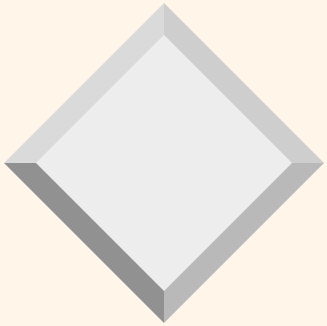


Outlier Detection in Personalized Medicine

Raymond T. Ng
Chief Informatics Officer, Proof Centre of Excellence
Professor, Computer Science, University British Columbia

PROOF Centre: Who are we? What Problems are we Tackling?

- Centre for the **PR**evention **O**f **O**rgan **F**ailures is a non-profit organization established in 2008
- Federally and industry funded to create a world-class national centre of excellence
- Canadian annual expenditures exceed $70 billions on end-stage heart, lung and kidney diseases
- Growing burden due to aging populations

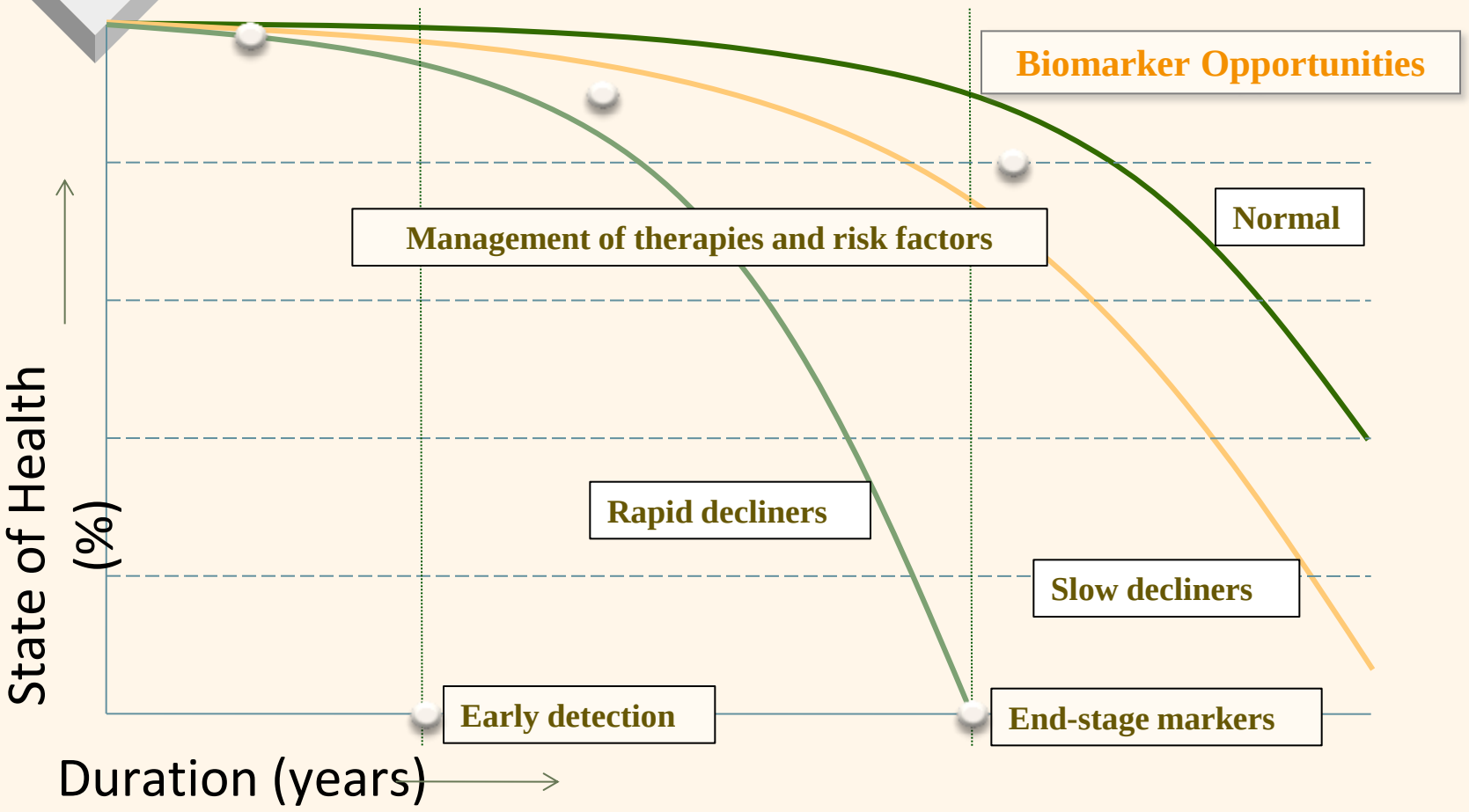
Our Approach: Biomarker Solutions

Distinct biological indicators (cellular, biochemical or molecular) of a process, event or condition that can be measured reliably in tissues, cells or fluids

BIOMARKER SOLUTIONS

***Sets* of genes and/or proteins in the *blood* identified using whole genome technologies**

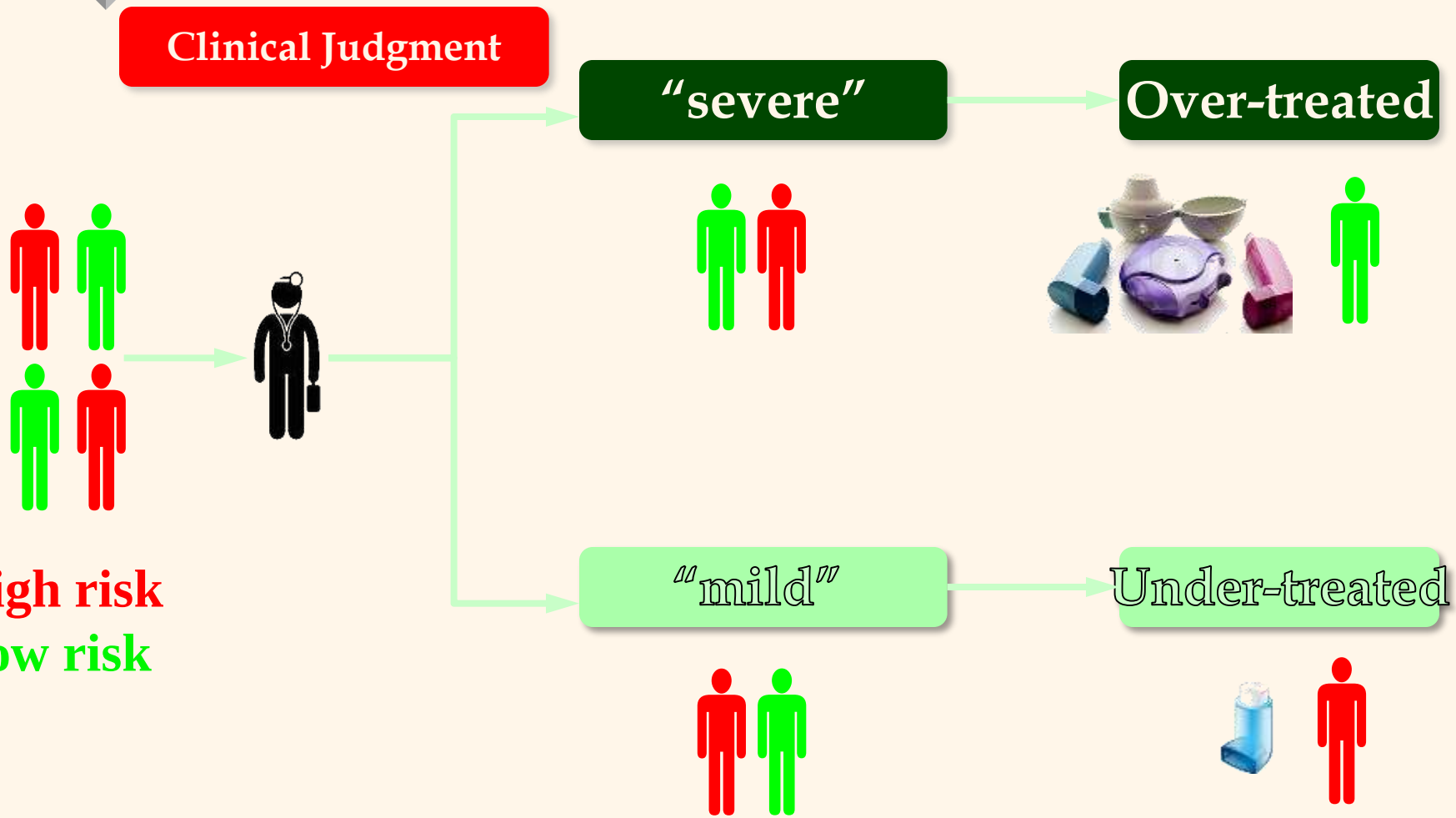
Impacting the Life Cycle of Organ Failures



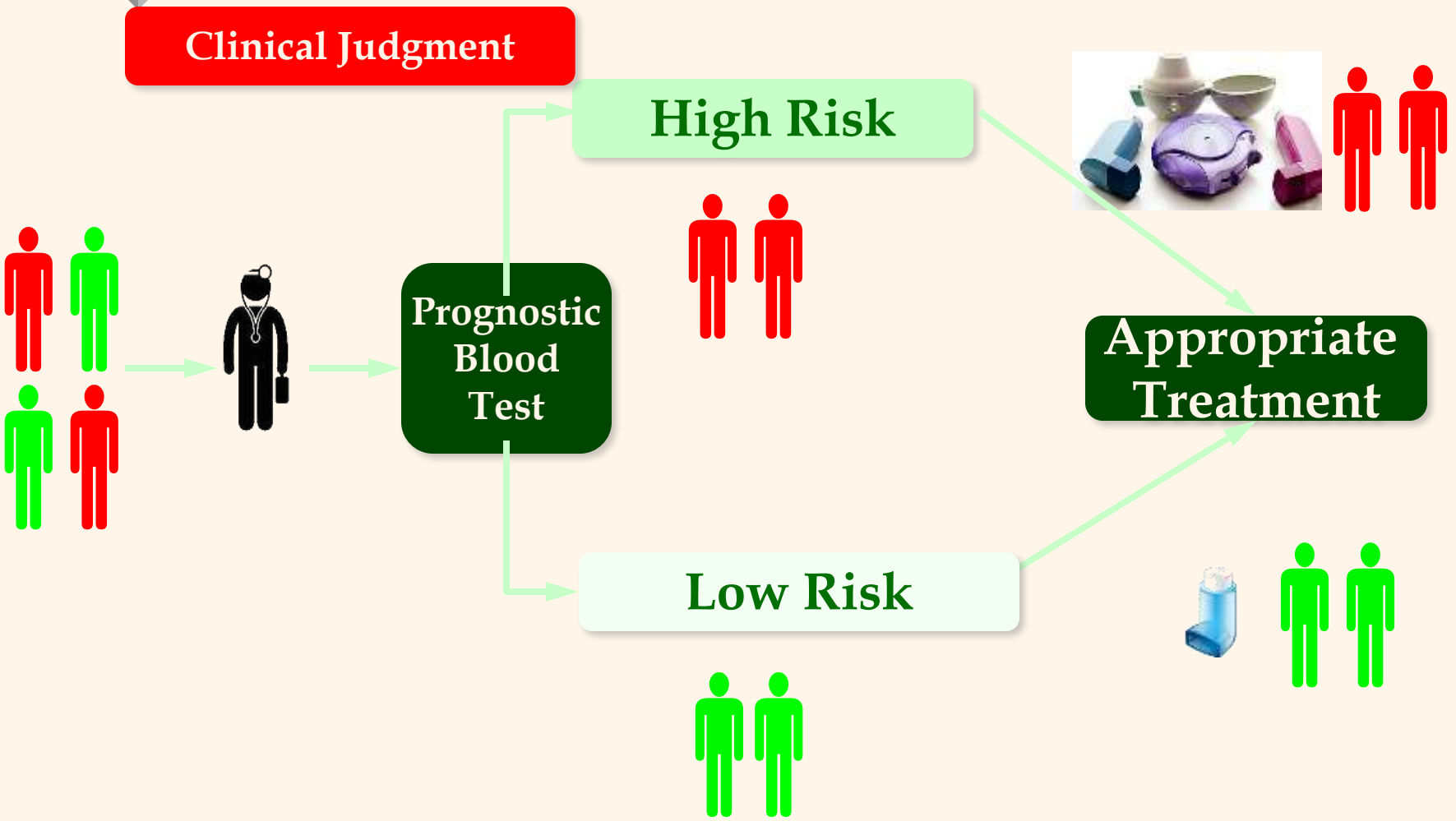
Our Program in Chronic Obstructive Pulmonary Diseases (COPD)

- Over 1.4M Canadians and 14M Americans suffer from COPD
- Exacerbations ("lung attacks") can be lethal and require extensive hospitalization
- Currently difficult to predict who will progress from no exacerbation to frequent exacerbations
- We are developing a prognostic blood test to help identify "rapid decliners"

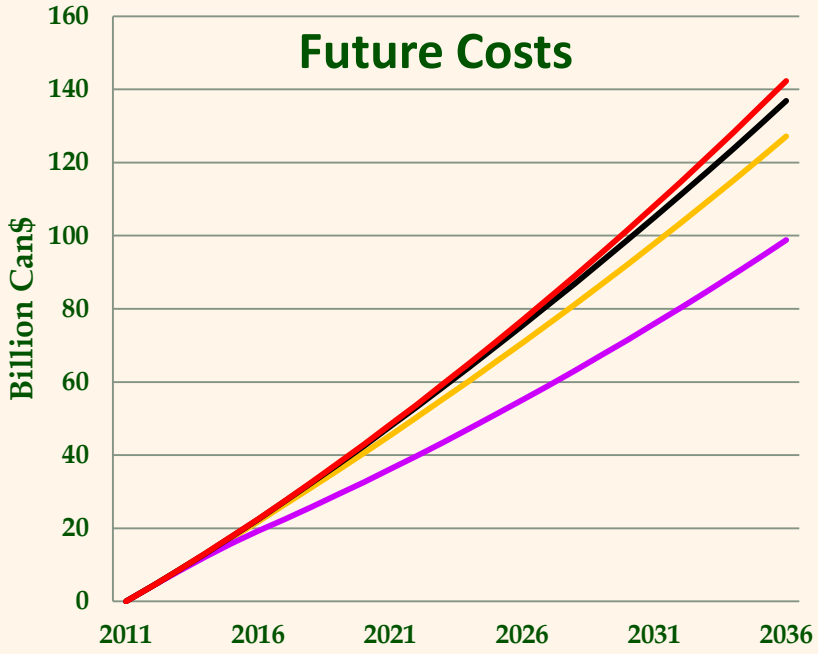
Current Management of COPD

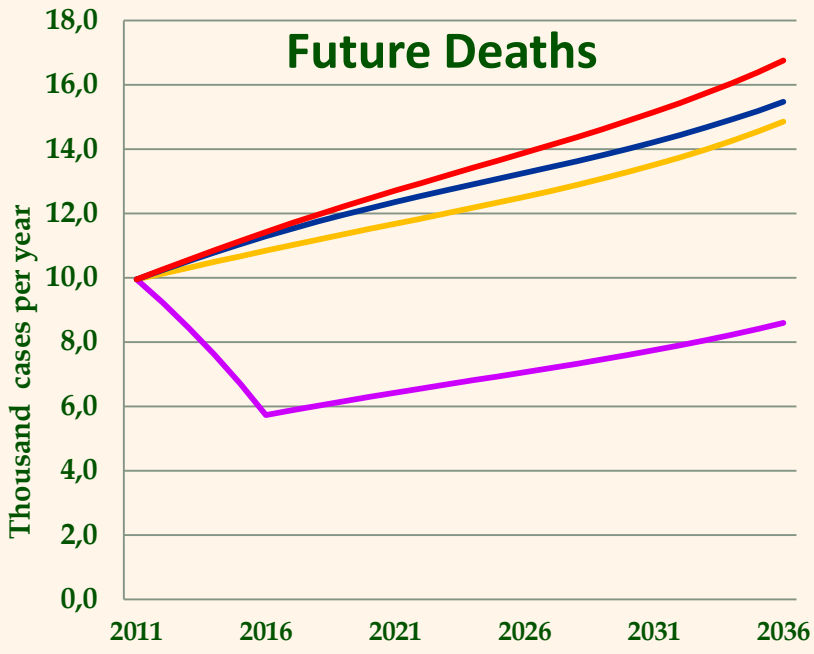


How will a prognostic test change care?



Impact of a COPD Exacerbation Blood Test





Base case
Smoking cessation
New drugs to enhance lung function
New tests to prevent exacerbation

PROOF Centre Mandate

- Working with health care providers, industries and governments to bring these blood tests to patients within 3 years
- Hope to obtain FDA approval in 5 years
- Cheaper and yet better medicine

- *What do all these have to do with Outliers and Outlier Explanations??*

A. Detecting Potential Labeling Errors

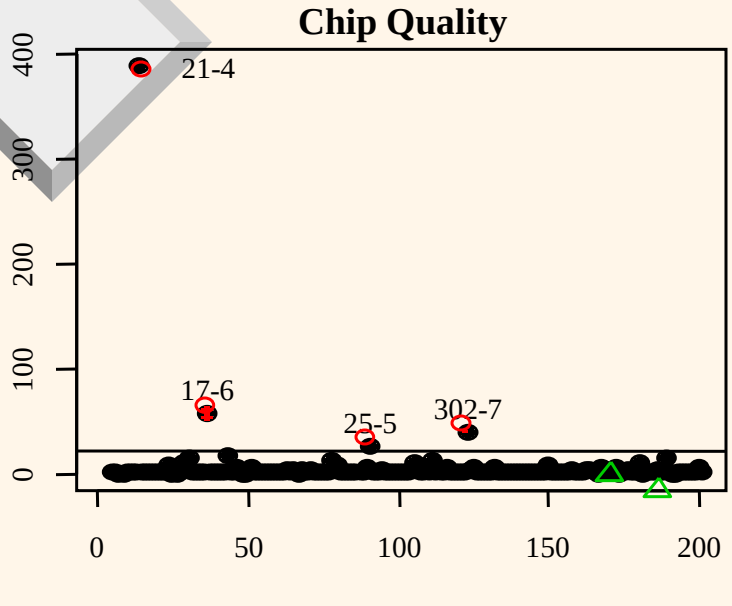
- Biomedical data can be very noisy:
  - Laboratory environment could change
  - Diagnostic decisions are not completely objective
  - Different "gold-standards" are used for grading
- Essential to check for label (e.g., severity of condition) consistency

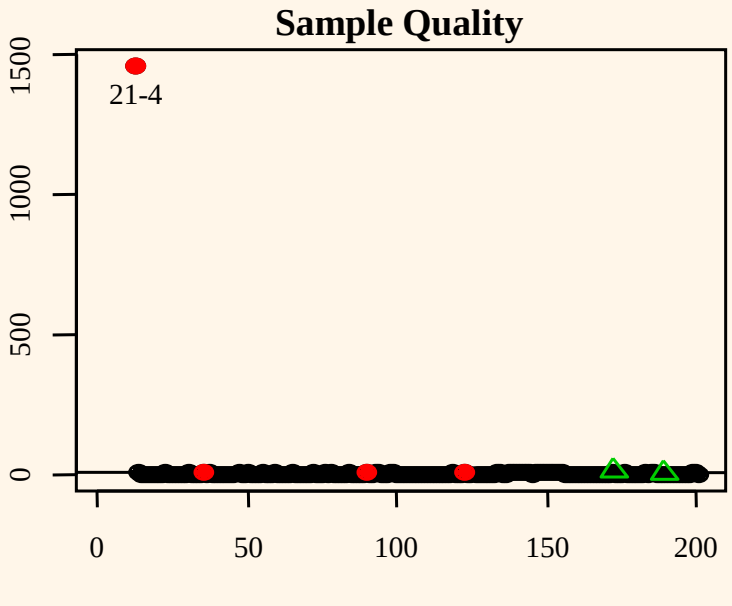
A. Our Approach [MBN06]

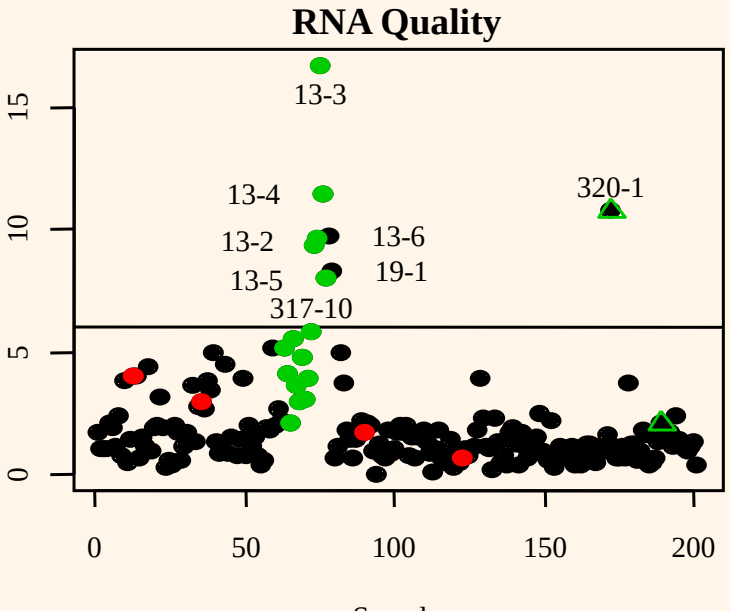
- Propose a leave-one-out perturbed classification matrix:
  - Flip every training sample and compare the resulting classifier with the classifier trained on the original training set
  - A training sample A is a suspect of mislabeling if flipping A's label significantly increases cross validation accuracy (using SVM)
- Effectiveness shown on 3 real microarray data sets with ground truths
- Identified a few suspicious cases in PROOF's projects, e.g., timing of blood draws, leading to a journal publication

B. MDQC: Our Approach to Microarray Quality Control [FZN+07]

- A QC report inspects a microarray in isolation
- Goal: to identify outlying arrays that are not evident from inspection of individual arrays
  - High-throughput whole genome technologies process arrays in batches of 96
- collapses all the values in QC reports into measures to assess the quality of the array, the sample, and the RNA
- measures the distances of each array to an "average" array in the study, adjusting for covariances


Chip Quality
21-4
17-6
25-5
302-7
0
100
200
300
400
0
50
100
150
200
Sample


Sample Quality
21-4
0
500
1000
1500
0
50
100
150
200


RNA Quality
13-3
13-4
13-2
13-5
13-6
19-1
317-10
320-1
0
5
10
15
0
50
100
150
200
Sample

B. MDQC Advantages

- Performs a *multidimensional* analysis and not requiring absolute thresholds (which are often arbitrary)
- Easy to implement and visualize, and computationally inexpensive (as compared with Affy PLM)
- Can suggest potential sources of problems and possible batch effects

Summary

- Personalized medicine relies critically on very clean data; outlier detection plays a valuable role
- Whole genome technologies rapidly advancing; but platforms not yet stable; important application domain for outlier detection
- Finding "intelligent" information associated with outliers help the user to understand the outliers, track down potential problems, and intervene as early as possible