Rancho bioscience

Summer 13

Accelerated Cure Project: a Case Study

for Rancho BioSciences

What follows is a description of a large data curation project undertaken by Rancho BioSciences for the Accelerated Cure Project for Multiple Sclerosis (ACP). ACP is a non-profit focused on the discovery of a cure for multiple sclerosis, and related demyelinating disorders, by providing the research community with tools to enable scientists to remove barriers that otherwise would slow the conduct of groundbreaking research.

Stephen Wicks, Ph.D., Rancho BioSciences

The Accelerated Cure Project for Multiple Sclerosis (ACP) maintains a biorepository of blood samples from approximately 3200 patients with demyelinating diseases (primarily Multiple Sclerosis, but also including transverse myelitis (TM), neuromyelitis optica (NMO), acute disseminated encephalomyelitis (ADEM), optic neuritis (ON)) and associated control samples. The source individuals for these samples have been characterized through the collection of an extensive case-report form (CRF). This makes the dataset particularly valuable to the research community, who solicit access to the blood samples and associated CRF data in order to test particular hypotheses that arise in the course of research programs focused on finding causes of and cures for demyelinating diseases. ACP provides this material with the proviso that any data collected from the biorepository by the research entity be returned to ACP in a timely fashion so that it, in turn, might be made available for analysis and consideration by other researchers. Repository samples have been distributed to more than 70 entities, and data from more than 35 completed studies has been returned to ACP.

ACP is also a founding member of the Orion Bionetwork. This organization is an alliance of members with access to tools and data that can contribute to computational models of neurological disease progression. As part of its relationship with Orion, ACP has committed to making the ACP repository data available for analysis. This data would include the CRF data, as well as data derived from various scientific studies. Currently, the CRF data resides in an SQL database, and the research data resides in various forms depending on the nature of the individual research study. Research studies vary widely in scope and focus and include gene expression datasets (CEL files), GWAS studies (plink output files), qRT-PCR data (C*t* files), proteomic studies (csv or tab-delimited text files), and a variety of focued biomarker studies (spreadsheets or text files). Some of this data is partially processed and some of the data is completely raw.

Orion has selected tranSMART as a data repository to house MS-related data from several organizations including ACP for analysis. tranSMART is a data warehousing and analysis application first developed at Johnson & Johnson with Recombinant Data Corporation in 2009. It was open-sourced in late 2012, and has accrued significant developer and user interest as a potential solution to the problem of data availability and compatibility in the age of high-throughput biological analysis. tranSMART can house high dimensional biological datasets (gene expression, sequence, SNP, protein etc), clinical data (low dimensional biomarker data), and clinical/demographic data. Additionally, it can house other arbitrary data forms (images, PDFs etc.) and provides sophisticated selection and search abilities across data types. For high-dimensional datasets, workflows that exploit analysis modules, written in R, such as hierarchical clustering, survival analysis, heatmap visualization, etc. have been developed and are incorporated into the tranSMART interface.

The true strength of this warehouse is evidenced by its cohort selection abilities. Since it houses all data forms associated with a given study, cohort selection based on clinical variables is trivial. It would be relatively simple to, for example, compare the gene expression profiles of those individuals from the ACP study who have relapsing-remitting MS, have left handed paternal grandfathers, never smoked, but lived in close proximity to a toxic substance emitting facility during the 90s, to the gene expression profiles of similar individuals whose paternal grandfathers were right-handed.

Deliverables:

ACP partnered with Rancho BioSciences through the Orion BioNetwork to curate the CRF data as of May 20, 2013, and to provide a series of files representing that data, that were suitable for loading into a tranSMART database. In addition, Rancho committed to the curation and loading of a set of 4 research studies (minimum), and again providing that data as tranSMART-ready files. Rancho committed to provide a full-time on-site curator from late May through the end of October 2013 to accomplish this task.

Data Description:

At the outset of the curation project, the CRF data resided in an SQL database, and the research data resided in various forms depending on the nature of the individual research study. These studies included gene expression datasets (CEL files), GWAS studies (plink output files), proteomic studies (csv or tab-delimited text files), and a variety of spreadsheets. Some of this data was partially processed and some of the data was completely raw. Rancho BioSciences partnered with ACP to curate all the CRF data and data from 4 studies as part of this contract.

CRF data:

Description. The CRF data is stored in an SQL database. The database consists of distinct tables, approximately 70 of which correspond to data from the CRF forms filled out by patients. These tables draw entries from various parts of the CRF. The full structure of the CRF data is outlined in Appendix 2. The 16 main divisions of the primary CRF are as follows:

1.1. Subject Entry
1.2. Inclusion/Exclusion
1.3. Laboratory Assessment
1.4. Interview Information
1.5. Study Completion
1.6. Demographic Information
1.7. Ethnic Background

1.8. Family History
1.9. Demyelinating Disease
1.10. Medical History
1.11. Reproductive Health
1.12. Environmental Exposures
1.13. Medications and Nutritional Supplements
1.14. Conclusion
1.15. Study Ascertainment
1.16. Investigator Signature

Some subjects (n=469) were seen at two distinct clinical visits, and those patients have a second CRF form associated with the second visit. The following information was collected from those patients that received a second clinical visit:

2.1. Subject Entry
2.2. Inclusion/Exclusion
2.3. Laboratory Assessment
2.4. Interview Information
2.5. Study Completion
2.6. Demographic Information
2.7. Family History
2.8. Demyelinating Disease
2.9. Medical History
2.10. Reproductive Health
2.11. Environmental Exposures
2.12. Medications/Supplements
2.13. Stress/Anxiety
2.14. Conclusion
2.15. Study Ascertainment
2.16. NMO IgG
2.17. Neuro-Opthalmologic Tests
2.18. Investigator Signature

Each of these categories has one or more subcategories. For example, the Family history section can be expanded as follows:

1.8.1. Parents and Grandparents (6 rows)
1.8.2. Siblings (22 rows)
1.8.3. Half Siblings (4 rows)
1.8.4. Multiple Birth
1.8.5. Additional Blood Relatives (8 rows)
1.8.6. Partners the participant had children with (23 rows)
1.8.7. Children (9 rows)

and for each of these categories a number of questions were posed, roughly corresponding to the parenthetical number beside each category. Since these questions are asked of each instance of the category (for example, a patient with 16 siblings would have to provide answers to the same 22 questions for each sibling), these tables can be large. The un-curated visit 1 family history report has 576 columns, by approximately 3200 rows (one per patient), or approximately 1.8 million cells of data.

Data preprocessing: In order to extract the CRF data from the SQL database, a number of steps were required. The database was provided as a single SQL dump file at the start of the project. First, the database was converted to an SQLite form in a flask environment, using python scripts written by an exchange intern who had previously worked with ACP (R. Koenig). Target fields were extracted from the SQLite database with a selection interface that allowed individual field selection and reporting. This tool (written in python) exported a csv-formatted file as follows,

'AC000632','No',?,'Yes, specify below','Coal power plant',…
'AC000487','No',?,'No',?,…

Note that according to file specifications, csv files should protect all instances of certain characters (for example, an internal comma, or internal single quotes) either with double quote containment, or an explicit escape character. The exported SQL data was inconsistent in this regard, and consequently when it was imported into a spreadsheet for curation, it would parse each line to a different number of fields. To rectify this problem, a pre-processing pipeline was established. Eleven serial grep substitutions were performed on each exported report to escape commas that were creating this problem. A further three serial grep replacements were used to remove internal un-escaped commas and single quote characters.

Next, several date formats were used in different pats of the CRF form. For example, the following were all present in exported CSV files:

1999
--/1999
--/--/1999
Dec 1999
10/--/1999
01/12/1999

Several of these formats were not recognized as dates by downstream curation tools, or worse, were converted to nonsensical dates by tool import functions (Excel was particularly problematic). The following grep substitution (or variants thereof, with appropriate post processing) converted all of these to a standard format that would be unambiguously interpreted by any version of excel (since excel 1997):

Substitute: '([-0-9][-0-9])/([-0-9][-0-9])/([-\d]+)'
With: '\3-\1-\2'

The resulting dates conformed to the ISO standard, and reported dates in the form "yyyy-mm-dd". When no month or day information was reported by patients, but was required by the CRF question, we used Jun-01 as a mock date. For example, for a field in which people were asked when they last took a MS disease-modifying therapy, and the field expected a 'dd/mm/yyyy' format, and responded with, say,

"'86", this data would be transformed to 1986-06-01. This was necessary since subsequent calculations needed during curation—for example, "calculate the elapsed number of months since the drug was taken prior to a blood draw"—required a reported month. If a field only expect a year value, no mock day or month were used.

Finally, not every individual in the ACP database has complete record information. In particular, some individuals have nearly incomplete information. Most or even all of the tables for those individuals are empty. A set of these were identified by the lack of any demographic information (age, gender, etc.) early in the project and a file 'exclude.txt' was created, containing the BarcodeID of these subjects, to filter these rows from the exported data, all tables were subjected to the following grep command:

grep –Fvf excluded.txt *inputname.csv* > *outputname.txt*

Files treated with these steps consistently imported into a spreadsheet as .csv files and parsed flawlessly.

Research Study data:

Study data comes back to ACP regularly. In June 2013, ACP had data from 26 studies. By August 2013, there was data from 36 studies available. Some of these datasets would be relatively trivial to load as clinical data points (for example, ELISA results from 3-4 biomarkers in a subset of blood samples from the ACP Repository). Data from other research studies was much more elaborate. The main challenge from the perspective of curating this data is the variability of study type. While the ETL loading procedures for gene expression data are well documented, the documentation for loading GWAS, or raw sequence data is less well documented. Such documentation has been slow to appear in the public domain.

Sample data: The data formats of several studies chosen for curation will be used to illustrate the variability in data formats in this project. The research studies vary widely in scope. Some will be prepared as high dimensional data, and other studies will be loaded as clinical endpoints.

1) Biogen gene expression study: Biogen IDEC returned data for 230 samples selected from the ACP repository. Sample selection was intended to address the following objective: To investigate the whole blood gene-expression profiles of Secondary Progressive MS patients (SPMS), to better understand the molecular features that distinguish SPMS vs. the relapsing-remitting form of the disease (RRMS). RNA was extracted from PAXgene tubes, purified and hybridized to the Affymetrix ht-hgu133plusPM gene-chip arrays.

   The data returned to ACP was heavily processed. In particular, the data was treated for background variation, normalized, and extracted into a

spreadsheet: no CEL files were made available in this case. The file has 231 columns (230 samples and one ID column) and 54716 rows (one header row, and 54715 probe IDs). Appendix 1 contains a sample of some results derived from this data processed in tranSMART.

2) Vanderbilt University RT-PCR data. Samples from 151 patients in the ACP repository were withdrawn, and RNA was extracted and purified. 48 distinct RT-PCR reactions was run on these samples. The data was returned to ACP in processed and normalized form as an excel spreadsheet. The study design was focused on finding gene expression signatures that could be leveraged to design a diagnostic blood test for demyelinating diseases.

3) Somalogic aptamer binding data. Somalogic has developed a set of aptamers (they call somamers) that are modified DNA molecules that bind specific proteins with high affinity and specificity. They have developed a platform based on this technology. Somalogic withdrew samples from 114 patients, extracted proteins and hybridized them to this platform. The returned data for 113 subjects that met QC for relative protein concentrations for 1046 proteins. We designed a platform file for this data that describes what proteins were interrogated with this chip.

4) Protagen data. The German company protagen has a large in-house library of recombinant clones of human proteins. They utilize a technology that scores blood samples for the presence of auto-antigens to human proteins. In auto immune diseases these antibody titres can become elevated. Protagen requested blood samples from 40 individuals, and returned a spreadsheet containing auto-antigen measures for 384 proteins.

5) 23andme GWAS data. The personal genotyping company 23andme requested samples for 115 individuals, and returned genome wide SNP typing data for 114 individuals. This data provides genotype calls for 934671 locations. The data was returned in a modified form of a transformed plink file for each individual. Plink was used to adjust the data format to form that could be loaded into transmart.

6) Glycominds data. Based upon samples drawn from ACP and other sources, the medical diagnostics company, Glycominds, has developed a blood based diagnostic assay (**Glycominds**' gMS™ Dx test for **MS)** based on alterations in protein glycosylation that are seen in MS. The assay uses anti-bodies that detect differences in glycosylation states of proteins. Glycominds used 1206 ACP samples, and returned data for 10 distinct measures that were used to develop this MS assay. The data was returned as an excel spreadsheet.

7) Diogenix data. Diogenix, another company with a goal of an MS diagnostic assay based on gene expression profiles, selected 26 samples from the ACP database for a feasibility study. These samples were run through the

Affymetrix U133 platform and, in contrast to the Biogen gene expression study, results were returned to ACP for 26 individuals as raw CEL files. The curator processed files in the software package "R" to prepare them for loading into tranSMART.

Challenges and Solutions:

While the project is not yet complete, there are several examples that illustrate how data curation was central to assisting ACP in meeting its commitment to Orion BioNetworks. In some cases, curation merely served to simplify a potential analysis; in other cases, curation enabled entirely novel ways to explore the ACP data, by coding free text data, or awkwardly encoded data into a form that anticipated likely questions. A few of these are outlined here:

1) Uniform availability data: Disparate data types collected by ACP for the Repository project have been made available for study in a single database. Previously, clinical workups, demographic data, and study ascertainment data was keep in an SQL database, whereas the experimental data extracted from patient samples were kept in a variety of *ad hoc* forms. Since the contents of the full CRF data is housed in the same warehouse as the experimental data, the experimental data can now be evaluated within the context of complex cohort selections based upon factors that *even the scientists that requested the data and performed the experiments* could not have considered. Furthermore, the complete CRF dataset has tremendous potential value even without the associated experimental datasets.

2) Recoding fields of complex data types: ACP captured very deep data on a wide variety of measures, including drug/medication histories and environmental exposure histories for each patient in the repository. However, the data fields were not set up in a way that was conducive to cohort formation and data mining. We recoded the data into tables that allowed access to this information in a convenient format. For example, the subjects provided information on previous prescribed and OTC medications (non-MS specific) with the following fields: Drug, Route of delivery, Dose (including units), Frequency, Start Date, Stop Date, Reasons Stopped. Some patients filled out these fields for more than 80 distinct drugs. The data was made available in the ACP SQL database for that set of fields for the "first drug", "second drug", 'third drug"…etc.. However, there was no consistent interpretation across patients of the order of drug reporting: some patients interpreted this roughly chronologically, some used roughly alphabetical reporting, some used medical significance of the drug, while most were apparently reported in the order in which they were recalled. Any given drug could have been in any given rank order of reporting, so if one wanted to look at women taking, say hormonal birth control medication, and compare

them with women not taking oral hormones, it would have been nearly impossible to do so. We recoded the data such that it would be possible to analyze the data by drug, rather than by the ordinal value.

3) Assignment of free-text values to controlled vocabularies: It was often necessary to recode free text data responses in the ACP data. Some examples illustrate this point, but the following are not a complete list of cases in which data was recoded. It is important to note that in all three examples given, the reported data was completely un-minable prior to the recoding.

   A) Patients were asked to report on ownership of animals. “Dog”, “god”, “dogs”, “terrier”, “DOG”, “puppies”, “four dogs and a cat”, and innumerable other responses were recoded to “Dog” in the database (the latter also being assigned to “Cat”.) A controlled vocabulary was designed to describe the range of pets reported on in this dataset. 551 unique reported values were reduced to 31 animal classifications. After recoding to controlled vocabularies, patient cohorts could be formed based on not only whether a patient was ever exposed to a pet, but also based on the kind of pet owned.

Alligator	Crayfish	Possum
Amphibian	Dog	Rabbit
Assorted	Farm Animal	Raccoon
Badger	Ferret	Reptile
Bird	Fish	Rodent
Butterflies	Harbor Seal	Skunk
Cat	Hedgehog	Snake
Chameleon	Hermit Crab	Spider
Chimpanzee	Lizard	Sting Ray
Chinchilla	Monkey	Turtle
		Worm

Table 1. Animal/Pet coding for the ACP project: Approximately 550 unique animal or animals were reported by patients, and were assigned to 31 categories by the curator.

   B) Participants were asked to report experience with athletics. Nearly every patient reported participating in one or more sports in their lifetimes. As with the case of drug reporting described above, the order of reporting of sport participation was not particularly systematic or informative, but that is

how the data was encoded. Furthermore, responses such as “Cardio”, “workout”, “GYM membership”, “weights and cardio”, and the numerous variations could all be considered equivalent. All sports were recoded to a curator-created library of 29 sports.

Aerobics Animal-assisted Sport Band Basketball Boat-assisted Sport Bowling Cycling Dance Extreme Sport Field Hockey	Football Golf Gymnastics Machine-assisted Sport Martial Art Other Sport Racquet Sport Running Sedentary/Precision Sport	Skating Sport Skiing Soccer Softball/Baseball Track and Field Volleyball Walking Water Sport Weight/Cardio Yoga

Table 2. Sport coding for the ACP project: Approximately 800 unique “sports” were reported by patients, and were assigned to 29 categories by the curator.

C) Drug dosage information: Patients were asked to provide dosages of disease-modifying drugs that they had been prescribed. This information was provided in a free text field by the patient. First, it was immediately clear that patients are not the correct way to obtain this information. Patients often reported clearly incorrect information, with dosages off by factors of a thousand or more from recommended dosages (usually associated with obvious confusion over unit measurements), or in units that were inappropriate given the route of delivery of the medication in question. Furthermore, drug frequency was also reported separately, and was also reported as free text. No guidelines for reporting drug dosage or regime were provided and consequently people reported these variables in a very idiosyncratic fashion. For example, there were 43 distinct ways for a patient to indicate “Once a week”. Drug dosage, units, and frequency were combined and recoded by the curator to a single field with three values (Dosage: High, Standard, Low) for import into tranSMART. Start and Stop dates associated with these drugs were recoded to a values of “Duration of Dosing” and “Last Taken (Months Prior to Blood Draw)”.

4) Data acquisition procedures: ACP has ceased enrolling subjects in the Biorepository, and most clinical sites have been shut down. However, ACP is

gearing up a second, more ambitious protocol, centered on longitudinal sample and data acquisition from MS patients. In preparation for this protocol, ACP has adopted a series of recommendations from Rancho, based on its curation of the Biorepository dataset. For example, we recommend that data regarding medications usage (when prescribed) be drawn from health records and medical health professionals, rather than relying on patient-reported data as well as further specific recommendations on how drug dosage data be acquired. Another example arose out of attempts to determine potential toxin exposure by asking respondents to list, for each address in their history, their proximity to a toxic-substance emitting facility. "Proximity" was defined as "within 1 mile", and while the patients were asked to specify the nature of the "Nearby substance emitting facilities (within 1 mile)", the free text responses were generally not helpful in defining any exposure risk. Researchers interested in developing risk models for the impact of chronic toxin exposure on MS progression would extract information from geographic information systems (GIS) databases such as the EPA-TRI (Environmental Protection Agency-Toxics Release Inventory) Program database on substance release information from known sources. Reasonably accurate models could be generated if street level residential information was available (Note: the ACP data does not encode residential information beyond *city, state*, to help ensure that patient de-identification is maintained.)

5) Date coding: While, on occasion, there are reasons to be concerned about the time and date of an event (for example, in the case of the ACP data, given that there is a known link between Vitamin D metabolism and MS progression and susceptibility, and since Vitamin D levels are determined, in part, by sunlight exposure, one might be concerned with seasonality of significant events such as exacerbations), in general, the salient information contained in a date stamp from an analysis point of view is simply the elapsed time to or from a primary event. In the case of the ACP data, the primary event of interest was the date of the specimen acquisition. As such, other events, such as dates of significant trauma, hospitalizations, exacerbation/relapses, drug schedules, etc. were coded in terms of elapsed Years/Months/Days (as appropriate) to the date of the relevant blood draw.

Next Steps

Once the ACP CRF and experimental datasets have been curated, they will be loaded into an internal Rancho BioSciences tranSMART instance in order to perform quality control checks on the file deliverables. Currently, with more than 90% of the CRF data and four research studies curated, the ACP repository consists of 44 flat tab-delimited files. Loading these into an active tranSMART instance will allow us to ensure that the files delivered to ACP at project completion will be tranSMART

compliant. We will derive a variety of cohorts based upon clinical parameters of potential interest, and use those cohorts to segment and analyze a single high dimensional dataset (gene expression), with advanced R-based workflows in tranSMART.

Orion will provide a tranSMART instance to host the ACP data. Rancho BioSciences has been contracted to load the ACP files into this Orion-hosted tranSMART for use by the wider Orion computational community. It is anticipated that the depth and breadth of the ACP CRF data will prove useful in defining predictors of MS onset or progression. In addition, ACP will engage a biostatistician to analyze the curated CRF data outside of the tranSMART context, with the hope of identifying previously unrecognized relationships in the data.

Conclusions

The ACP Repository data is a rich potential source of insights into the etiology and progression of demyelinating diseases. Approximately twelve million dollars was invested over a period of seven years at clinical sites across the US to gather data from approximately 3200 patients either with demyelinating diseases or unaffected controls. Each subject provided blood samples (and for a subset, a two blood samples)at one clinical visit and completed between 45 and 80 pages of background and medical information, covering topics including, but not limited to, demographics, family history, medication and supplement history, a comprehensive medical history, environmental exposures, hobbies, study ascertainment, and laboratory assessment. After curation, responses to the CRF by 3200 respondents totals approximately 26 million data points. Each participant in the biorepository thus represents an investment of between $3500 and $4000. Samples from many of these patients have been leveraged multiple times by more than 70 research organizations interested in developing MS diagnostics, understanding etiology, and identifying markers of disease activity. These studies ranged from the determination of a small number of measures from a limited subset of patients, to large scale OMICS projects, such as GWAS, gene expression analysis, proteomic analysis, and whole-exome sequencing. The data from these studies is slowly being returned to ACP for integration with the clinical workup.

The scope of the data is undoubtedly impressive. Before statistical analysis of the data could be implemented, however, several key steps had to taken. Results from research studies returned to ACP were not subject to clear guidelines that outlined how data should be treated or documented when returned to ACP. Consequently significant curation of the research sets was required. Additionally, the CRF data required significant treatment prior to being in a state that would be conducive to statistical analysis: fields required harmonization and consolidation, controlled vocabularies needed to be applied, and free text responses required expert curation. The relatively affordable step of curation was deferred until CRF data collection was

nearly complete. Then, over the course of five months, the dataset was transformed into a rich, pliable source of potential insights into MS and related demyelinating diseases.

The data was represented in files for loading into a tranSMART instance. The unique power of tranSMART, to enable exquisite circumscription of cohort selection by clinical measures is particularly important in the case of a disease with a complex and unknown etiology like Multiple Sclerosis. Consider that until recently, Neuromyelitis Optica was thought to be a type of Multiple Sclerosis. It seems likely that other subsets of the population of people currently diagnosed with MS, in fact have what will come to be recognized as one of a number of distinct yet related conditions. That is, MS likely represents a cluster of related diseases, each with a distinct etiology, but sharing similar clinical outcomes. The ability to parse out subsets of a population using pre-clinical measures and clinical endpoints, and analyze the high-dimensional OMICS data from those sub-populations will likely be critical to understanding the nature of the etiologies of these related diseases.

Appendix 1: Sample Experimental Data

Biogen's gene expression study:

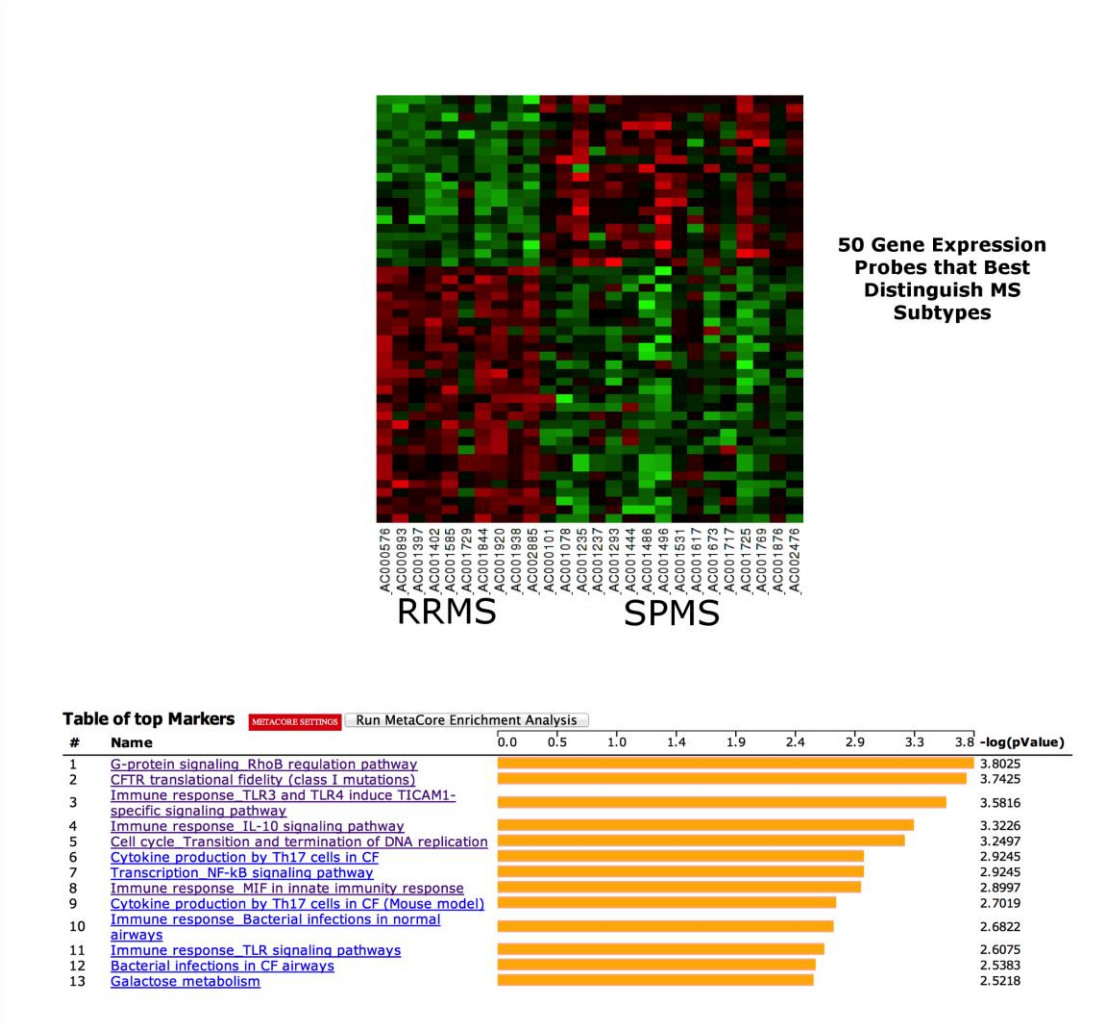

Table of top Markers METACORE SETTINGS Run MetaCore Enrichment Analysis

#	Name	-log(pValue)
1	G-protein signaling_RhoB regulation pathway	3.8025
2	CFTR translational fidelity (class I mutations)	3.7425
3	Immune response_TLR3 and TLR4 induce TICAM1-specific signaling pathway	3.5816
4	Immune response_IL-10 signaling pathway	3.3226
5	Cell cycle_Transition and termination of DNA replication	3.2497
6	Cytokine production by Th17 cells in CF	2.9245
7	Transcription_NF-kB signaling pathway	2.9245
8	Immune response_MIF in innate immunity response	2.8997
9	Cytokine production by Th17 cells in CF (Mouse model)	2.7019
10	Immune response_Bacterial infections in normal airways	2.6822
11	Immune response_TLR signaling pathways	2.6075
12	Bacterial infections in CF airways	2.5383
13	Galactose metabolism	2.5218

We loaded Biogen's gene expression data set into tranSMART and derived a cohort from disease status. We ran "Marker Selection" on 26 samples from the gene expression dataset to look for fifty gene-expression probes that best distinguished RRMS (Relapsing-Remitting; 10 samples) from SPMS (Secondary-Progressive; 16 samples). These probes were visualized in a heatmap, and the set of protein pathways associated with those probes were identified via Meta-Core enrichment in tranSMART. Many of the identified pathways are immunity-related, consistent with the nature of the Multiple Sclerosis.

Appendix 2: ACP CRF data structure.

Note that pages 1-53 refer to data collected on clinical visit one for all subjects, while pages 201 on refer to data collected only for a subset of the subjects on a second clinical visit.

InclusionExclusion
- Page 1
  - Survey
  - Case Inclusion
  - Case Exclusion
  - Control Inclusion
  - Control Exclusion

Laboratory Assessment
- Page 2
  - Laboratory Assessment

Study Completion
- Page 3
  - Interview Information
  - Study Completion

Demography
- Page 4
  - Demography
  - Social

Ethnic Background
- Page 5
  - Ethnicity
  - Location

Family History
- Page 6
  - Multiple Sclerosis Parent
- Page 7
  - Multiple Sclerosis Sibling
- Page 8
  - Multiple Sclerosis Half Siblings
  - Multiple Sclerosis Sibling
  - Multiple Sclerosis Other Relatives
- Page 9
  - Multiple Sclerosis Partner
- Page 10
  - Multiple Sclerosis Children

Demyelinating Disease
- Page 11
  - Demyelinating Disease MS/CIS
- Page 11
  - Demyelinating Disease TM

- Page 252
  - Study Ascertainment Evoked
  - Study Ascertainment Biopsy
  - Study Ascertainment Standard2
- Pages 253, 254, 255
  - Study Ascertainment Differential
  - Asia Impairment Scale
  - IgG -Neuro-opthalmologic
- Page 256
  - Nmo Igg
  - Neuro-Opthalmologic Tests

Investigator Signature

- Page 257
  - Case Report Form Verification