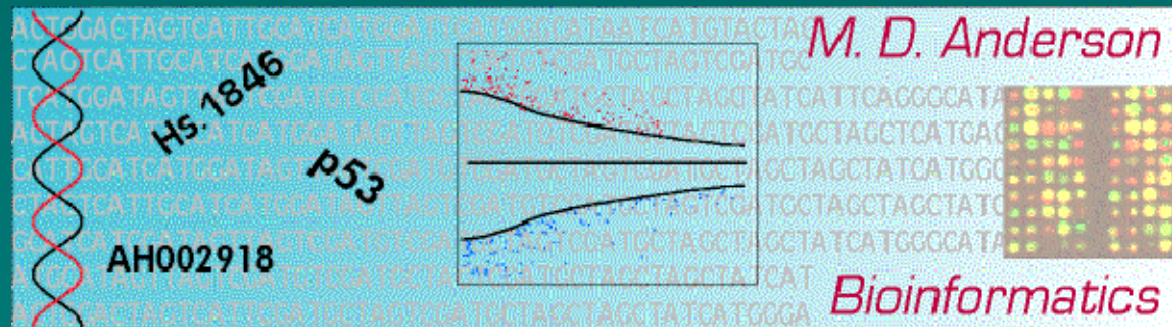


Protein Biomarkers: Panel Discussion

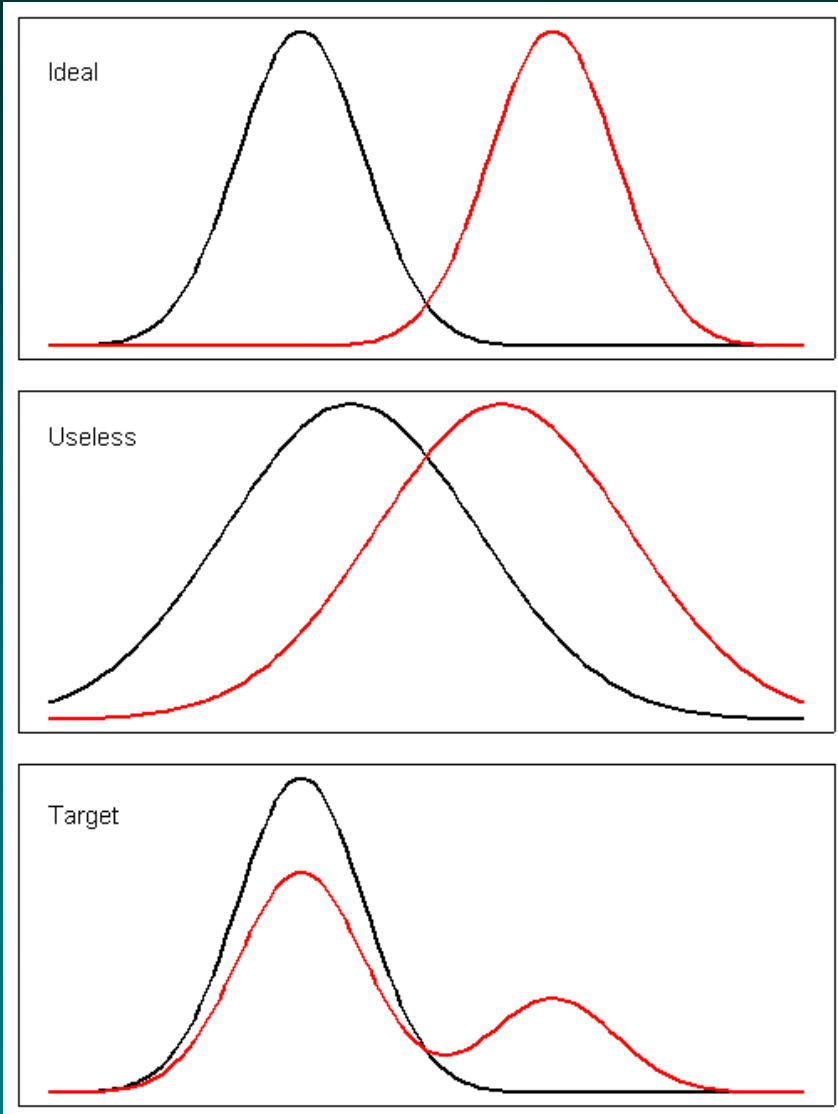
Kevin R. Coombes

Section of Bioinformatics

Department of Biostatistics and Applied Mathematics



Sample size for biomarker discovery



- Class comparison: easy
 - 10-20 per group
- Class prediction: hard
 - Validation needed even at initial discovery
 - Split-sample design (training/test) --- at least 30 per group in both training and test
 - Cross-validation (n-fold repeated) --- at least 50 per group

Patterns

- Cons
 - Multivariate combinatorial explosion
 - 5 out of 10000 = 8.3×10^{17}
 - 10 out of 10000 = 2.7×10^{33}
 - Overfitting
 - Lack of identifiability: cannot shift to more accurate, clinically useful technologies
- Pros:
 - We've been doing single biomarkers for decades. It takes too long, and superb single markers are too rare.

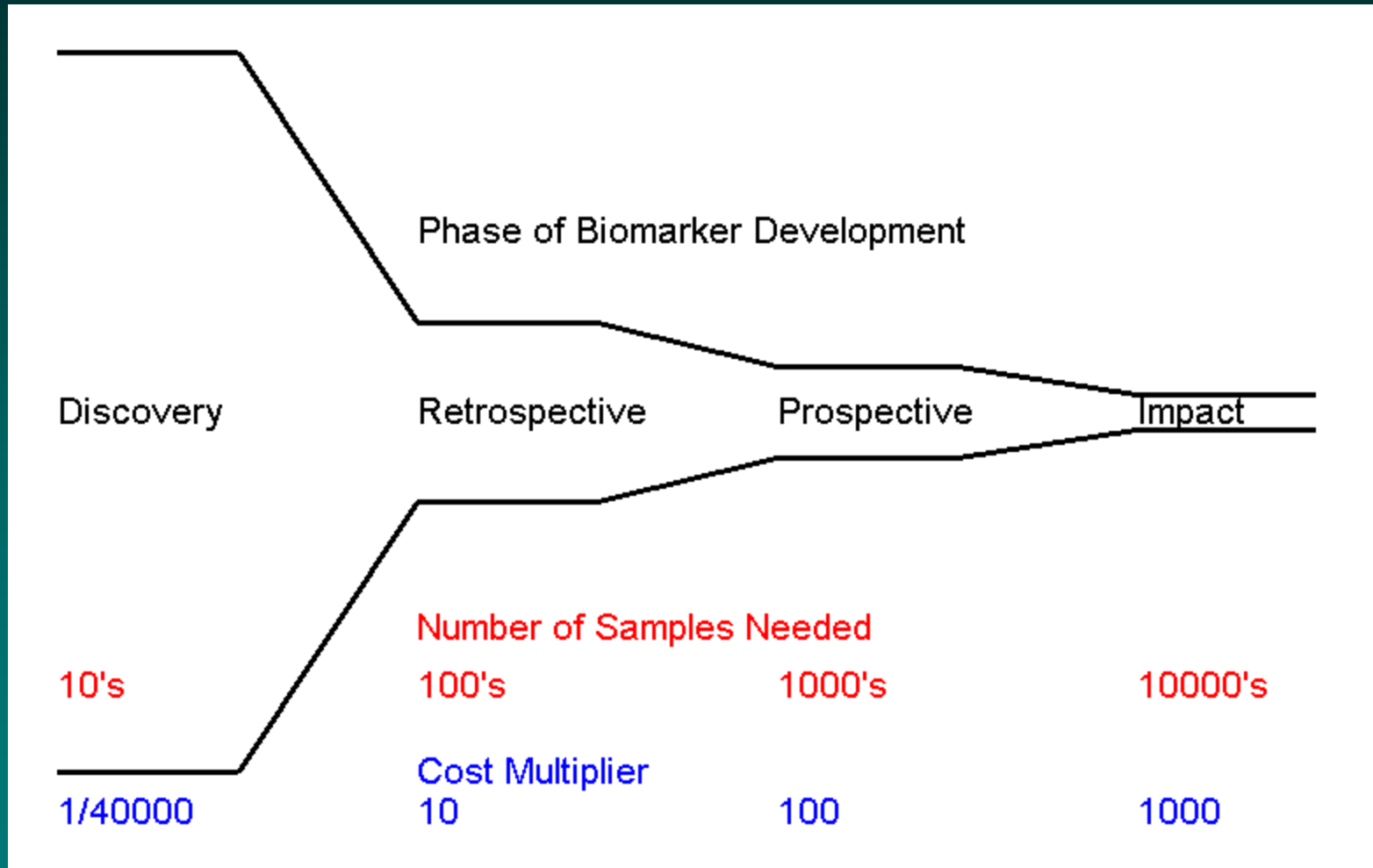
Panels of biomarkers

- GHI study of breast cancer (Paik, NEJM, 2004)
 - PCR on formalin-fixed tissue
 - 21 genes (16 target; 5 reference)
 - Explicit model built on three preliminary studies, then tested on independent study
- Easier to find markers with high specificity (99%) but modest sensitivity (30%)
- Define patient subsets
- Lots of weak markers are at least as good as a single strong marker

Unbiased vs. hypothesis-driven biomarker discovery

- Bayesian statistics makes explicit use of prior information
 - “Unbiased discovery” = “uninformative prior” = every protein is equally likely to be useful
 - “Hypothesis-driven” = highly informative prior = only the proteins I like will be useful
 - Truth probably lies somewhere in between
- Throughput
 - Hypothesis-driven is inherently low throughput
 - “Unbiased” can be high throughput
 - High throughput has a cost in terms of false discoveries

Sample size for validation: Biomarker pipeline



Open access

- Requirements for open access to microarray data have already been imposed by top journals
 - MIAME standard is a start
- NIH is pushing for open access to clinical trial results
- Open access to proteomics data sets should be the goal
 - Ciphergen software produces XML files that could be used as a first draft of a standard for data exchange