Published in final edited form as:

Clin Exp Metastasis. 2012 October ; 29(7): 641–643. doi:10.1007/s10585-012-9508-y.

Rigorous Quantitative Sciences Integration — The Foundation of High-Dimensional Genomic Research

Yu Shyr, Ph.D.

Yu Shyr: yu.shyr@vanderbilt.edu

Overview

One of the greatest challenges facing biostatisticians and bioinformaticians today is the high-dimensionality of genomic data. Fifteen years ago, cDNA data with 5,000 variables was high-dimensional; biotech companies then developed microarray platforms with 45 to 50 thousand probes, followed by SNP arrays that further increased the number of variables to half a million, then a million. Next-generation sequencing can generate 3 billion variables. In short, how high is ‘high’-dimensional? The answer continues to grow, keeping pace with improvements in technology that generate ever more rapid and affordable sequence data.

Such technological advancements bring with them a great challenge for data analysis. Processing 10GB of RNA-seq data requires 30GB of storage space; processing 100GB of whole genome sequence data requires 200 to 300GB. As the volume of data continues to increase, traditional methods of data management and analysis may no longer be practical; for example, when 20 million rows of data must be read, reading the entirety of the data prior to starting the analysis may no longer be resource-efficient. This challenge will only grow, as next-generation sequencing gives way to third-generation sequencing.

Experimental Design

The challenge of data size demands an efficient study design. Genomic studies tend to fall into three general designs, and the success of a study hinges on choosing a design that best fits the study goals and resources of the lab.

1) Patients (training) → Genes → Patients (testing) → Biology

Ten years ago, many studies started with microarray analysis of patient bio-specimens to identify ‘winner’ genes (i.e., candidate biomarkers with differential expression in cases vs. controls) [1–4]. Winners validated in an independent patient cohort were examined further in the lab, to investigate the biology behind the markers. This strategy, however, was largely unsuccessful. The variation across patients and tumors was too great, sample size tended to be too small, and many covariates were not controlled; the data were often over-fitted to model winners, and this resulted in many failed studies.

2) Biology (cell lines, animal model) → Genes → Patients

A more successful study design begins with biology. Genes with biological functions of potential relevance for the disease of interest are identified and studied in cell lines or animal models, prior to moving to the patient. This approach eliminates the fishing aspect of design (1), due to the biological rationale underlying the signature of interest.

As an example of this type of design, my group collaborated on a study published in *Gastroenterology* (2010) [5], which began with development of a highly metastatic

derivative of the MC-38 cell line in a mouse model, then analysis of MC-38 parental and MC-38met cells to identify 300 significant genes involved in invasion and metastasis. These 300 genes were analyzed for directional concordance of expression in MC-38met cells and a training dataset of 19 high-risk (metastatic colon cancer) patients at Vanderbilt Medical Center; this refinement step yielded a 34-gene recurrence classifier. Applying this classifier to a testing cohort of 177 patients from Moffitt Cancer Center, we generated metastasis scores that proved significant for overall and disease-specific survival. In addition, the 34-gene metastasis score was able to discriminate high and low risk of recurrence in patients with stage II and III colon cancer, thus demonstrating potential for translation of the gene signature to the clinical setting for treatment selection.

3) Multiple data sets (GEO) → Data mining → Biology (cell lines)

Another successful study design is one ideal for a small lab without access to animals and patients. Thousands of publicly available datasets, such as those curated by the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) projects, may be downloaded and mined for data of interest. When the data yield promising findings, these findings may be taken back to the lab to confirm the underlying biology in cell lines. In this method, in contrast to method (1), the enormous available sample size overcomes the potential stumbling block of patient and tumor variability.

A study from the Vanderbilt-Ingram Cancer Center published in *Journal of Clinical Investigation* (2011) [6] used this approach to identify potential subgroups in triple negative breast cancer and then find therapeutic methods to target each subgroup.

We started with data from 2,500 patients in 21 publicly available breast cancer databases, and normalized across the data to correct for differences among platforms, studies, and patient baselines. In the training dataset, we identified six subgroups of triple negative breast cancer; these six subgroups were confirmed in our validation dataset, as well as in vitro in cell line studies. These six subgroups, a novel classification of triple negative breast cancer, may be translated clinically to inform treatment options for patients—and the data used to develop the gene signature was derived entirely from publicly available datasets.

Lessons from Duke University

The challenges for genomic studies extend beyond study design, as dataset size also has the effect of increasing the risk of basic informatics mistakes during data processing. One of the most prominent examples of such an error arose in clinical trials conducted by Duke University, which involved use of gene signatures to predict clinical outcome in cancer patients. The original paper to describe the findings of these trials, published in *Nature Medicine* (2006) [7] by Potti et al., appeared at first blush to promise an exciting development in cancer treatment. This development, however, was soon debunked, when M.D. Anderson Cancer Center biostatisticians Dr. Keith Baggerly and Dr. Kevin Coombes [8] were unable to replicate the results reported by Potti et al. Instead, they found a series of errors in Duke's data analysis, including gene mislabeling and an "off-by-one" error, in which certain data had been copied from one spreadsheet to another without a necessary shifting of values by one row to account for column headings; this off-by-one error resulted in mismatches between genes and gene probe identifiers.

As a result of these errors, the validity of the gene signatures findings was called into question, as was the safety of the associated clinical trials. The trials were suspended, but later restarted after an external review of the methodology. The resumption of the trials raised great concern in the biostatistical community and prompted a joint letter from 33

biostatisticians to National Cancer Institute Director Dr. Harold Varmus, expressing concern over the prediction model used in these trials and urging the NCI to re-suspend the trials.

The trials were permanently suspended; the NCI requested that the Institute of Medicine investigate the research group; and 10 papers published by Potti, including those in *Nature Medicine*, the *Journal of Clinical Oncology*, and the *New England Journal of Medicine*, based on these findings have been retracted with numerous corrections and partial retractions to others.

This disaster was not the result of poor study design or incorrect statistical methodologies but simple informatics mistakes made during the handling of huge quantities of data. Thus, this case provides an object lesson in the necessity of quantitative sciences integration, to bring together the disciplines required to store, process, manage, and analyze high-dimensional data accurately and successfully. Having a single biostatistician analyze a dataset is no longer sufficient in the context of genomic studies; the datasets are simply too large, and it is too easy to make costly mistakes.

Integration

With the size of high-dimensional datasets today—and with the inevitable further expansion of this size, even in the immediate future—we must develop a model for genomics studies that includes data management and analysis by an integrated team from a variety of backgrounds, combining expertise across fields of quantitative science. Biomedical informatics, bioinformatics, biostatistics, and computer science all are key disciplines—with biomedical informaticians to design databases and develop systems for linking raw molecular data with clinical data; bioinformaticians to focus on the pre-processing of data as well as pathway and network analysis; biostatisticians to develop and validate algorithms for modeling complex data; and computer scientists to write efficient and accurate code for dealing with millions of rows of data.

All of these disciplines are critical to the success of genomics studies—and, in turn, successful collaboration among these disciplines requires a leader who can oversee and encourage communication and cooperation among these specialists. In short, accuracy, success, and productivity in the area of genomics will increasingly come to depend on an integration of quantitative sciences.

References

1. Shyr Y. Statistical strategies for analyzing the microarray data in human lung cancer. *Lung Cancer* 2003; 2003; 41:90–91.
2. Yamagata N, Shyr Y, Yanagisawa K, Edgerton M, Dang TP, Gonzalez A, Nadaf S, Larsen P, Roberts JR, Nesbitt JC, Jensen R, Levy S, Moore JH, Minna JD, Carbone DP. A training-testing approach to the molecular classification of resected non-small cell lung cancer. *Clinical Cancer Research*. 2003; 9(13):4695–4704. [PubMed: 14581339]
3. Yanagisawa K, Shyr Y, Xu BJ, Massion PP, Larsen PH, White BC, Roberts JR, Edgerton M, Gonzalez A, Nadaf S, Moore JH, Caprioli RM, Carbone DP. Proteomic patterns of tumour subsets in non-small-cell lung cancer. *Lancet*. 2003; 362(9382):433–439. [PubMed: 12927430]
4. Xu BJ, Gonzalez AL, Kikuchi T, Yanagisawa K, Massion PP, Wu H, Mason SE, Olson SJ, Shyr Y, Carbone DP, Caprioli RM. MALDI-MS derived prognostic protein markers for resected non-small cell lung cancer. *Proteomics – Clinical Applications*. 2008; 2(10):1508–1517. [PubMed: 21136798]
5. Smith JJ, Deane NG, Wu F, Merchant NB, Zhang B, Jiang A, Lu P, Johnson JC, Schmidt C, Bailey CE, Eschrich S, Kis C, Levy S, Washington MK, Heslin MJ, Coffey RJ, Yeatman TJ, Shyr Y, Beauchamp RD. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology*. 2010; 138(3):958–968. [PubMed: 19914252]

6. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, Pietenpol JA. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *Journal of Clinical Investigation*. 2011; 121(7):2750–2767. PMID: PMC3127435. [PubMed: 21633166]
7. Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R, Cragun J, Cottrill H, Kelley MJ, Petersen R, Harpole D, Marks J, Berchuck A, Ginsburg GS, Febbo P, Lancaster J, Nevins JR. Genomic signatures to guide the use of chemotherapeutics. *Nature Medicine*. 2006; 12(11):1294–1300.
8. Baggerly KA, Coombes KR. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Annals of Applied Statistics*. 2009; 3(4):1309–1334.