Vaccine xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Vaccine

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

High-throughput data analysis and data integration for vaccine trials

January Weiner 3rd*, Stefan H.E. Kaufmann**, Jeroen Maertzdorf

Department of Immunology, Max Planck Institute for Infection Biology, Charitéplatz 1, D-10117, Berlin, Germany

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Data integration
Systems vaccinology
Systems biology

ABSTRACT

Rational vaccine development can benefit from biomarker studies, which help to predict, optimize and evaluate the immunogenicity of vaccines and ultimately provide surrogate endpoints for vaccine trials. Systems biology approaches facilitate acquisition of both simple biomarkers and complex biosignatures. Yet, evaluation of high-throughput (HT) data requires a plethora of tools for data integration and analysis. In this review, we present an overview of methods for evaluation and integration of large amounts of data collected in vaccine trials from similar and divergent molecular HT techniques, such as transcriptomic, proteomic and metabolic profiling. We will describe a selection of relevant statistical and bioinformatic approaches that are frequently associated with systems biology. We will present data dimension reduction techniques, functional analysis approaches and methods of integrating heterogeneous HT data. Finally, we will provide a few examples of applications of these techniques in vaccine research and development.

1. Introduction

To an ever increasing extent rational vaccine design depends on an understanding of the underlying biological system by means of methodical, high throughput (HT) data acquisition and analysis at various levels [1]. While it is sometimes stated that all the "easy vaccines have been discovered" [2], it is hoped that systematic large-scale analyses will provide new hypotheses and a faster screening process and that novel computational and statistical approaches will be game changers in the vaccine field [1]. The human immune response can be monitored at various levels, each of which may provide an independent perspective and novel information about the system [3]. For example, blood RNA, frequently used for transcriptomics, is derived from peripheral blood cells; in contrast, metabolites isolated from blood may be released by any tissue or organ. Transcriptomics is likely the most frequently applied HT approach in vaccinology (see [4] for examples). Yet, other types of HT data have also been analyzed, including proteomic and genomic analyses [5].

While the general topic of data integration is frequently raised in a biological context, its precise meaning remains somewhat elusive, or, rather, the term refers to a broad spectrum of questions [6]. Generally speaking, data integration can be understood technically and pragmatically as an attempt to curate or juxtapose heterogeneous data sets in a single database or interface on the one hand, and as an attempt to reconstruct and understand the biological system by using different data sets on the other. While the former presents formidable technical challenges, the latter is primarily interesting from a scientific standpoint.

Different technical approaches can be chosen for integrating heterogeneous data sets for common use and presentation. In data warehousing, the different data sets are stored and linked in a single database that can be accessed by users. Other approaches involve federated databases such as BioMart or integrated web views of different databases. However useful, this type of data integration is usually post hoc, and hence does not connect platforms within a specific experimental setup involving multiple data types.

In this review we focus on data integration in the context of functional analysis and systems biology. That is, we do not merely describe computational approaches used to solidify and unify data sets (for example, by standardization and data base design). More importantly, we outline a data analysis plan that makes heterogeneous and multidimensional data sets generated in vaccine trials a biomedical reality, in order to generate and test hypotheses about relevant biological mechanisms. In the following, we consider systems biology approaches and functional analysis of data in the context of data integration and heterogeneous data sets.

* Corresponding author. Tel.: +49 3028460514; fax: +49 3028460505.
** Corresponding author. Tel.: +49 3028460500; fax: +49 3028460501.
E-mail addresses: january.weiner@mpiib-berlin.mpg.de (J. Weiner 3rd), kaufmann@mpiib-berlin.mpg.de, sk_editor@mpiib-berlin.mpg.de (S.H.E. Kaufmann).

http://dx.doi.org/10.1016/j.vaccine.2015.04.096
0264-410X/

2. Sample size selection for HT data

Before collecting data it is necessary to select sample sizes suitable for the planned comparisons. Although this review is primarily concerned with data integration and analysis, this issue needs to be considered since a successful analysis depends on choosing a suitable experimental setup. Moreover, the complexity of HT data analysis influences the required experimental design and should be taken into account at the planning stage of a clinical trial. Data integration and systems biology should be considered at the earliest planning stages of any trial.

Analysis of statistical power provides the solution in cases with few variables, but the complexity of the problem grows in HT settings. In a typical experimental design, a minimal effect size is set, and based on known estimates of variability, a specific sample size is chosen. In a multivariate setting, in which several thousands or hundreds of thousands of effects are estimated, other variables need to be similarly considered, accounting for both the procedure of detecting significant changes that follow (e.g., changes in gene expression) and for downstream functional analyses.

In sample size calculations for HT analyses, it is assumed that a certain fraction of variables shows a statistically significant treatment when compared to a placebo group. This effect can be defined by standard effect size calculations or, for example, by log fold changes. An attempt to select a group size that would allow detection of all such variables would result in a heavily overpowered study [7]. Therefore, it is necessary to specify a threshold setting for the percentage of regulated variables (e.g., genes) that are to be detected within the particular set-up. For some platforms, notably RNASeq, implementation of methods for sample size calculation exist. For example Hart et al. [8] have constructed a method for sample size calculation in the case of RNASeq data and implemented it in the form of an R package and an Excel sheet. Also for RNASeq, Ching et al. [9] used publicly available data sets to create an R script for sample size estimation. Another online tool, RNAseqPS, for RNASeq sample size and power calculation has been provided by Guo et al. [10] based on their previously published theoretical methods for sample size calculations [11]. In general, however, sample size determination for HT analysis in a vaccine trial may require a simulated analysis based on either a pilot study or on previous trials. Such a pilot study-based method, which allows sample size calculation even in cases of complex experimental setups (beyond a simple two-group comparison), has recently been presented by van Iterson et al. [7] and implemented in the R package SSPA.

3. Multivariate techniques in analysis of HT data

By their very nature, HT data are highly dimensional; that is, the number of variables (e.g., genes for which the transcriptional profiles are collected) is by orders of magnitude higher than the number of samples. In a basic approach, each variable is treated and tested independently, for example, by running a statistical test (such as a *t*-test or ANOVA) for each gene in a transcriptomics data set. However, in this approach, the sheer number of variables greatly impacts on the statistical power. A large number of variables results in a large number of observations, which will have a low p-value only due to chance (false positives). Therefore it is necessary to apply either a correction for multiple testing (such as the Bonferroni correction) or a method to control the false discovery rate (FDR). By far the most frequently used method in the

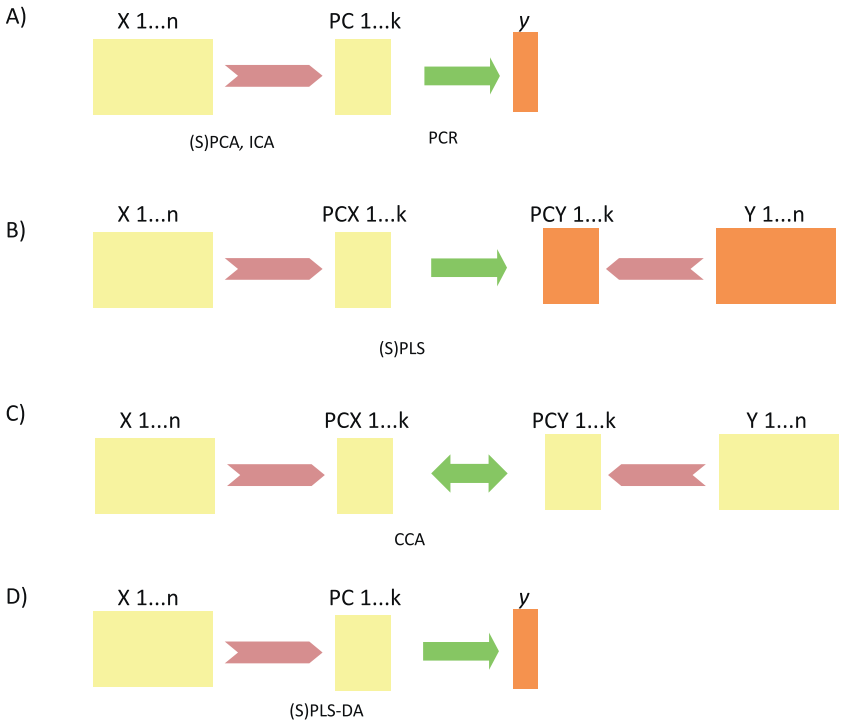

Fig. 1. Dimension reduction techniques for the analysis of highly dimensional data. Yellow X, independent (predictor) variables. Orange Y, dependent (response) variables. Horizontal size of the rectangles indicates the number of variables, vertical size of the rectangles indicate the number of samples. Pink arrows indicate dimension reduction. Green arrows indicate modeling a relationship between data sets. Techniques such as principal component analysis (PCA) and independent component analysis (ICA) (A) allow conversion of a highly dimensional data set into one with a reduced number of variables (dimensions). The principal components (PCs) can then be related to a response variable using PCA regression. Partial least squares (PLS) (B) and canonical correlation analysis (CCA) (C) are techniques in which two data sets are simultaneously reduced in dimensionality. In PLS, there is a clear distinction between dependent and independent variables, while in CCA, the two data sets are symmetric. PLS-discriminant analysis (PLS-DA) (D) is a PLS-based machine learning technique, in which a categorical variable is the modeled response. Abbreviations: sparse partial least squares, (S)PLS; sparse principal component analysis, (S)PCA.

Please cite this article in press as: Weiner 3rd J, et al. High-throughput data analysis and data integration for vaccine trials. Vaccine (2015), http://dx.doi.org/10.1016/j.vaccine.2015.04.096