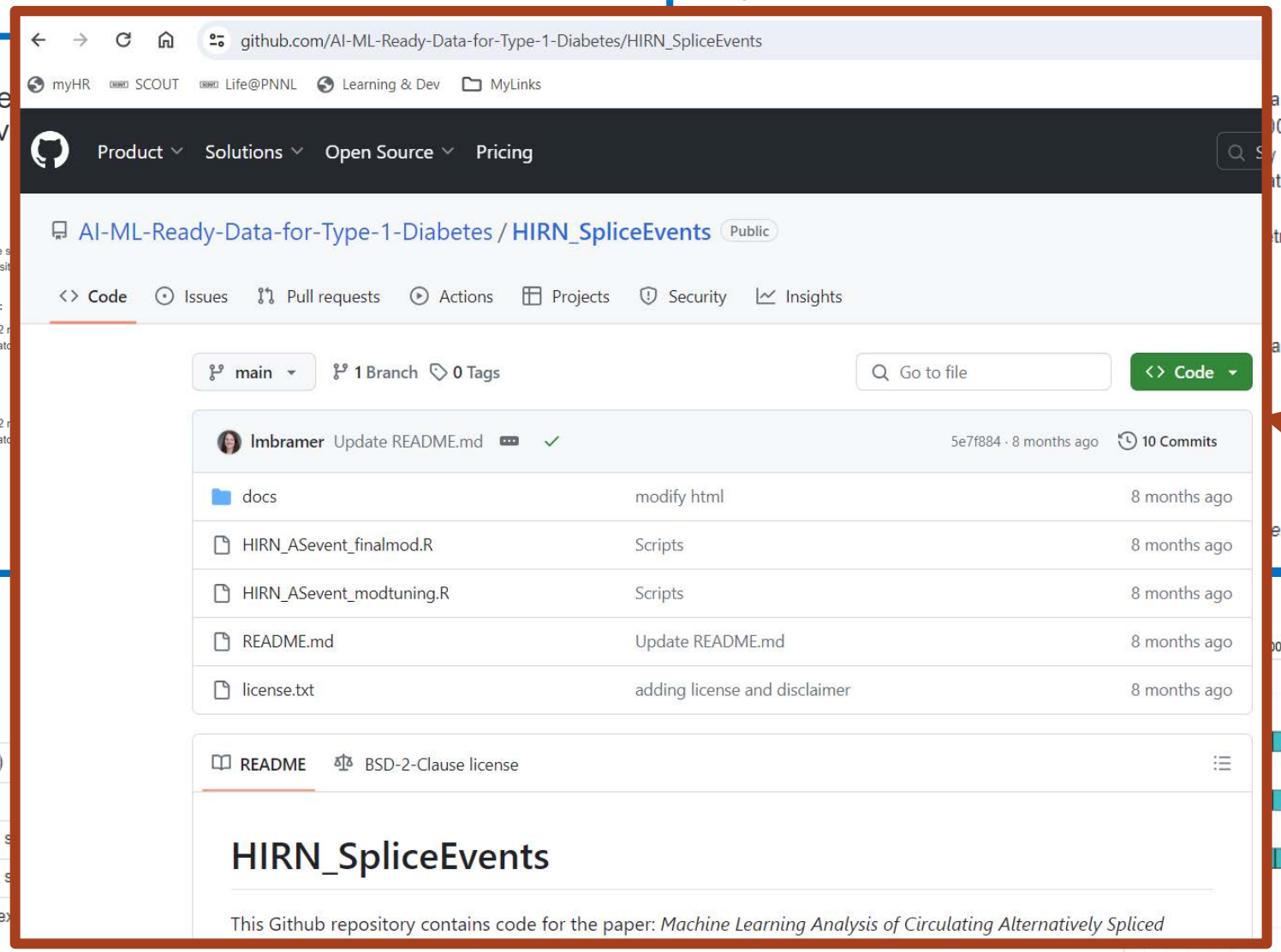
- 12 healthy controls; 12 r…
- cases and controls matc… recorded.
- 150 million reads

…and …01)

…t

…tric

…alidation

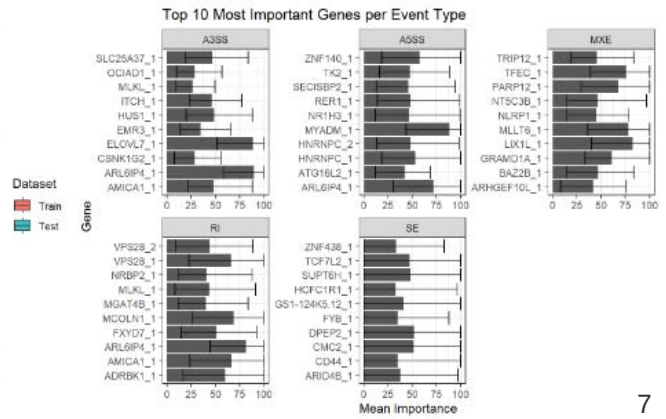


Quality Preparation

…el tuning, and final model fitting/evaluation is avail…ble on GitHub.

Results

Event
Retained Intron (RI)
Skipped Exon (SE)
Alternative 5' splice s…
Alternative 3' splice s…
Mutually exclusive ex…

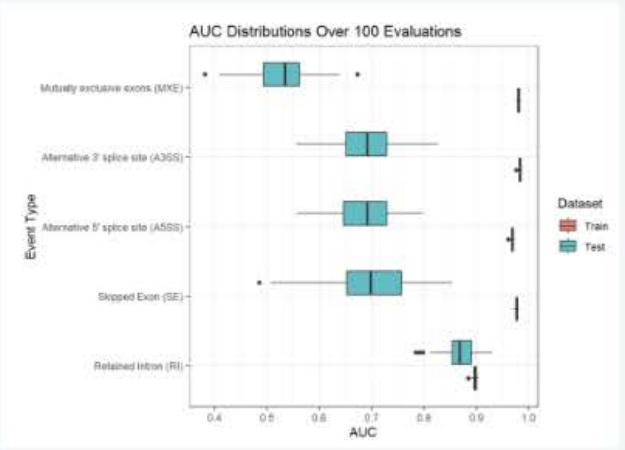


DATASET - Transcriptomics

Human Islet Research Network (HIRN): Alternative Splicing Events

BIOLOGY
HUMAN HEALTH
DATA ANALYTICS & MACHINE LEARNING

Download



TYPE 1 DIABETES
ALTERNATIVE SPLICING
MACHINE LEARNING
PREDICTIVE MODELING

Inclusion levels of alternative splicing (AS) events of five different varieties (i.e. skipped exon (SE), retained intron (RI), alternative 5' splice site (A5SS), alternative 3' splice site (A3SS), and mutually exclusive exons (MXE)) were measured in human blood samples from two separate cohorts of patients.

Cohort 1 (Training Cohort):

- 12 healthy controls; 12 new onset type 1 diabetic (T1D) cases
- cases and controls matched on biological sex, age, and body mass index (BMI)
- 180 million reads

Cohort 2 (Testing Cohort):

- 12 healthy controls; 12 new onset type 1 diabetic (T1D) cases
- cases and controls matched on biological sex and age. BMI not recorded.
- 150 million reads

Language
English

Data Dictionary

HIRN Splice Event Data Guide.docx

Projects (1) +

data.pnnl.gov/group/nodes/dataset/33494

myHR SCOUT Life@PNNL Learning & Dev MyLinks

DATASET - Transcriptomics

Human Islet Research Netwo

BIOLOGY
HUMAN HEALTH
DATA ANALYTICS & MACHINE LEARNING

TYPE 1 DIABETES
ALTERNATIVE SPLICING
MACHINE LEARNING
PREDICTIVE MODELING

Inclusion levels of alternative splicing (AS) events of fi
(i.e. skipped exon (SE), retained intron (RI), alternativ
alternative 3' splice site (A3SS), and mutually exclusiv
measured in human blood samples from two separate

Cohort 1 (Training Cohort):

- 12 healthy controls; 12 new onset type 1 diabetic (T
- cases and controls matched on biological sex, age,
- 180 million reads

Cohort 2 (Testing Cohort):

- 12 healthy controls; 12 new onset type 1 diabetic (T
- cases and controls matched on biological sex and a
- 150 million reads

Language
English

Data Dictionary

HIRN Splice Event Data Guide.docx

Projects (1)

1 of 7

HIRN Splice Event Data Guide

Overview

The following R dataframes (.rds) are contained within the directory:

1. ***a3ss_data.rds***: contains response and predictor data for all measured A3SS-type splice events
2. ***a3ss_metadata.rds***: contains associated metadata for all measured A3SS-type splice events
3. ***a5ss_data.rds***: contains response and predictor data for all measured A5SS-type splice events
4. ***a5ss_metadata.rds***: contains associated metadata for all measured A5SS-type splice events
5. ***mxe_data.rds***: contains response and predictor data for all measured MXE-type splice events
6. ***mxe_metadata.rds***: contains associated metadata for all measured MXE-type splice events
7. ***ri_data.rds***: contains response and predictor data for all measured RI-type splice events
8. ***ri_metadata.rds***: contains associated metadata for all measured RI-type splice events
9. ***se_data.rds***: contains response and predictor data for all measured SE-type splice events
10. ***se_metadata.rds***: contains associated metadata for all measured SE-type splice events

A3SS refers to an alternative 3' splice junction being used in the alternative splicing; A5SS to an alternative 5' splice junction; MXE denotes a mutually exclusive exon event; RI a retained intron event; and SE a skipped exon event.

Dataset details

All data/metadata .rds pairs are formatted the same and contain largely the same set of variables, only specific to the corresponding splicing event. Nonetheless, descriptions of all of the contents of each dataset are subsequently provided.

1. ***a3ss_data.rds***
   a. 8894 rows: Each corresponds to a unique A3SS splice event
   b. 265 columns:
      i. **Status_TRAIN_1 – Status_TRAIN_24**: columns containing the response data (i.e. case-control status) for each of the 24 samples in the training cohort.
      ii. **MergeID**: Column containing the unique identifier for each splice event. This variable is used to merge a3ss_data.rds and a3ss_metadata.rds.
      iii. **Status_TEST_1 – Status_TEST_24**: columns containing the response data (i.e. case-control status) for each of the 24 samples in the testing cohort.
      iv. **IncLevel_TRAIN_1 – IncLevel_TRAIN_24**: columns containing inclusion level predictor data for each of the 24 training samples.
      v. **IncLevel_TEST_1 – IncLevel_TEST_24**: columns containing inclusion level predictor data for each of the 24 test samples, with missing values not imputed.
      vi. **IncLevel_TRAIN_imputed_1 – IncLevel_TRAIN_imputed_24**: columns containing inclusion level predictor data for each of the 24 test samples, with missing values imputed based on the average of the observed training sample data.
      vii. **IJC_TRAIN_1 – IJC_TRAIN_24**: columns containing inclusion junction count predictor data for each of the 24 training samples.

viii. **IJC_TEST_1 – IJC_TEST_24**: columns containing inclusion junction count

Documentation Preparation

data.pnnl.gov/group/nodes/dataset/33494

myHR SCOUT Life@PNNL Learning & Dev MyLinks

Access

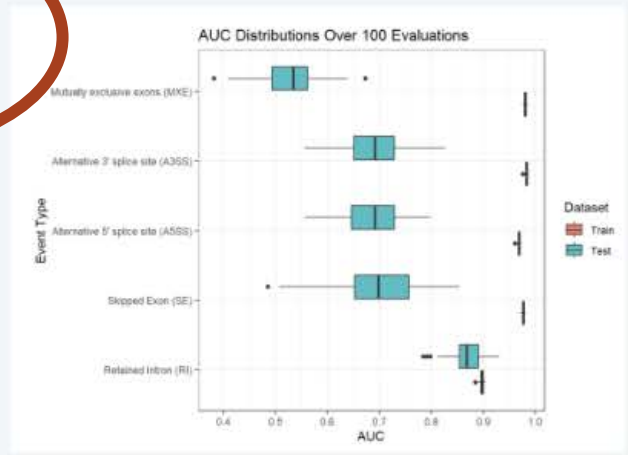
DATASET - Transcriptomics

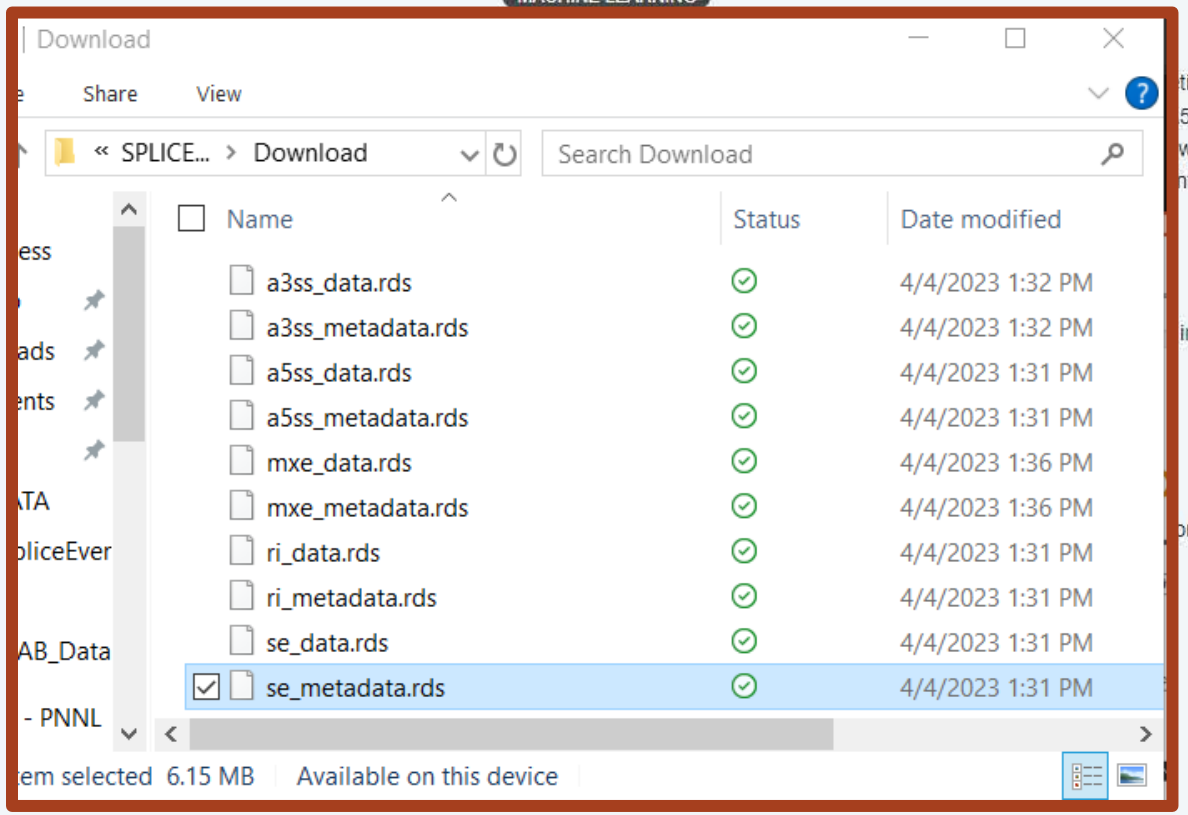
Human Islet Research Network (HIRN): Alternative Splicing Events

BIOLOGY
HUMAN HEALTH
DATA ANALYTICS & MACHINE LEARNING

TYPE 1 DIABETES
ALTERNATIVE SPLICING
MACHINE LEARNING

Download





ties
5SS),
were
nts.

index (BMI)

orded.

Projects (1)

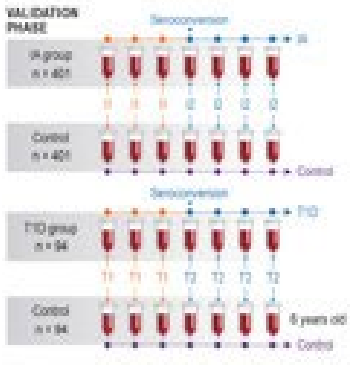
Benchmark Proteomic Data for Batch Correction

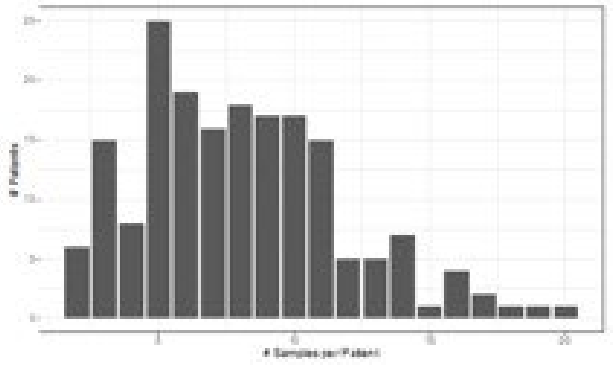
TEDDY
The Environmental Determinants of Diabetes in the Young

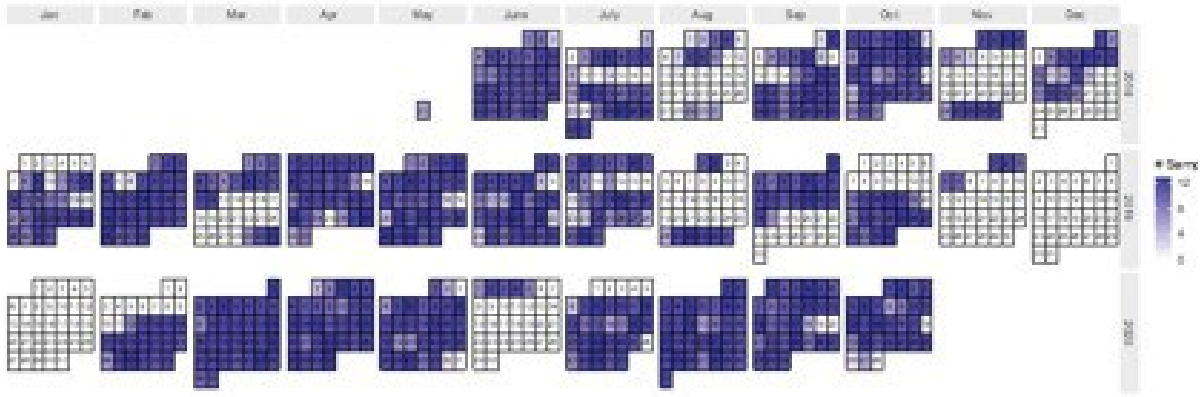
- Quality control (QC) samples were comprised of 6 pooled plasma samples from TEDDY and 1 commercial pooled plasma sample from BioIVT
- 811 peptides were selected for proteomics assay development with 694 successfully monitored

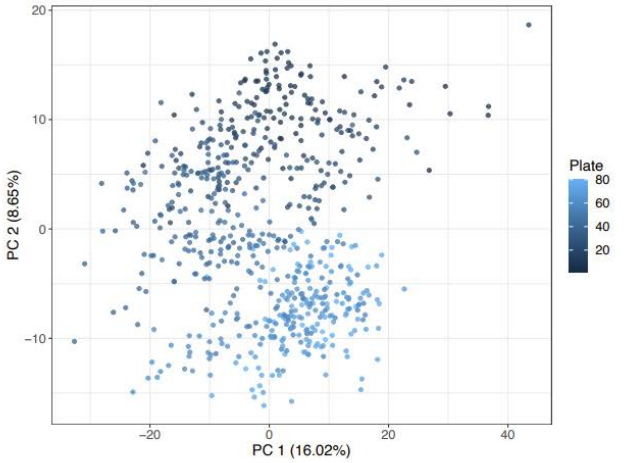
Study Design

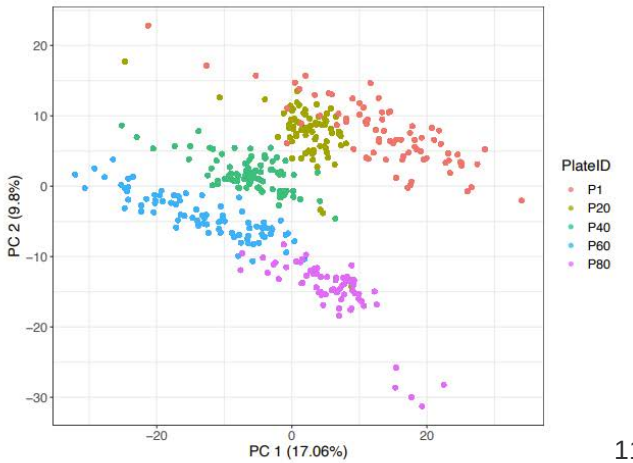
- Nested case-control study from TEDDY is comprised of over 8,000 individuals from 7 centers (Germany, Sweden, and Finland in Europe; and Denver, Georgia, Florida, and Washington in the USA) from the ages of 0–6 years old.
- From this cohort, we selected 401 individuals who developed islet autoimmunity (IA) and 94 who developed Type 1 diabetes (T1D), each paired to a matched control.











Single Slide of DataHub – Just to show consistency in approach

DATAHUB
Pacific Northwest

Search by keywords, authors and much more...

Log in

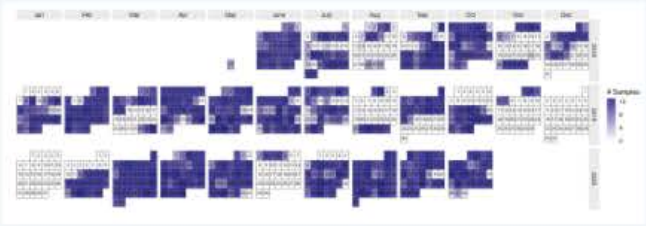
Categories Datasets Data Sources Projects Publications People

DATASET - Proteomics

TEDDY Targeted Proteomics Study Data

SCIENTIFIC DISCOVERY
BIOLOGY
HUMAN HEALTH
COMPUTATIONAL MATHEMATICS & STATISTICS
DATA ANALYTICS & MACHINE LEARNING

Download



MACHINE LEARNING

Comprised of 6,426 sample runs, The Environmental Determinants of Diabetes in the Young (TEDDY) proteomics validation study constitutes one of the largest targeted proteomics studies in the literature to date. Making quality control (QC) and donor sample data available to researchers aligns with TEDDY's commitment to sharing data with the scientific community. The data presented here can be used as a resource for new computational method developments such as batch correction as well as benchmarking and comparing the performance of different methods/tools.

Language
English

Projects (2) −

TEDDY AI/ML Ready Datasets and Models →
HIRN AI/ML Ready Datasets and Models →

Pacific Northwest NATIONAL LABORATORY

Challenges

- Evaluating data sources as adequate for AI/ML
  - Sample size and Replication
  - Quality
  - Source (genomics, proteomics, etc.)
  - Defining level of data to capture
- Integrating data release, notes, code

AI/ML-ready data is following a set of principles, there is no standard

Future Work

HiRN HUMAN ISLET RESEARCH NETWORK

DAISY

TEDDY The Environmental Determinants of Diabetes in the Young

nPOD Network for Pancreatic Organ Donors with Diabetes

Autoimmunity Screening for Kids ask A SIMPLE TEST TO DETECT Type 1 Diabetes (T1D) + Celiac

Type 1 Diabetes TrialNet

Acknowledgements

Raghu Mirmira (University of Chicago)

Tom Metz (PNNL)

Ernesto Nakayasu (PNNL)

Wenting Wu (Indiana University)

Lisa Bramer (PNNL)

Javier Flores (PNNL)

TEDDY
The Environmental Determinants of Diabetes in the Young

And many more….

Pacific Northwest
NATIONAL LABORATORY

Thank you