AUTOMATED AND SCALABLE CLOUD-BASED COMPUTATIONAL PIPELINE FOR LARGE-SCALE UNBIASED PLASMA PROTEOMICS STUDY

Joon-Yong Lee, Jinlyung Choi, Sara Nouri Golmaei, Yuntao Hu, Sai Ramaswamy, Dijana Vitko, Wan-Fang Chou, Megan Mora, Jessica Chan, Mark Marispini, Benjamin Ta, Peter Spiro, Hoda Malekpour, Ajinkya Kokate, Robert Zawada, Purva Ranjan, Bruce Wilcox, Chinmay Belthangady, Manway Liu, Philip Ma
PrognomiQ, Inc., San Mateo, CA

INTRODUCTION

- Liquid-chromatography/mass-spectrometry (LC/MS) for untargeted plasma proteomics has become a valuable method to identify potential cancer biomarkers in liquid biopsy samples
- However, discovering reliable cancer biomarkers in plasma requires extensive proteomic profiling of numerous samples to obtain statistically robust and generalizable results while accounting for confounding factors
- Automated and scalable workflows are lacking, posing a challenge to the application of LC/MS to large-scale studies
- To address this challenge, we have developed a highly flexible, cloud-based pipeline for deep and unbiased proteomic profiling of Proteograph™ (Seer Inc.)-processed plasma samples utilizing EvosepOne coupled to Bruker timsTOF HT

METHODS

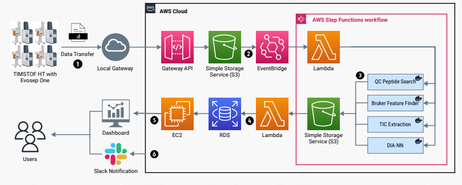
- We collected >3,000 plasma samples as part of an ongoing cancer biomarker discovery study
- All samples were processed with the Proteograph™ 5-nanoparticle workflow to achieve deeper plasma proteome coverage (>5,000 protein groups)
- We conducted data-independent acquisition schemes parallel accumulation-serial fragmentation (dia-PASEF®) analysis on each processed sample using an EvosepOne LC system coupled with a Bruker timsTOF HT mass spectrometer, acquiring data with a 21-minute gradient
- We leveraged Amazon Web Services (AWS) infrastructure to build an automated and scalable computational pipeline (**Figure 1**) including:
  - Automated data transfer
  - Quality Control (QC) data processing
    - QC peptide search
    - Chromatogram extraction
    - Feature finding
    - Peptide/protein identification
    - Notification logic
    - Web-based user interface

KEY RESULTS

- An automated and scalable pipeline was built in the cloud to support the large-scale study with real-time data processing from multiple timsTOF HT instruments (**Figure 1**)
- A computational pipeline for QC data processing and QC metric ingestion was implemented with the AWS Step function (**Figure 2**)
- The automated pipeline significantly improved data acquisition workflow (**Figure 3**)
- The utilization of parallel mode in the AWS Step function for data-independent acquisition-neural network (DIA-NN) search resulted in a significant improvement in data processing time; processing 1,000 files now only takes ~2 hours, whereas it previously took about 2-3 days on a local workstation (**Figure 4**)

RESULTS

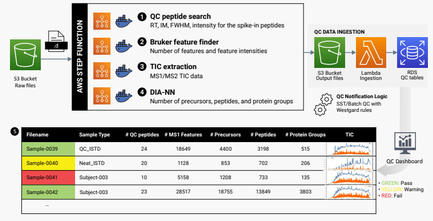
FIGURE 1. Overall data workflow with the AWS infrastructure for an automated data transfer and computational pipeline.



API, Application Programming Interface; AWS, Amazon Web Services; DIA-NN, data-independent acquisition neural network; EC2, Amazon Elastic Compute Cloud; QC, quality control; RDS, Relational Database Service; TIC, total ion current.

- Overall Data Workflow (**Figure 1**)
  1. Raw .D files from multiple instruments are automatically transferred to AWS Simple Storage Service (S3) via the AWS Gateway upon file generation
  2. Upon transfer of raw .D files into S3, the AWS Step function for data processing pipeline is triggered by S3 events
  3. Output data from each processing unit are then stored in S3
  4. All the QC-related information is automatically ingested into AWS Relational Database Service (RDS)
  5. Users can monitor QC and other metrics for each sample run in real-time through a web-based dashboard hosted in the cloud
  6. All the custom events from AWS (e.g., job failures) are notified via Slack messages to users

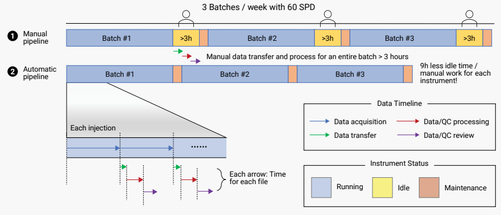
FIGURE 2. Computational pipeline for QC metrics with the AWS Step function and QC data ingestion.



AWS, Amazon Web Services; DIA-NN, data-independent acquisition neural network; FWHM, full width at half maximum; IM, ion mobility; ISTD, internal standard; MS1/MS2, first/second stage mass spectrometry; QC, quality control; RDS, Relational Database Service; RT, retention time; S3, Simple Storage Service; SST, system sustainability test; TIC, total ion current.

- The pipeline consists of 4 independent modules 1 through 4, which are containerized and executed as the AWS Batch job in parallel
- All resulting QC data are automatically ingested into the database
- The QC Dashboard displays the QC metrics for each injection in a table format 5, and the QC notification logic evaluates the data quality based on historical data to provide users with a green(Pass)/yellow(Warning)/red(Fail) flag for quick evaluation

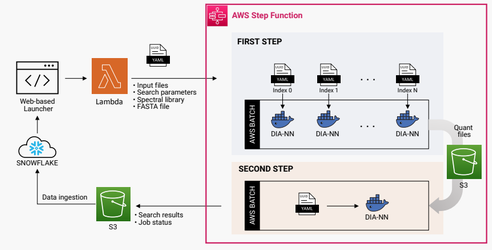
FIGURE 3. Data acquisition timeline for manual and automated pipelines.



h, hours; QC, quality control; SPD, samples per day.

- In the manual pipeline without automation, human operators are required to manually upload and download data files from each instrument to the processing computer environment 1
- However, in the automated pipeline, file transfer and data processing occur simultaneously and data review can be conducted in real-time 2
- Standardized file naming, data processing parameters, and worklist files in the automated pipeline minimized human errors and improved throughput

FIGURE 4. Reproducible and scalable data processing pipeline for DIA-NN search in parallel.



DIA-NN, data-independent acquisition-neural network; S3, Simple Storage Service.

- Users easily initiate a DIA-NN search job by clicking a button on a web-based application
- This web-based launcher has a feature for users to query file locations in S3 for data files of interest from the executed worklist data in Snowflake database
- When the user launches a job, a YAML file is generated automatically and saved in S3 with a unique run ID
- The AWS step function is then used to start *N* number of DIA-NN batch jobs for *N* number of input files in parallel and collect the quant files generated from the initial search
- The second step is then executed, leveraging these quant files to produce the final DIA-NN outputs
- All outputs are saved automatically in S3 and Snowflake for downstream analysis and can be tracked easily using the unique run ID

CONCLUSIONS

- We created a cloud-based computational pipeline by leveraging AWS infrastructure to automatically transfer and process data at scale in support of large-scale untargeted proteomics studies
- The pipeline capability to process data and provide quality control in real time was demonstrated with a >3000 subject biomarker study comprising >20,000 injections
- We improved the efficiency of data acquisition workflow by reducing labor-intensive manual work and idle-time of LC/MS systems using this automated pipeline

DISCLOSURES:
Study funded by PrognomiQ, Inc. All authors are current or former employees of PrognomiQ, Inc.

ACKNOWLEDGEMENTS:
Funded by PrognomiQ, Inc (San Mateo, CA). Editorial and graphical assistance provided by Prescott Medical Communications Group (Chicago, IL)