Applied & Translational Genomics 6 (2015) 42–44

Contents lists available at ScienceDirect

Applied & Translational Genomics

journal homepage: www.elsevier.com/locate/atg

tranSMART Foundation Datathon 1.0: The cross neurodegenerative diseases challenge

Carol Isaacson Barash [a], Keith Elliston [b], Rudy Potenzone [b]

[a] Helix Health Advisors, United States
[b] tranSMART Foundation, United States

1. Introduction

The tranSMART Foundation's inaugural Datathon took place June 30–July 2 at the Thomson Reuters offices in Boston, MA. The overall aim of the Datathon was to determine the feasibility of using the tranSMART platform to explore multiple large datasets on a customized cloud server to support a Datathon that could generate new research findings. The goal of this Datathon was to identify similarities and differences across different neurodegenerative diseases, specifically Alzheimer's disease and Parkinson's disease, and to discover new insights into these diseases.

Specific objectives were to identify:

- Common biomarker changes across Parkinson and Alzheimer disease
- Common pathway changes across Parkinson and Alzheimer disease
- The normal distribution of imaging and fluid biomarkers across controls
- Novel hypotheses, research findings or conclusions about these neurodegenerative diseases.

2. Design

The tranSMART Foundation, the Michael J. Fox Foundation, the University of Luxembourg and the University of Michigan worked together with the Laboratory of Neuro Imaging (LONI) to install tranSMART v1.2.4 on cloud servers at LONI, and to install the 14 datasets to be used for the Datathon. The ADNI, PPMI, LRRK2 and BioFIND datasets were curated and loaded by Thomson Reuters, working with the Michael J. Fox Foundation. Due to restrictions on access to and redistribution of the ADNI, PPMI, LRRK2 and BioFIND datasets, data use agreements were executed with the Alzheimer's Data Neuroimaging Initiative, the Parkinson's Progression Markers Initiative, the LRRK2 dataset, led by the Michael J. Fox foundation, and the BIOFIND dataset. Ten datasets that were curated and loaded by the University of Luxembourg originated in GEO, and did not require any data use agreements. This Datathon marked the first time that these datasets have been made available in a single analytic platform. Together the datasets represent over $500 million investment in data generation.

ADNI: The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal, multicenter study to develop genetic, biochemical, clinical, and imaging biomarkers for the early detection and progression tracking of Alzheimer's disease. 3142 patients are currently enrolled.

PPMI: The Parkinson's Progression Markers Initiative is a longitudinal, multimodal observational study of a large patient population. The dataset contains biological sampling, advanced imaging, clinical and behavioral assessments; i.e. the Movement Disorder Society—Unified Parkinson's Disease Rating Scales (MD-UPDRS), Montreal Cognitive Assessment (MoCA) and the University of Pennsylvania Smell Identification (UPSIT). 1334 patients are currently enrolled.

LRRK2: Michael J. Fox Foundation has established a LRRK2 Cohort Consortium to undertake an innovative approach to design and streamline drug development around the LRRK2 gene, a promising target. 2824 patients are currently enrolled.

BIOFIND: is a clinical observational study designed to discover and validate novel biomarkers for Parkinson's disease. 229 patients are currently enrolled.

10 Parkinson's Disease (PD) studies from GEO: The NCBI Gene Expression Omnibus (GEO—http://www.ncbi.nlm.nig.gov/). In attempt to exact valuable knowledge, data scientists from the Luxembourg Centre for Systems Biomedicine (http://wwwen.uni.lu/lcsb), University of Luxembourg (http://wwwen.uni.lu) manually curated 10 PD studies from GEO, which are selected based on having good amount of clinical data apart from gene expression data. These studies were curated in the context of ongoing Innovative Medicine Initiate (IMI) project and eTRIKS (http://www.etriks.org). Data from these studies were passed through following workflow: data acquisition, parsing, manual inspections, data standardization, semantic alignment and mapping, the generated structured files are ready to be used as input for the tranSMART ETL (Extraction, Transformation and Loading) operations. The structured files were loaded into tranSMART using the Pentaho Kettle ETL scripts.

The tranSMART Foundation, working with the University of Michigan and LONI, installed the platform on LONI cloud servers, and coordinated the installation of the curated datasets onto these servers with Thomson Reuters, the Michael J. Fox Foundation and the University of Luxembourg. The latest tranSMART platform, v1.2.4, was employed for these efforts. Access to the databases permitted participants to evaluate whether modifications were needed to make the data more usable.

http://dx.doi.org/10.1016/j.atg.2015.08.004

Twenty-five scientists from leading institutions in the US and Europe were selected from a pool of over seventy applicants five teams.[1] In addition to the tranSMART platform, various third-party analytic tools were employed, including MetaCore™, R interface, Spotfire, E-Workbook, MatLab, and REFS™.

3. Approaches & results

Each team applied a distinctive and innovative approach to exploration and analysis of these data. The sum of their findings validated the feasibility of their approaches and suggest possible new approaches for the research community to adopt in studying these diseases.

Team 1: Challenge: Identify novel biomarkers that predict progression of Parkinson's disease using the PPMI dataset. Team members sought to 1) identify significant SNPs, and SNP interactions, that predict progression of Parkinson's disease, and 2) evaluate the implications of findings for an improved understanding of the biology of the disease as well as for better clinical trial design..

They assessed the data to construct a data frame to model key variables in the database. This included the data frame composition, transforming outcome variables of Parkinson's Progression and evaluating the genetic data pipeline. The genetic data pipeline identified 210k SNPS, with higher criticism statistical pruning revealing 293 MoCA related SNPS and 108 UPDRS related SNPS. The final data frame contained 109 known PD SNPS, with 263 SNPS associated with MOCa scores and 668 SNPS associated with UPDRSI-III scores. Diverse rates of progression were found.

REFS™ Machine Learning was employed to do predictive analytics and build a model to predict the rate of change for MOCA decline. REFS™ model building proceeded in three steps: enumeration, optimization and simulation, to predict variables and relationships that influence outcomes.

Findings:

- Several genetic markers and one epistatic interaction were found to have stronger network consensus than age of onset.
- SNP that has an association with Multiple Sclerosis severity
- SNP that has an association with brain cancer
- A SNP associated with both Parkinson's Disease and Alzheimer disease
- Evidence for a conditional epistatic interaction.

Conclusion:

Newly identified genetic markers can explain Parkinson's disease rate of progression. The identified epistatic interaction enables improved patient groupings and enhanced risk profiles. Continued work would ideally include RNA data and multi-modal omics, both of which would deepen biological understanding of rate of progression.

Team 2: Challenge: To Build Disease Profiles for Parkinson's disease and Alzheimer disease through Biomarker Signatures Discovery. The overall aim was to identify neurodegenerative disease signatures and profiles in pre-symptomatic Parkinson's disease and Alzheimer disease. Specifically, the group sought to identify biomarkers that distinguish pre-symptomatic from normal individuals and map a unique stage of disease signature to prodromal and patient populations, both for the purpose of developing a pre-symptomatic diagnostic that is less expensive than the current imaging technologies.

Findings:

1. Analysis of ADNI data found that converters display high cerebral spinal fluid (CSF) T-Tau levels and lower precuneus thickness.
2. Potential new biomarker signatures for neurodegenerative disease, shared by Parkinson's disease and Alzheimer's disease, are:
   - CSF total Tau is significantly higher in Alzheimer's disease patients versus controls while ⟨-Synuclein levels are the same as healthy controls
   - PD patients have lower monocyte level than prodromal and healthy controls
   - PD patients and prodromals have higher levels of neutrophils than healthy controls.
3. Differentially expressed genes are markers of underlying pathology in Alzheimer's disease
   - Differential expression between normal_normal and normal_MCI (mild cognitive impairment) appears to be:
     i. Potential biomarkers
     ii. Contributors to disease process.
   - Differential expression was found in non-coding RNA, transcription factors and Wnt signaling pathways
4. Overlaying differential expression genes on the Parkinson disease map revealed two areas of mitochondrial dysfunction.

Conclusion: These findings are valuable and should be extended by further research. With added data content, further data mining is likely to identify important as yet unknown results.

Team 3: Challenge: Investigate whether there's anything in common between Gene Expression Analysis of PD and AD Blood Samples. The group explored data in the tranSMART platform to formulate their hypotheses and generate its approach to investigating possible commonalities between the molecular data, including pathways, and correlations between expression and clinical outcome. They found that PIP4K2A blood expression is lower in both PD and ADNI dementia. Applying Thomson Reuter's tools for pathway analysis pointed to TGF-Beta and SMAD3 as key players. Further, direct interactions of PD/AD pathway-enriched genes were observed.

Findings:

1. The TGFb pathway is altered in the whole blood of both PD and AD patients
2. LARGE transcript expression correlates with 20% slower increasing likelihood of developing dementia.

Conclusion: Based on these results, proposed next steps are to examine integrated gene signature of the TGBf pathway activity and to correlate TGBf pathway and LARGE expression with clinical outcome.

Team 4: Challenge: To identify a gene signature that differentiates Parkinson's disease from healthy normals and to compare findings with Alzheimer's disease. The group explored GEO studies of whole blood and post mortem brain samples in search of a shared biomarkers that could differential PD from healthy subjects. Expression data from 8 studies were retrieved with tranSMART R-client and assembled into a single matrix. The analysis sets the following parameters: expression data were presented by tranSMART as z-scores, fold changes were approximated as the difference of z-score means between cohorts, loose thresholds were used so that an adequate number of genes were considered to permit evaluating overlap interactions between several studies. The gene significance was inferred from the number of studies the gene was differentially expressed in.

It was revealed by pair wise comparison of the number of common differentially expressed genes for each study that this is a significant

[1] Team One: Jason Eshleman, IO Informatics, Boris Hayete, GNS Healthcare, Matthew Valko, GNS Healthcare, Vasco Verissimo, University of Luxembourg, Daniel Weaver, PerkinElmer, Team Two: Hiroko Dodge, University of Michigan, Ken Kubota, Michael Jay Fox Foundation, Thomas Misko, Takeda, Venkata Satagopam, University of Luxemoborg, Venus So, Roche, Jieping Ye, University of Michigan, Team Three: Milan Ganguly, IDBS, Andrew Krueger, Takeda, Adam Palermo, Sanofi, Jose Cruz, IO Informatics, Jian Zhu, University of Michigan, Team Four: Eric Aslakson, Poiema, LLC, Ivo Dinov, University of Michigan, Mark Frasier, Michael Jay Fox Foundation, Eugene Myshkin, Thomson Reuters, Alexandria Papa, Pfizer, Alex Yuzhakov, Ulon (Akson Pharmaceuticals), Team Five: Christian Ebeling, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Zhuma Feng, Biogen, Pinghua Gong, University of Michigan, Wei Gu, University of Luxembourg, Jong-min Lee, Massachusetts General Hospital/Harvard University, Janneke Schoots van der Ploeg, The Hyve.

overlap of differentially expressed genes identified from studies done on brain samples. Yet there was almost no overlap between the differentially expressed genes from blood samples with brain samples, indicative of the role of blood brain barrier and suggesting that researchers should be careful when using blood biomarkers as predictors for brain diseases.

The top differentially expressed genes found common to all 8 studies under investigation were: RBX1 (RING-box protein 1), RIOK3 (serine/threonine-protein kinase RIO3), UCHL1 (ubiquitin carboxyl-terminal esterase L1) and UGP2 (UDP-glucose pyrophosphorylase 2). Of them, only UCHL1 was previously reported in the literature to be associated with PD, the other three genes can be considered novel PD biomarkers never before associated with it. 534 genes were differentially expressed in at least six studies.

Pathway enrichment analysis with MetaCore™ software indicated that this 534 gene signature is enriched with pathways specific for neurodegenerative diseases and PD. The top pathways involved were oxidative phosphorylation, neurofilament cytoskeleton remodeling dynein–dynactin motor complex in an axonal transport. Among the top pathways also enriched with over connected hubs, that can be considered upstream regulators of that gene signature, as suggested by MetaCore™ Key Pathway Advisor, was the "LRRK2 in neurons in Parkinson's disease". Interestingly, LRRK2 is the most commonly known genetic cause of PD and a number of genes from that 534 gene signature are involved in its signaling.

Findings:

1. The 534 gene signature is enriched with genes associated with neurodegenerative diseases in general and LRRK2 signaling in particular
2. Many genes not yet identified as associated with PD are in fact potential novel biomarkers.
3. These results can be applied to subtype Alzheimer's patients. A subset of the 534 gene signature when applied to blood sample data from ADNI was able to differentiate between two cohorts, which have yet to be identified.

The team formulated a second project to categorize PPMI subjects through the use of clinical data in the PPMI tranSMART dataset, including additional brain shape information derived by the neuroimaging biomarker dataset from the PPMI cohort.

Conclusion: This second project is currently ongoing and is evaluating dystonia and dyskinesia as a measure of disease severity compared to brain shape information.

They intend to perform a latent class analysis followed by a statistical validation of the clinically derived subgroups against SNPs.

Team 5: The team compared differential gene expression for PD versus controls with microarray data and checked the effects of those genes in AD. The goal was to identify biomarkers that are associated with triggering the same etiologies in hopes of identifying causes of rapid progression. The team evaluated the top 100 differentially expressed markers in each study and checked the markers that are shared by at least two studies. They looked extensively at the genes that are also differentially expressed in AD (ADNI dataset) as well as one gene that shows differences in both brain tissue and blood.

Findings:

- 16 biomarkers were identified from at least 2–3 different studies
- 1 biomarker was identified from 2 studies evaluating both blood sample and brain tissue samples, indicating a potential diagnostic marker
- Both the PD and AD data sets showed biomarkers associated with the same triggering etiology.

Conclusion: Further evaluation of shared biomarkers associated with triggering etiology is a promising research strategy.

4. General observations on the Datathon

4.1. Lessons learned

Several teams used the tranSMART platform for data browsing and patient cohort selection, and then downloaded and exported data files to other analytical tools outside the platform. Others used the built-in analytics for most or all of their analysis. Some noted specific data collection needs that will improve data analysis capabilities, such as the noted absence of CSF in the PD prodromal cohort.

5. Conclusion

Bringing data scientists, neuroscientists and biostatisticians together to leverage a large, integrated cross-neurodegenerative disease dataset in tranSMART validated the utility for the platform, and the value of these data when integrated. The Datathon produced an innovative approach using machine learning, new biomarker findings, and scientifically thorough pathway analyses. It also successfully demonstrated that data sharing and the Datathon approach in particular can expedite discovery, as well as offer new fruitful ways to explore datasets. These new approaches are available for the research community to adopt.

Further, the event demonstrated that the tranSMART Platform can support many different large and distributed datasets as well as meet varied end-user needs simultaneously. The platform's ability to organize these datasets in useful fashions enabled participants to generate new findings and approaches for future research. In doing so, the Platform proved to be a powerful tool in expediting research in a cost efficient way. The value of bringing together experts from different disciplines and organizations to work in newly formed teams on integrated datasets deployed in the tranSMART platform was evident. The Datathon also provided excellent technical and scientific feedback on how to improve the tranSMART platform for use by the research community in real scientific projects. Finally, all 5 teams stated that they intended to continue the work started at the Datathon, and to continue to work with their Datathon team members.

Thank you to our sponsors, IDBS, PerkinElmer and Thomson Reuters for helping make the Datathon happen.