Vaccine xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Vaccine

journal homepage: www.elsevier.com/locate/vaccine

Lessons learned in the analysis of high-dimensional data in vaccinomics

Q1 Ann L. Oberg[a,d], Brett A. McKinney[b], Daniel J. Schaid[a,d], V. Shane Pankratz[c], Richard B. Kennedy[d], Gregory A. Poland[d,*]

Q2 [a] Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
[b] Tandy School of Computer Science, Department of Mathematics, University of Tulsa, Tulsa, OK, USA
[c] UNM Health Sciences Library & Informatics Center, Division of Nephrology, University of New Mexico, Albuquerque, NM, USA
[d] Mayo Clinic Vaccine Research Group, Mayo Clinic, Rochester, MN, USA

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Vaccines
Vaccination
Systems biology
Immunogenetics
Q3 Data interpretation, statistical

ABSTRACT

The field of vaccinology is increasingly moving toward the generation, analysis, and modeling of extremely large and complex high-dimensional datasets. We have used data such as these in the development and advancement of the field of vaccinomics to enable prediction of vaccine responses and to develop new vaccine candidates. However, the application of systems biology to what has been termed "big data," or "high-dimensional data," is not without significant challenges—chief among them a paucity of gold standard analysis and modeling paradigms with which to interpret the data. In this article, we relate some of the lessons we have learned over the last decade of working with high-dimensional, high-throughput data as applied to the field of vaccinomics. The value of such efforts, however, is ultimately to better understand the immune mechanisms by which protective and non-protective responses to vaccines are generated, and to use this information to support a personalized vaccinology approach in creating better, and safer, vaccines for the public health.

1. Introduction

Q4 Like personalized medicine, personalized vaccinology aims to provide the right vaccine, to the right patient, at the right time, to achieve protection from disease, while being safe (*i.e.*, free from unintended side effects). The science through which such a vision can be realized is a field we have developed and termed "vaccinomics," which is grounded within the immune response network theory [1–5]. The immune response network theory states that "the response to a vaccine is the cumulative result of interactions driven by a host of genes and their interactions, and is theoretically predictable."[1] Further, the theory recognizes the impact of metagenomics, epigenetics, complementarity, epistasis, co-infections, and other factors including polymorphic plasticity. These factors, and others, explain the temporal, genetic, and immune responses that are deterministic and predictive of the immune response. In fact, we have published an initial equation describing this outcome [4]. Vaccinomics then uses this information to reverse engineer new vaccine candidates that overcome genetic or other barriers to the development of protective immunity.

Importantly, the above requires systems-level high-dimensional data in order to understand, at the whole-system level, the many perturbations a vaccine might induce in a host that result in immunity. Like any network, the immune response is composed of connected genetic features and networks of feedback loops. While exciting science, it makes application of systems biology to immune responses generated by vaccines in the individual challenging.

High-dimensional assays are generally resource intensive and typically used to perform an unbiased assessment of a biological system in order to generate hypotheses for further investigation. For example, measuring mRNA expression via next generation sequencing (NGS) allows one to screen approximately 20,000 genes for association with an outcome such as vaccine response. Although results from such screening studies must be replicated or functionally validated, it is nonetheless critical to avoid false-positive and false-negative findings by paying close attention to study design and analysis plans [6,7]. Our goal herein is to convey some of the lessons we have learned over the last decade regarding sound principles of study design and analysis in experiments utilizing

* Corresponding author. Tel.: +1 507 284 4968; fax: +1 507 266 4716.
E-mail addresses: poland.gregory@mayo.edu, vitse.caroline@mayo.edu (G.A. Poland).

http://dx.doi.org/10.1016/j.vaccine.2015.04.088
0264-410X/

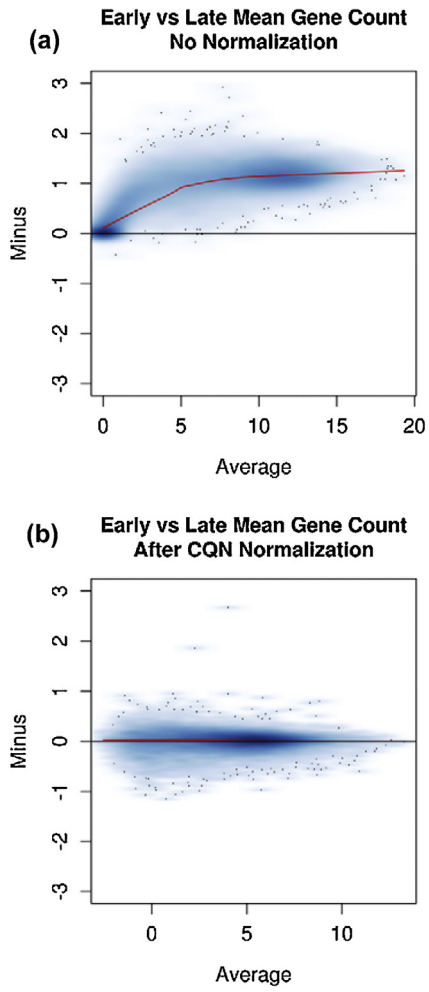

Fig. 1. (a) Minus versus average (MVA) plot demonstrating the effect of change in reagents and sequencing software. There is one data point for every feature measured on the assay. The *x*-axis is the average of each feature over all specimens in the study. Generally, the *y*-axis is the difference of each feature from the mean. Thus, if the observations are identical to the mean, all data points would lie on the $y = 0$ line. Here, the *y*-axis is the difference between the before reagent change mean and the after reagent change mean. A reference line for $y = 0$ as well as a loess smoother are included on the plot. If the smoother overlays the $y = 0$ line, no normalization is needed. If the smoother is parallel to the $y = 0$ line but shifted up or down, this indicates that between specimen biases are similar for all abundance levels and a linear normalization is needed. The nonlinear smoother demonstrates that nonlinear bias is present. (b). The same study shown after normalization and filtering out genes with median count <32. The fact that the smoother is now straight and lies on the $y = 0$ line demonstrates that the bias has been removed.

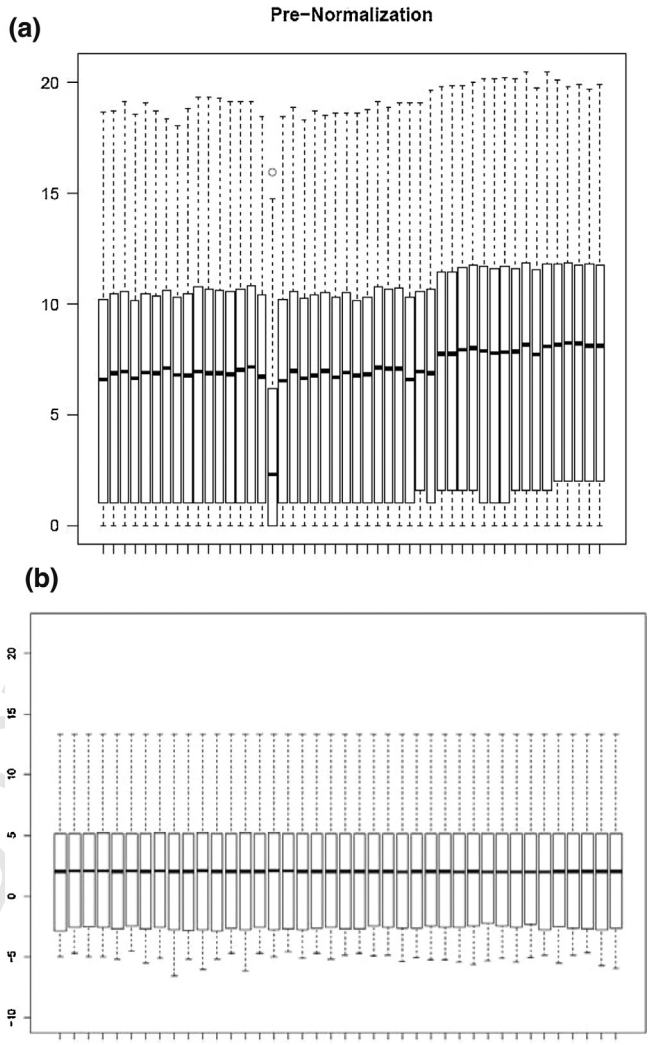

Fig. 2. Box-and-whisker plots showing global distribution of per-gene counts on the log scale (*y*-axis) by lane (*x*-axis) sorted by assay order. Top, mid-line and bottom of boxes indicate 75th, 50th and 25th percentiles, respectively. (a) Pre-normalization. The total counts/lane increased from ~150 million to ~200 million after reagent and software upgrades. This is evident from the general shift up approximately two-thirds of the way across the plot. A failed specimen with median nearly half that of the neighboring specimens is evident about one-third of the way across the plot. The failed specimen was deleted in subsequent analyses. (b) Post-normalization. After normalization via Conditional Quantile Normalization (CQN) [81], the distributions of the specimens are aligned exactly at the maximum, 75th and 50th percentiles as expected. The lower counts are not exactly aligned since the smallest counts are not adjusted in CQN.

high-dimensional assays such as gene expression or genome wide SNP association studies, as applied to the field of vaccinology. While some of the points may appear simple and straightforward, which makes them easy to overlook, they require effort to implement in practice. We discuss study design, normalization, modeling, and determining statistical significance.

2. Study design

The relative abundance measures produced by most high-dimensional assays are susceptible to experimental artifacts such as batch effects [8]. Such artifacts add uninteresting, and often misleading, variation to the data. For example, in an mRNA Seq study we performed involving over 450 specimens, reagents were upgraded midway through the study, increasing the reads/lane by about 50% (Figs. 1 and 2). Usually, the causes of batch effects are less obvious than a reagent change. Thus, randomization, balance, and blocking are vital to ensure that biological effects and experimental artifacts can be distinguished from one another. Further information and detailed examples for how to implement these methods in practice are available [6,9–12].

It is critical that study outcomes are appropriately defined. When studying vaccine response, known or established correlates of protection are typically used [13]. For example, an antibody level of at least 0.15 μg/ml is considered protective against Haemophilus influenza type b. In other cases, commonly accepted levels of immunity are used as surrogates of protection (e.g., a hemagglutination inhibition antibody (HAI) titer of 1:40 for influenza or a neutralizing antibody titer of 1:32 for smallpox). These surrogates of protection represent a best guess at what level of immune response is sufficient to protect against overt disease at the population level.

Just as each pathogen and disease is different, so too are the immunologic responses critical for protection. Therefore, multiple

Please cite this article in press as: Oberg AL, et al. Lessons learned in the analysis of high-dimensional data in vaccinomics. Vaccine (2015), http://dx.doi.org/10.1016/j.vaccine.2015.04.088