Genomics Biomarker Discovery Pipeline – EDRN/JPL viewpoint

Ashish Mahabal, Luca Cinquini, Dan Crichton (PI), Thomas Fuchs, Heather Kinkaid, EDRN-JPL team, Joe Perez-Rogesrs, Marc Lenburg (BU- PI), Ania Tassinari, BU team

Supporting the Science Data Lifecycle

- *Ingestion of data*:
- *Cataloging of Structured and Unstructured Data*:
- ***Data Processing*: Scalable**
- ***Data Management*: Construction and management of metadata catalogs and data (often distributed);**
- *Data Discovery*:
- *Data Access*:
- ***Data Distribution, Computation and Analysis*: Support for analysis and services (e.g., subsetting) on the data; move towards automated data discovery**

Moving Towards a New Paradigm

Early Detection Research Network

NASA JPL Jet Propulsion Laboratory California Institute of Technology

Development of automated pipelines built on workflows

Constructing highly distributed, multi-organizational systems

Sharing of data and services which allow for the discovery, access, and transformation of data

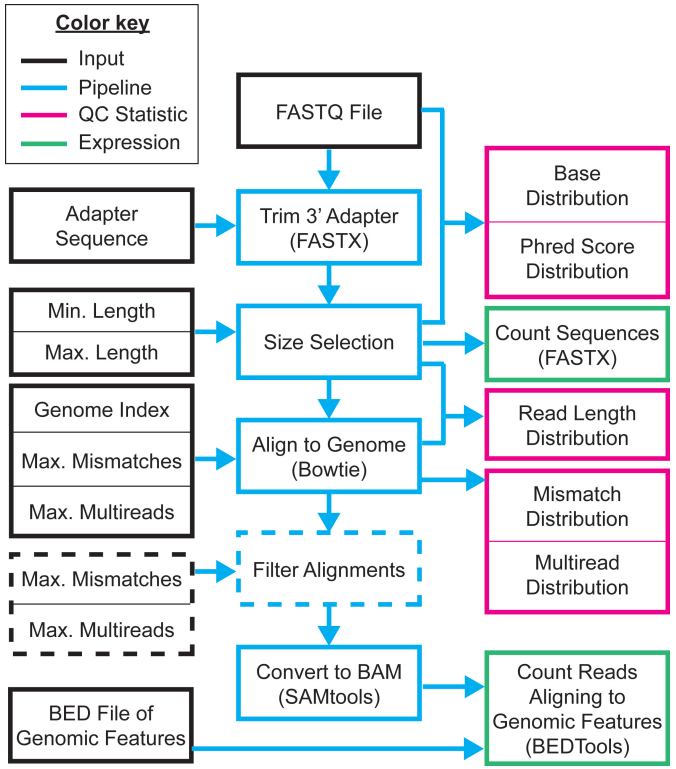
Addressing complex modeling, inter-disciplinary science and decision support needs

Changing the way in which data analysis is performed

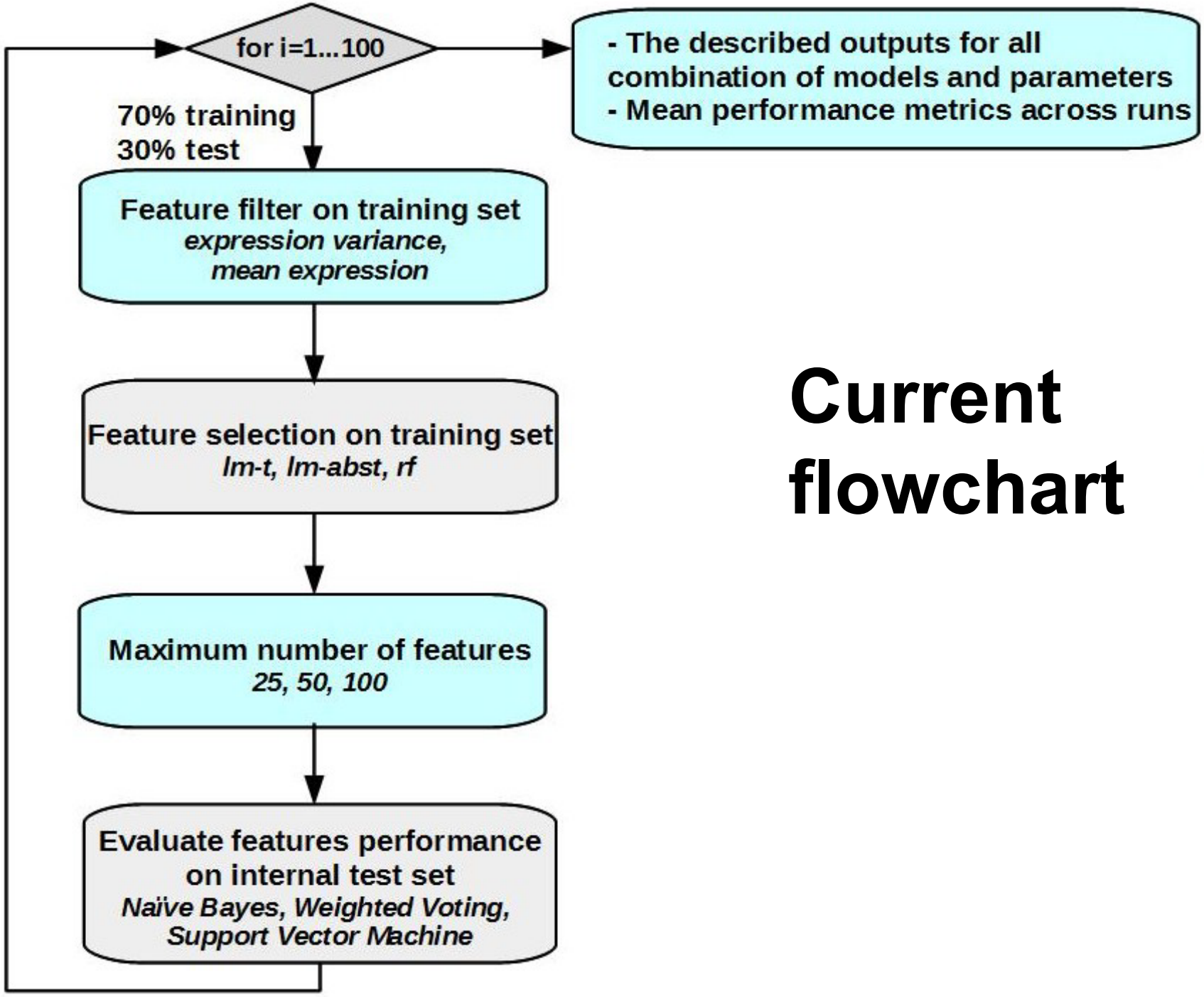
Boston University (PI - Marc Lenburg)

The group at Boston University is seeking to extend on their previous work on mRNA biomarkers detected in bronchial epithelium samples to mRNA and microRNA biomarkers that can be detected in less invasively collected nasal epithelium samples.

We start by choosing one pipeline



Small RNA (e.g., microRNA) sequencing workflow



Current flowchart

sion Laboratory
Institute of Technology

Current setup (at BU)

PBS scripts (sh) submitting R codes

5 modules

1. Discovery set split (q samples, p variables, m genes; 80/20, 70/30 etc. splits)
2. Gene filter (class independent characteristics)
3. Feature selection (prioritize features based on phenotype)
4. Biomarker size (select 25/50/100 etc. features)
5. Prediction method (naïve bayes, SVM etc.)

Ingredients of the for loop

ff=c("var","mean”)

fs=c("lm-t","lm-abst","rf”)

bs=c(25,50,100)

cl=c("wv","nb","svm")

(feature filter, feature selection, biomarker size, classification)

n(for iterations) = n(ff)*n(fs)*n(bs)*n(cl)

Early Detection Research Network

Sample partial output (summary)

JPL Jet Propulsion Laboratory California Institute of Technology

ind	FF	FS	BS	CL	TrAUC	TestAUC	TrainNA	TestNA
1	var	lm-t	25	wv	1	0.933	0	0
2	var	lm-t	25	nb	1	0.966	0	0
3	var	lm-t	25	svm	1	0.875	0	0
4	var	lm-t	50	wv	1	0.891	0	0
5	var	lm-t	50	nb	1	0.966	0	0
6	var	lm-t	50	svm	1	0.608	0	0
7	var	lm-t	100	wv	1	0.783	0	0
8	var	lm-t	100	nb	1	0.925	0	0
9	var	lm-t	100	svm	1	0.525	0	0
10	var	lm-abst	25	wv	0.992	0.933	0	0

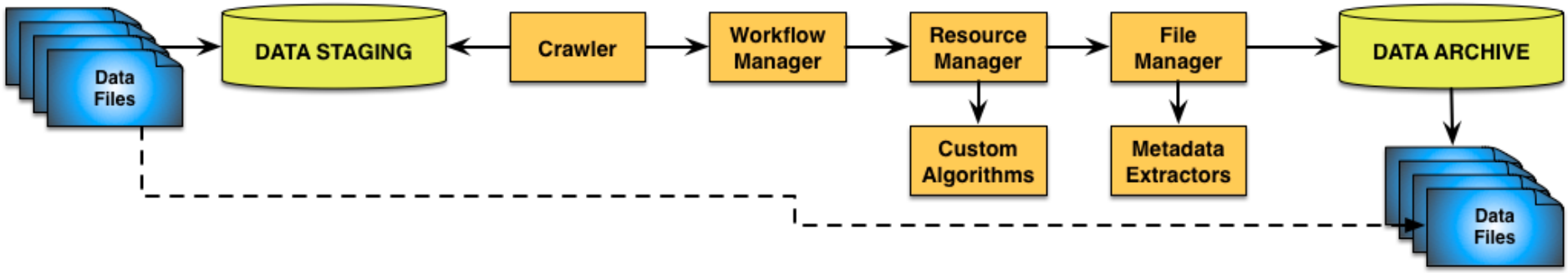
The science inferences/conclusions use such output as a starting point. Streamlining that through web-based visual analytics can be a future goal.

Using Object Oriented Data Technology (OODT)

Early Detection Research Network

NASA JPL Jet Propulsion Laboratory California Institute of Technology

- Open Source
- Modular management and processing framework
- Workflow manager
- Resource manager



Early Detection Research Network

OODT setup

NASA JPL Jet Propulsion Laboratory California Institute of Technology

- Outer loop parallelized using workflow/ resource managers
- Concurrent runs being explored (simultaneous read/write/process may be possible)
- Inner (for) loop to be parallelized within R or through a split

OODT details – qsub implementation

- First OODT task is used to submit the R scripts to qsub
- qsub distributes the R script execution across available node (one iteration on each node)
- Second OODT task waits for summary file output to be available, then publishes it to the data archive and metadata catalog

Early Detection Research Network

Pure OODT implementation

NASA JPL Jet Propulsion Laboratory California Institute of Technology

- First OODT task is used to distribute computation across available nodes (one iteration on each node), as separate sub-workflows
- Each OODT sub-workflow running on a node executes
- R script to process one iteration
- Another OODT task runs the R script to generate the summary file
- Final OODT task waits for summary file output to be available, then publishes it to the data archive and metadata catalog

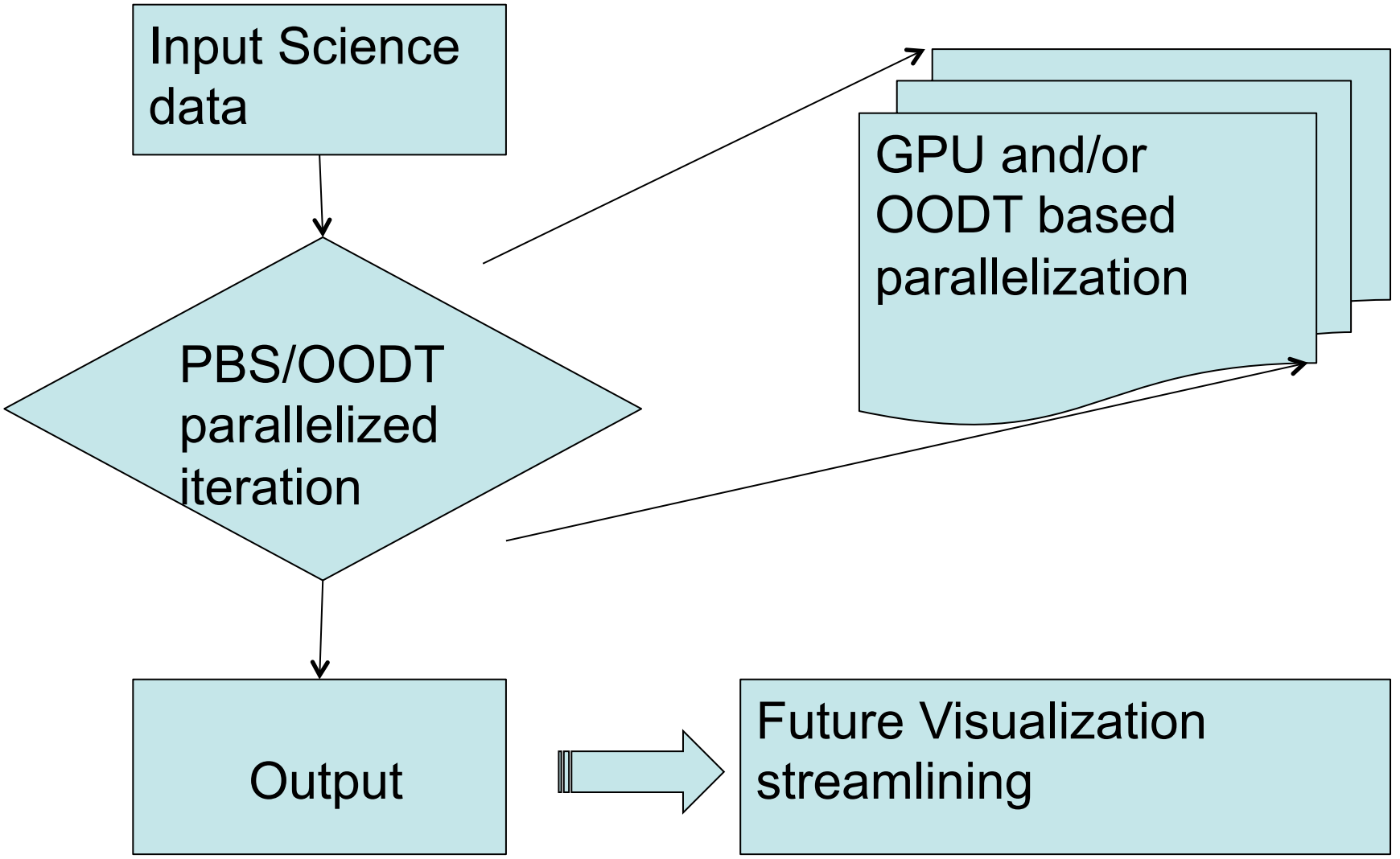
Early Detection Research Network

GPU explorations

NASA JPL Jet Propulsion Laboratory California Institute of Technology

- Biggest gain would be possible be using individual processors in a GPU for each iteration of the for loop
- This has the potential to speed up the computation by 100 to 1000 times
- This will bring down execution time from o(week) to o(hour)

Summary


Input Science data
PBS/OODT parallelized iteration
GPU and/or OODT based parallelization
Output
Future Visualization streamlining