Tiered Computation

Raymond Ng
CIO, Proof Centre
Acting Head, Department of Computer Science, UBC

Government of Canada Gouvernement du Canada
PROOF Centre of | Centre d' EXCELLENCE
In affiliation with: THE UNIVERSITY OF BRITISH COLUMBIA

BiT Biomarker Discovery Strategy

"Omics" Tools and Approaches

TRANSCRIPTOMICS

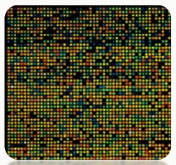
PAXgene Whole Blood

↓

RNA Extraction

↓

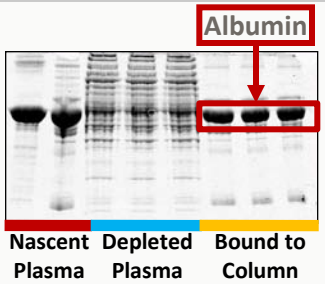
Affymetrix Microarray Analysis



Microarray Core Laboratory, Children's Hospital, LA, CA

PROTEOMICS

Plasma



Plasma Depletion

↓

ABI 4800 iTRAQ Analysis

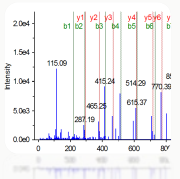
UVic-Genome BC Proteomics Platform, Victoria, BC

METABOLOMICS

Serum and Urine

↓

NMR & Mass Spec Analysis



U of Alberta Metabolomics Platform, Edmonton, AB

QA/QC – All sample collection and processing is done to SOP

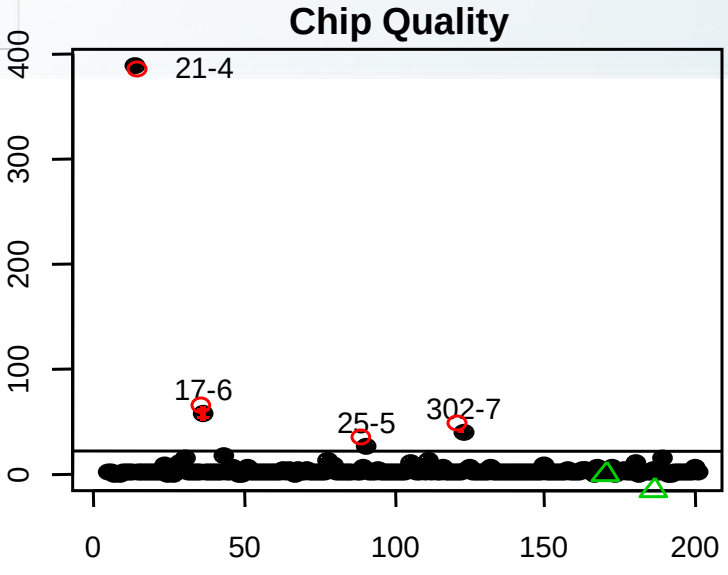
Importance of Data Cleansing and Pre-processing

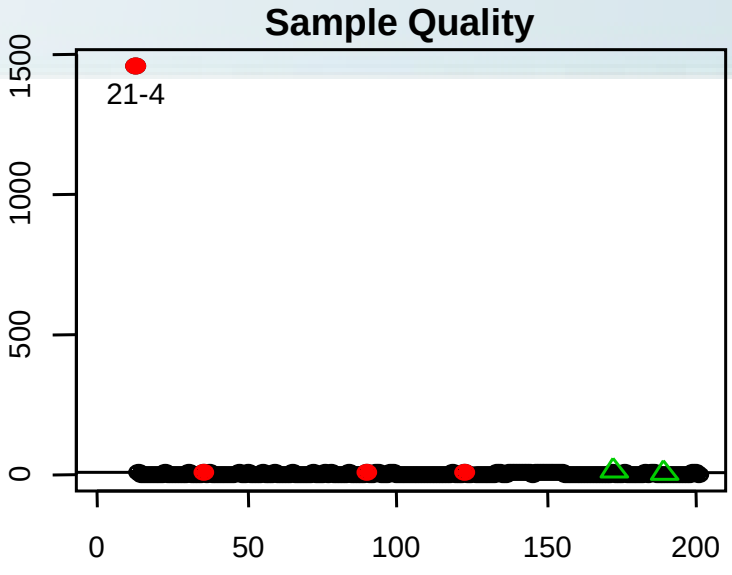
A. Clinical: *"Detecting potential labeling errors in microarrays by data perturbation,"* Bioinformatics 2006 (Malossini, Blanzieri)

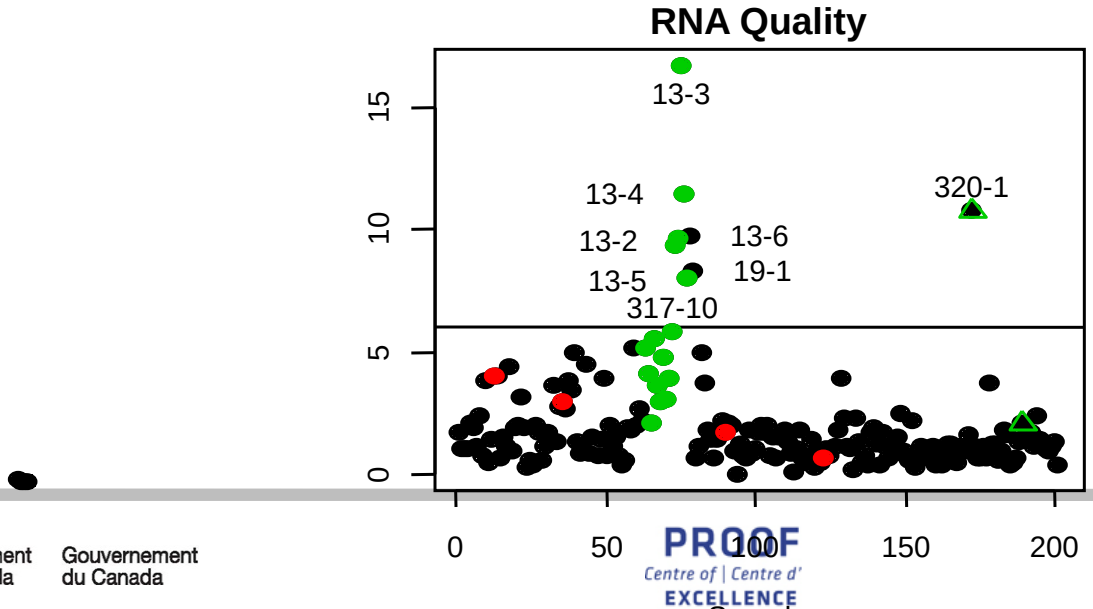
B. mRNA: *"MDQC: a new quality assessment method for microarrays based on quality control reports,"* Bioinformatics 2007 (Cohen-Freue, Hollander et al.)

C. DNA: *"Modelling Recurrent DNA Copy Number Alterations in array CGH Data,"* Bioinformatics 2006, 2007 (Shah, Murphy, Lam)

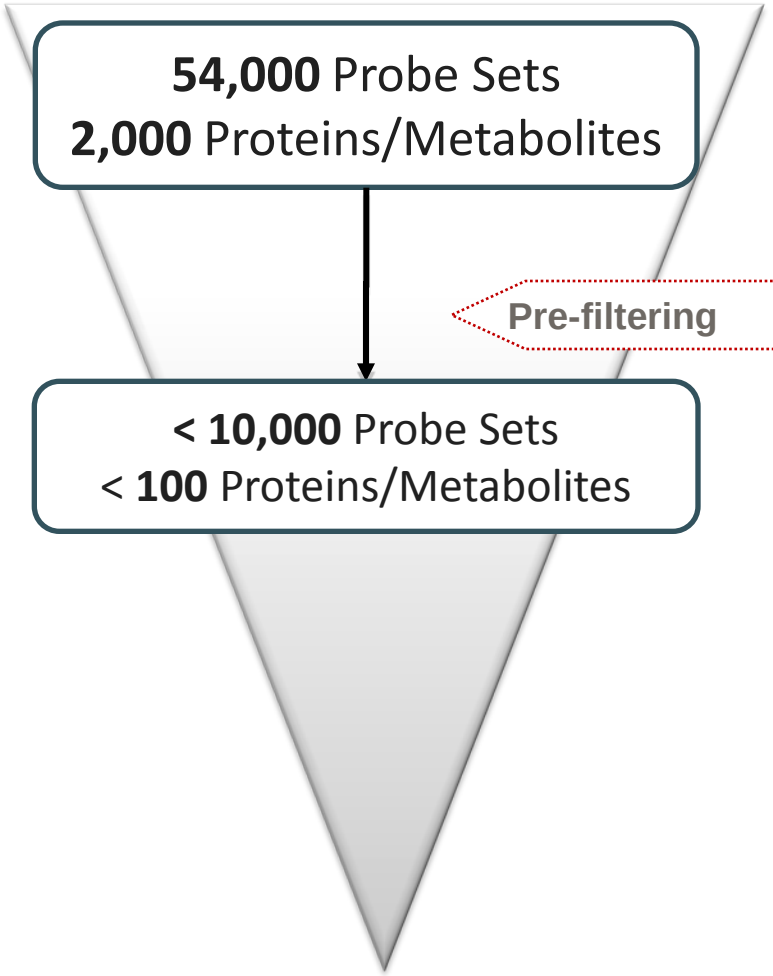
Microarray Quality Control Assessment Tool





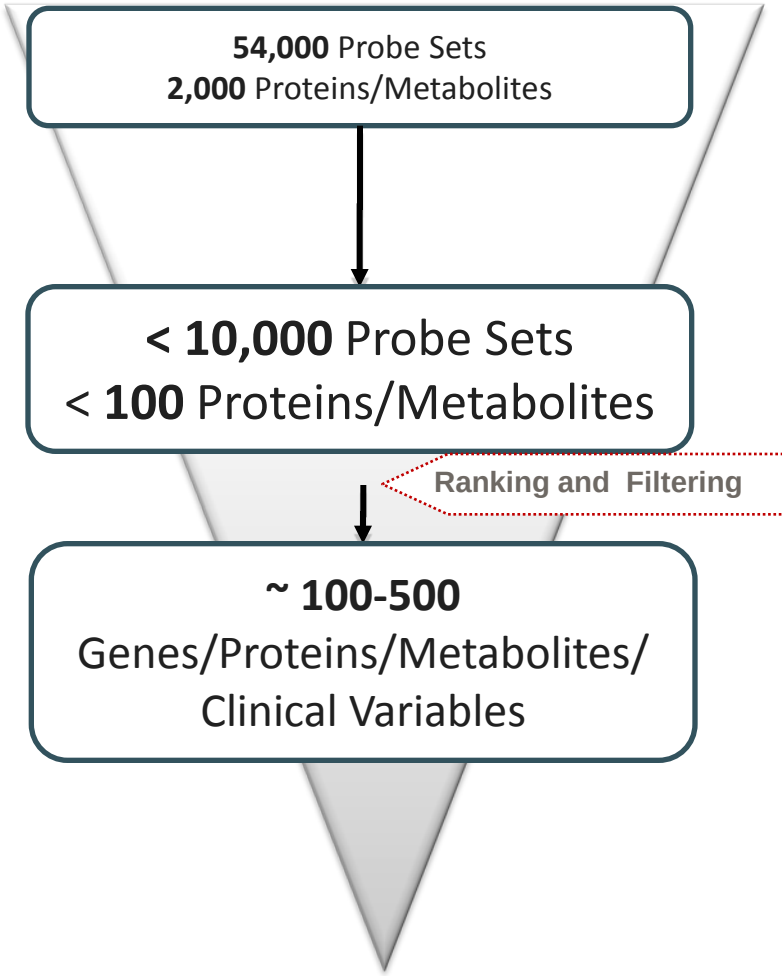


Finding "Needles in a Haystack"



I. **Remove features with small variations across all samples (rejection or otherwise)**

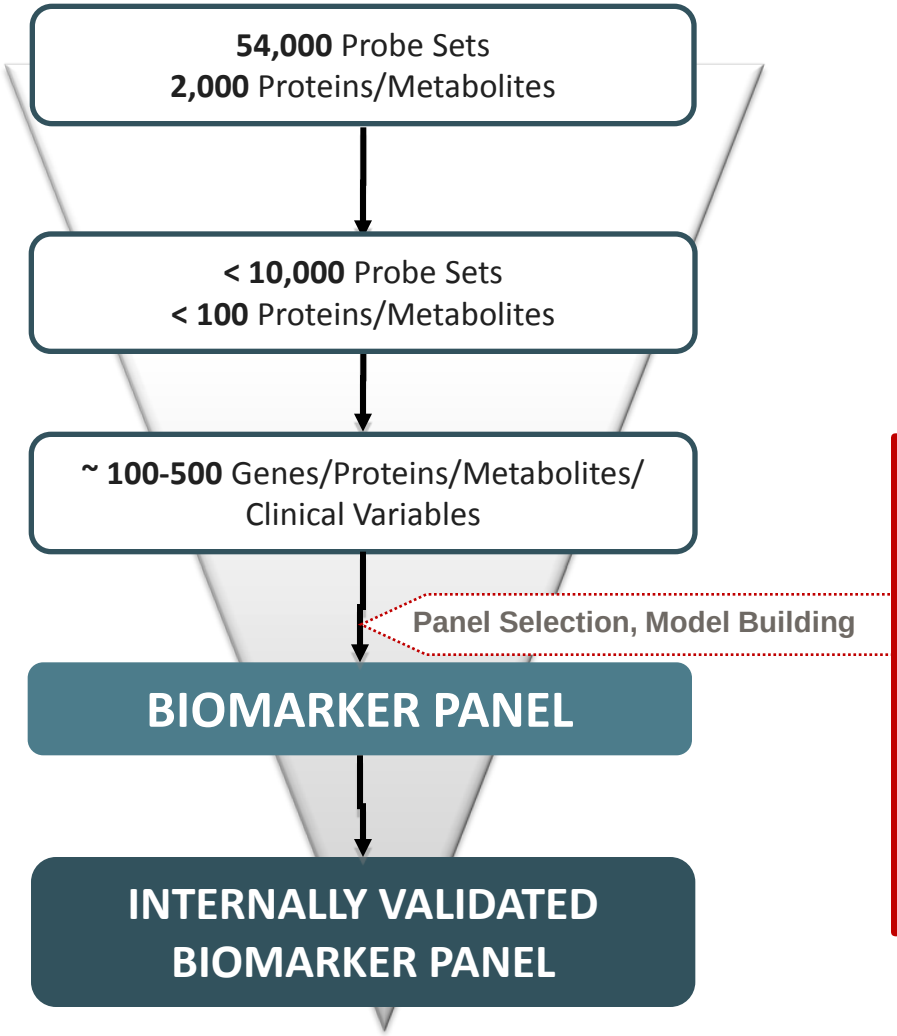
“Needles in the Haystack” (cont)



II. (Univariate) Rank each individual feature on how well it discriminates the rejection samples from non-rejection samples

III. (Multi-variate) Rank groups of features together on their joint discrimination power, taking correlation into account

“Needles in the Haystack” (cont)



IV. Select features to be included in the panel, possibly assigning different weights to different features

Rich Space for Choices

Pre-filtering (remove probe-sets with low variability)	1)	k samples above absolute threshold
	2)	First half using inter-quartile range
	3)	First half using empirical central mass range
Uni-variate ranking (FDR-based; per probe-set)	1)	Maximum of LIMMA, robust LIMMA and SAM
	2)	LIMMA
	3)	Robust LIMMA
Uni-variate filtering (per probe set)	1)	FDR cut-off (FDR<0.01)
	2)	Size cut-off: Top 50 probe-sets
	3)	Combination rule: FDR<0.05 but at least 50 and at most 500 probe sets
Multi-variate ranking (optional)	1)	Stepwise Discriminant Analysis
	2)	SVM-based ranking (one step)
	3)	Recursive Feature Elimination (multi-step)
	4)	Elastic Net-based (coefficients)
Multi-variate filtering (optional)	1)	Significance of improvement cut-off
	2)	Top 50 (as returned by multi-variate ranking)
	3)	Non-zero coefficients (Elastic Net)
Classifier Generation	1)	Linear Discriminant Analysis
	2)	Support Vector Machine
	3)	Random Forest
	4)	Elastic Net
	5)	Logistic regression