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, Berlin, Germany

***Correspondence:**

Stefan Kaufmann. Max Planck Institute for Infection Biology, Department of Immunology.

Charitéplatz 1, D-10117 Berlin, Germany. kaufmann@mpiib-berlin.mpg.de.

January Weiner 3rd. Max Planck Institute for Infection Biology, Department of Immunology.

Charitéplatz 1, D-10117 Berlin, Germany. january.weiner@mpiib-berlin.mpg.de.

Running title: Data integration from vaccine trials

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 towards 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.

Keywords: Data integration, systems vaccinology, systems biology

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 already 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 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 which 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.

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.

Multivariate techniques in analysis of high throughput 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 HT field is the Benjamini–Hochberg (B-H) method to control FDR [12]. Although the B-H method is more powerful than the Bonferroni correction, it still substantially decreases the statistical power available for discovery.

In molecular HT techniques, variables are often strongly correlated: functionally related genes show similar expression patterns and changes between conditions; the same can be said about metabolites and proteins. Therefore, it is often advantageous or even necessary to reduce data dimensionality in order to “digest” usable information. Applications of principal component analysis (PCA) include regression (PCA regression) (Figure 1A) and clustering; furthermore, linear discriminant analysis (LDA) is another closely related machine learning algorithm.

Frequently, HT data are transformed using PCA in which the sample space is transformed from one set of coordinates (e.g., gene expression values) to another (principal components; PCs). PCs are linearly independent (orthogonal), and the fraction of the total sample variance contained in a given PC decreases with its rank. In essence, a high number of variables is replaced by few “composite variables” – PCs, which capture most of the variance in the data set. This allows, foremost, for easy visualization of the data. Since differences between experimental groups are usually a substantial fraction of total variance, they often appear in one of the first PCs. While PCs are linearly independent, they need not be statistically independent. This requirement forms the basis for another method, the independent component analysis (ICA). In some applications, ICA has been shown to decompose into components that correspond better to biological entities (e.g., pathways) than PCA [13]. Additionally, many multivariate methods such as sparse PCA (SPCA) or sparse partial least squares (SPLS) allow variable selection using relevant statistical penalizations [14].

Worth mentioning is an additional complication that has not yet been sufficiently appreciated in vaccine systems biology data analysis. Many multivariate techniques produce data that are two-dimensional: the first dimension corresponds to the samples, and the second dimension to genes. In this case, each sample is a vector of values that correspond to the variables analyzed (e.g., a set of gene expression values read out from a microarray experiment for that sample). Frequently, data from vaccine trials are grouped in a time series, which adds a third dimension – time – to the existing two (Figure 2). In this way each sample corresponds to a *tensor* – a two-dimensional matrix of

variables, in which each variable itself consists of multiple measurements at different time points. Although application of relevant tensor techniques has yet to find its way into systems vaccinology, it is worth noting that methods exist that naturally expand on concepts of multivariate techniques and thus can be applied to vaccine studies involving time series. For example, the counterparts of PCA are canonical polyadic decomposition (CPD, also known as PARAFAC) or multilinear principal component analysis (MPCA; [15]).

These and related methods are natural choices in the analysis of heterogeneous data, not only because variable reduction and variance normalization allow balancing between data sets. Specific multivariate techniques link different multivariate data sets. For example, both partial least squares (PLS) and canonical correlation analysis (CCA) allow identification of related variables from two different data sets and can be used to identify biological entities (e.g., genes and metabolites or genes and proteins) that are correlated (Figure 1B and C). The classical PLS algorithm is the multivariate pendant for linear regression. However, a one-dimensional predictor variable (e.g., change in the expression of a gene in different samples) and one-dimensional response variable (e.g., a clinical parameter) are replaced by two-dimensional matrices, each consisting of a number of variables. In PLS, one then maximizes the covariance of transformed components derived from a predictor matrix (with multiple samples and multiple predictors) and a response matrix (with multiple samples and multiple response variables). If PLS corresponds in a way to linear regression, then CCA corresponds to correlation analysis, in that both matrices in CCA are equivalent.

Liquet et al. [16] used PLS and PLS-discriminant analysis (PLS-DA) (Figure 1D) in a two-step approach to analyse results from an HIV vaccine study. In the vaccine study, twelve HIV-negative participants were vaccinated with the HIV-1 LIPO-5 ANRS vaccine, and data (cytokine secretion from antigen stimulation assays and gene expression from stimulated PBMCs) were collected before and after vaccination. The analysis showed a discrimination of cells stimulated by different antigens based on

the readouts (cytokine concentrations and gene expression levels). Furthermore, the PLS analysis was used to connect changes in cytokines with changes in gene expression.

Several tools exist which implement these and other multivariate techniques. While PCA is implemented by virtually every statistical and gene expression analysis package, specialized software exists for PLS, CCA and others. In systems biology, a notable package is the mixOmics suite [17] available as both an R package [18] and an online interface.

Functional analysis of transcription data

Gene expression profiling, particularly of peripheral blood cells, remains one of the most frequent applications of HT techniques in clinical settings. Although both microarrays and RNAseq have been widely applied in this context, other HT and semi-HT techniques, such as Fluidigm and GeneString, are available as well. Typically, gene expression studies reveal a large number of genes that are differentially expressed between experimental conditions. However, to evaluate and to understand the biological meaning of these data, consolidation within higher-order functional schemes, such as gene ontologies, pathways or transcriptional modules, is necessary.

What do we know about genes and other entities?

Compared to functional bioinformatic analysis of non-coding RNAs or non-gene-associated methylation sites, gene expression data and proteomic data are much easier to interpret given the plethora of knowledge about the function of protein-coding genes. Note that this body of knowledge is heavily biased towards a few selected genes that have been the target of numerous studies, while at the same time a large number of genes have never been experimentally analyzed and their functions, if at all given, are only inferred from sequence homology analysis. The distribution of the number of articles that concern a given gene follows the power law: more than 75% of human genes have between 0 and 5 linked PubMed references, almost 50% have 0 linked PubMed reference at all,

while a few “top” genes, including TNF, IL6, APOE and TP53 have been studied in thousands of papers [19]. Similarly, the majority of genes in the human genome do not have an associated gene ontology (GO) term. At the same time, a notable proportion of significantly regulated genes observed in immunological analyses (Weiner and Maertzdorf, personal information) have no meaningful function attributed.

This trend presents a major obstacle for the interpretation of transcriptomic data. Despite an exponentially growing amount of transcriptomic and proteomic data, the functional annotations of genes lag behind. Moreover, despite the obvious lack of necessary annotations, research funding follows the same trend as literature, that is, genes with known functions receive exponentially more attention than genes for which no biological function has been associated. For other platforms, the picture may be even bleaker. Although many conserved microRNAs are known and microRNAs can be annotated by the set of genes which they target, the fraction of microRNAs which are annotated is even smaller [20]. Even less is known about long non-coding RNAs [21]. In case of metabolites, even though their chemistry and participation in the biological cellular metabolic pathways may be well understood, they still remain elusive: firstly, many of the uniquely identified mass spectra are not attributed to known metabolic compounds; secondly, the significance of their changes in serum or urine need not to be related to pathways in which they occur in the producer cells.

Statistical testing of biological functions

Transcriptional analysis of thousands of genes is typically followed by some kind of functional analysis that allows for generalizing the usually massive abundance of detected transcriptional changes. This is frequently done by applying a statistical test of the hypothesis that genes which are differentially regulated, are more likely to belong to a particular functional category, for example, they are more frequently involved in an immune response.

The hypergeometric test is frequently used in this context [22]. Genes under consideration (i.e., all genes spotted on a microarray or found by RNASeq) are split into two groups. The first group (sometimes called “foreground”) consists of genes that are differentially expressed between conditions under consideration, while the second (“background” or gene universe) consists of all genes. The hypergeometric test calculates the probability of observing a specific number of genes attributed to a function under the assumption that sampling of foreground from the universe is random. This test, although still widely spread, requires an arbitrary decision about which genes are defined as differentially expressed. Moreover, the outcome depends on the statistical power used to determine the differentially expressed genes. Finally, as main theoretical problem, the hypergeometric test assumes independence between genes – an assumption clearly at odds with the reality of gene expression regulation. Rejection of the null hypothesis by the hypergeometric test might not indicate enrichment, but rather reflect the fact that genes are indeed statistically correlated [23].

Other approaches do not suffer from these obstacles. Three other tests require a single, ordered list of genes (e.g., by the p-value from differential analysis) as the input. The Mann-Whitney U test can be used to compare ranks of genes belonging to a biological category with those not belonging to this category; this simple approach provides an easily interpretable effect size estimate in the form of the area under curve (AUC). Two other alternatives also involve a sorted list of genes. The GORilla algorithm [24] essentially calculates the p-value of the best possible partitioning of a list of genes into a foreground and background set. The microarray analysis package, limma, includes the geneSetTest algorithm using a bootstrapping approach [25].

These tests do not account for within-group variability of samples, as they use information distilled in the form of gene sets. However, methods exist to directly estimate variability between samples and directly account for the experimental design in an ANOVA-like manner. GSEA [26] and the program CAMERA from the limma package [25] both work this way. Also, other algorithms (such as roast from

limma) exist which are self contained rather than competitive. That is, instead of testing for enrichment of a function among differentially expressed genes compared to genes without this function, they only analyze whether a significant difference in gene expression in the given gene set exists [23].

Gene ontologies

Gene ontologies assign hierarchical categories to genes based on manual or automatic annotation and literature evaluation. The widely used eponymous gene ontology (GO) database [27] includes three hierarchies of gene categories: biological process (BP), molecular function (MF) and cellular component (CC). The structure of each of these three partitions is hierarchical; each partition starts at a generic root term, and each GO term can be associated with one or more different terms. Furthermore, each gene can be included in a number of GO terms. For example, the transcription factor STAT1, which responds to interferon signaling, is annotated by “signal transduction” and “regulation of transcription, DNA-templated” in BP, “sequence-specific DNA-binding transcription factor activity” and “signal transducer activity” in MF and “nucleus” in CC.

The association between genes and ontology terms is based on various types of evidence, including direct evidence from an experiment or a direct assay, as well as from an expression pattern, sequence similarity or protein structure [28], which allows a better understanding of the observed result and application of higher stringency. However, GO is the result of the accumulated scientific knowledge of genes and their functions; it reflects this knowledge well, which also includes the aforementioned bias intrinsic to that knowledge: few genes have complete and numerous GO annotations, while the majority of genes are associated with generic GO terms or with no GO terms at all. Other predefined data sets include pathways (notably KEGG pathways, [29]), signaling classifications (such as in the MsigDB, [30]) and groups of genes regulated by common transcription

factors. Particularly useful sets of genes that have been based on experimental evidence of gene co-expression will be discussed in the following section.

Transcriptional module analysis

Co-regulated and functionally related genes have similar expression profiles and show correlated changes in expression. In contrast to top-down driven GO, which relies on manual curation and literature research, transcriptional modules are clusters of co-expressed genes derived only from the expression profile data, which therefore may be less biased [31], or at least may not follow the power law trend of GO annotations. Blood transcriptional modules (BTM) analysis has been successfully applied to a number of tasks, including analysis of transcription profiles derived from subjects of vaccine trials [2;32]. Transcriptional modules are derived by clustering of genes that show similar expression patterns, as measured, for example, by the mutual information index [2;32].

Meaningful interpretation of transcriptional modules requires inclusion of a priori knowledge of gene functions. This is frequently done by testing the enrichment of modules in GO categories or other predefined data sets [2;32] and manual curation. Since this can lead to circular reasoning if BTMs are first generated and then applied to the same data set, caution is advised.

Combining multivariate approaches and functional analysis

Multivariate approaches, which reduce dimensionality of the data, can be directly plugged into functional analysis, omitting the step of differential analysis. For instance, in PCA, for each component each variable is assigned a gene. If a PC corresponds to the research question of interest (e.g., change in gene expression of the vaccine arm compared to the placebo arm of the trial), it is then possible to categorize or order the variables by their relative weight in that PC. By knowing associations between variables and functional categories, such as gene expression modules or

metabolic pathways, it is possible to perform an enrichment test on these ordered or categorized gene sets [13].

Figure 3 shows PCA of gene expression in tuberculosis (TB) patients and healthy subjects. TB patients form a distinct cluster that can be distinguished from the healthy controls along the third PC (horizontal axis). However, as the plot has been generated using PCs as coordinates, one wants to understand what functional differences can be attributed to each of the dimension. PCs are linear combinations of the variables (genes), and each component is influenced by some genes more than others. More specifically, each PC for a given sample is a sum of all gene expression values for this sample, with each value multiplied by a constant. These constants are called loadings and are different for each PC and each gene, but common for all samples. Genes that have the highest absolute loading in a given PC have the most influence on this PC.

In other words, we can ask whether genes with particularly large loadings (in terms of absolute values) belong to a particular functional category. On Figure 3, an enrichment analysis of BTMs has been applied to genes sorted by loadings in the third and fourth PCs of a PCA of TB patients and healthy controls. The enriched modules were then visualized using tag clouds generated with the tagcloud R package. PC3, which differentiates TB patients from healthy controls, turns out to be mostly influenced by genes related to the immune response, particularly the IFN response. This finding is in line with results published earlier [3], but did not require a differential analysis step.

Gaucher et al. [33] applied microarray-based transcriptomics analysis to RNA whole blood samples collected in three yellow fever vaccine trials, at baseline (day 0) and at six subsequent time points up to 2 months after vaccination. Using differential expression analysis and PCA, they defined a peak response on days 3 and 7. In order to understand the functional meaning of changes in gene expression, they applied an approach described by Teschendorff et al. [13]. First, they performed an ICA on the data set and for each PC, they selected 5% of the genes with the highest or lowest weights. Then, they tested the selected genes for enrichment in gene sets defined in the database

MsigDB [30]. In this way they could functionally characterize the response to the vaccines and show several distinct arms of the immune response regulated in a coordinated fashion by several master switch genes.

Data integration in molecular biology

Integration of heterogeneous data sets (e.g., combinations of transcription profiles, protein expression and metabolites) is not a trivial task for both statistical and biological reasons. First, data generated on different technical platforms tend to have diametrically different statistical properties. As an extreme case, while normalized transcription data from a microarray have a roughly normal distribution, DNA methylation signals after transformation show a logit distribution, since methylation ranges from 0 (no methylation at the site) to 1 and since fully methylated or non-methylated sites dominate the data. Second, as noted previously, the data sets do not always correspond to different facets of the same system (as it is for transcriptomes and proteomes derived from the same samples), but in fact can originate in different cells or even organs. Third, as data are acquired at various levels of the biological systems, there is a somewhat fuzzy distinction between predictor and response variables; in some cases, a set of variables can be considered to directly influence the other data set more profoundly than vice versa (e.g., transcriptomics and proteomics, or transcriptomics and vital life signs such as body temperature and blood pressure). Finally, it should be noted that the following sections are also applicable for low-dimensional data, such as data from antigen stimulation essays, in particularly for cytokine multiplex readouts.

Integrated data analysis can be roughly segregated into following approaches [34]:

Direct clustering

Relatedness between variables from two data sets can be measured by correlation coefficient or mutual information index. These, in turn, can be converted into distance, which can be used in

distance-based clustering methods. This approach works for closely related data sets, such as microRNA and transcription profiles from the same RNA isolates [3], but fails for other data types [34]. Specifically, variables originating from divergent data sets tend to cluster within the same data set – the data are dominated by variance between the platforms rather than by covariance of functionally related features from different platforms.

Functional analysis

The data sets in question are analyzed separately, and the information from all sets is integrated post hoc by comparing the results from functional analyses. For example, transcriptomic and proteomic data sets are analyzed independently at first, and then integrated using information on known protein–protein interactions and functional relationships, or by simply comparing lists of results. As an advantage, the approach does not involve direct statistical comparisons between incompatible data sets. However, it does not take advantage of the covariance structure of the data – in other words, novel information on how variables from different data sets are statistically dependent is inadvertently lost. This approach has been used by Lu et al. [35], who compared the results of analysis of differentially methylated DNA sites in a hepatitis B virus vaccination trial with the results of gene expression analysis from unrelated data sets, confirming two of their findings.

Multivariate statistical Approaches

Divergent data sets can be standardized using dimension reduction techniques such as PCA, or directly subjected to PLS analysis. Several approaches exist to date. Blanchet et al. [36] applied a two-step process for integrating proteomic and metabolomic data. In a first step, canonical variate analysis (CVA) was applied to the data sets independently to standardize the variables, reduce data set dimensionality and select variables. In a second step, they used a modified PCA to integrate both

data sets and found that the integrated data set performed extraordinarily well in discriminating the experimental groups.

Morine et al. [37] have used CCA to augment functional analysis. They started with two data sets, gene expression and plasma metabolic profiling from genetically obese mice with different diets. Both sets were analyzed with regularized CCA to find genes associated with plasma metabolic markers. Finally, to interpret the results, the genes for which the correlation score was above a threshold were tested for overrepresented pathways.

Modeling Approaches

For certain applications, for example, if the data sets are closely related (i.e., origin from the same source, like a certain cell population) and if additional information exists on the relations between variables from various platforms, sophisticated mathematical modeling of the changes may be possible. This approach requires substantial investment in generating and testing models, but can potentially provide unique insights missed by other methods [38].

Final remarks

While numerous statistical and computational methods allow data integration in a functional, biological context, no method fits all problems. Moreover, while in most cases it is possible to integrate radically different approaches, there is further room for improvement especially in dense multi-platform data which include more than two platforms. Finally, note that most methods do not specifically address the fact that HT data from vaccine trials often represent a time series.

Acknowledgments

The authors thank Mary Louise Grossman for help preparing the manuscript. Work in the authors' lab is supported by the European Union's Seventh Framework Programme (EU FP7) project "ADITEC" (HEALTH-F4-2011-280873; and EU FP8 project "TBVAC 2020" (grant 643381); the Innovative Medicines Initiative (IMI) Joint Undertaking (JU) "Biomarkers for Enhanced Vaccine Safety" project BioVacSafe (IMI JU Grant No. 115308); the Bill & Melinda Gates Foundation (BMGF) "Grand Challenges (GC) in Global Health" Program (BMGF GC6-2013, (#OPP 1055806 and #OPP 1065330); and the *Bundesministerium für Bildung und Forschung* (BMBF) project "Infect Control 2020."

Figures

Figure 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.

Figure 2. Data sets with different dimensionalities. (A) A single measurement (variable) for each sample. (B) A HT data set in which for each sample, several variables are measured. For example, for each RNA sample, the expression of several (or several thousand) genes is measured. Each sample is therefore a vector, and the whole data set is two-dimensional. (C) For each sample, each variable has been measured at four different time points. Such a data set can be regarded as three-dimensional, and each sample is a tensor with the two dimensions corresponding, to different variables (genes) and time points of measurement, respectively.

Figure 3. Components 3 (horizontal axis) and 4 (vertical axis) of a PCA of tuberculosis patients (TB, green triangles) and healthy, non-infected controls (NID, blue dots); data from [39]. The module

432 enrichment has been calculated using Mann-Whitney U test based on gene weights and modules
433 defined by [2]. Abbreviations: adenosine triphosphate, ATP; dendritic cell, DC; interferon, IFN;
434 natural killer, NK; to be announced, TBA; Toll-like receptor, TLR.

435

Reference List

- [1] Kennedy RB, Poland GA. The top five "game changers" in vaccinology: toward rational and directed vaccine development. *OMICS* 2011 Sep;15(9):533-7.
- [2] Li S, Nakaya HI, Kazmin DA, Oh JZ, Pulendran B. Systems biological approaches to measure and understand vaccine immunity in humans. *Semin Immunol* 2013 Oct 31;25(3):209-18.
- [3] Maertzdorf J, Weiner J, III, Kaufmann SH. Enabling biomarkers for tuberculosis control. *Int J Tuberc Lung Dis* 2012 Sep;16(9):1140-8.
- [4] Buonaguro L, Wang E, Tornesello ML, Buonaguro FM, Marincola FM. Systems biology applied to vaccine and immunotherapy development. *BMC Syst Biol* 2011;5:146.
- [5] Reif DM, Motsinger-Reif AA, McKinney BA, Rock MT, Crowe JE, Jr., Moore JH. Integrated analysis of genetic and proteomic data identifies biomarkers associated with adverse events following smallpox vaccination. *Genes Immun* 2009 Mar;10(2):112-9.
- [6] Gomez-Cabrero D, Abugessaisa I, Maier D, et al. Data integration in the era of omics: current and future challenges. *BMC Syst Biol* 2014;8 Suppl 2:11.
- [7] van Iterson M, van de Wiel MA, Boer JM, de Menezes RX. General power and sample size calculations for high-dimensional genomic data. *Stat Appl Genet Mol Biol* 2013 Aug;12(4):449-67.
- [8] Hart SN, Therneau TM, Zhang Y, Poland GA, Kocher JP. Calculating sample size estimates for RNA sequencing data. *J Comput Biol* 2013 Dec;20(12):970-8.
- [9] Ching T, Huang S, Garmire LX. Power analysis and sample size estimation for RNA-Seq differential expression. *RNA* 2014 Nov;20(11):1684-96.
- [10] Guo Y, Zhao S, Li CI, Sheng Q, Shyr Y. RNAseqPS: A Web Tool for Estimating Sample Size and Power for RNAseq Experiment. *Cancer Inform* 2014;13(Suppl 6):1-5.
- [11] Li CI, Su PF, Shyr Y. Sample size calculation based on exact test for assessing differential expression analysis in RNA-seq data. *BMC Bioinformatics* 2013;14:357.
- [12] Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Roy Stat Soc Ser B (Meth)* 1995;57(1):289-300.
- [13] Teschendorff AE, Journee M, Absil PA, Sepulchre R, Caldas C. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS Comput Biol* 2007 Aug;3(8):e161.
- [14] Le Cao KA, Rossouw D, Robert-Granie C, Besse P. A sparse PLS for variable selection when integrating omics data. *Stat Appl Genet Mol Biol* 2008;7(1):Article 35, doi: 10.2202/1544-6115.1390.

- 470 [15] Lu H, Plataniotis KN, Venetsanopoulos AN. Multilinear principal component analysis of tensor
471 objects for recognition. Department of Electrical and Computer Engineering, University of
472 Toronto, Toronto, ON M5S 3G4, Canada. 2006 p. 776-9.
- 473 [16] Lique B, Le Cao KA, Hocini H, Thiebaut R. A novel approach for biomarker selection and the
474 integration of repeated measures experiments from two assays. BMC Bioinformatics
475 2012;13:325.
- 476 [17] Gunther OP, Shin H, Ng RT, et al. Novel multivariate methods for integration of genomics and
477 proteomics data: applications in a kidney transplant rejection study. OMICS 2014
478 Nov;18(11):682-95.
- 479 [18] mixOmics: Omics data integration project [computer program]. Version Version 2-9 2011.
- 480 [19] Huss JW, III, Lindenbaum P, Martone M, et al. The Gene Wiki: community intelligence applied
481 to human gene annotation. Nucleic Acids Res 2010 Jan;38(Database issue):D633-D639.
- 482 [20] Ulitsky I, Laurent LC, Shamir R. Towards computational prediction of microRNA function and
483 activity. Nucleic Acids Res 2010 Aug;38(15):e160.
- 484 [21] Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. Nat Rev
485 Genet 2009 Mar;10(3):155-9.
- 486 [22] Falcon S, Gentleman R. Using GOSTats to test gene lists for GO term association.
487 Bioinformatics 2007 Jan 15;23(2):257-8.
- 488 [23] Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets:
489 methodological issues. Bioinformatics 2007 Apr 15;23(8):980-7.
- 490 [24] Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization
491 of enriched GO terms in ranked gene lists. BMC Bioinformatics 2009;10:48.
- 492 [25] Wu D, Smyth GK. Camera: a competitive gene set test accounting for inter-gene correlation.
493 Nucleic Acids Res 2012 Sep 1;40(17):e133.
- 494 [26] Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-
495 based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A
496 2005 Oct 25;102(43):15545-50.
- 497 [27] Harris MA, Clark J, Ireland A, et al. The Gene Ontology (GO) database and informatics
498 resource. Nucleic Acids Res 2004 Jan 1;32(Database issue):D258-D261.
- 499 [28] Blake JA. Ten quick tips for using the gene ontology. PLoS Comput Biol 2013;9(11):e1003343.
- 500 [29] Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering
501 the genome. Nucleic Acids Res 2004 Jan 1;32(Database issue):D277-D280.
- 502 [30] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular
503 signatures database (MSigDB) 3.0. Bioinformatics 2011 Jun 15;27(12):1739-40.
- 504 [31] Chaussabel D, Baldwin N. Democratizing systems immunology with modular transcriptional
505 repertoire analyses. Nat Rev Immunol 2014 Apr;14(4):271-80.

- [32] Li S, Roupheal N, Duraisingham S, et al. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nat Immunol* 2014 Feb;15(2):195-204.
- [33] Gaucher D, Therrien R, Kettaf N, et al. Yellow fever vaccine induces integrated multilineage and polyfunctional immune responses. *J Exp Med* 2008 Dec 22;205(13):3119-31.
- [34] Haider S, Pal R. Integrated analysis of transcriptomic and proteomic data. *Curr Genomics* 2013 Apr;14(2):91-110.
- [35] Lu Y, Cheng Y, Yan W, Nardini C. Exploring the molecular causes of hepatitis B virus vaccination response: an approach with epigenomic and transcriptomic data. *BMC Med Genomics* 2014;7:12.
- [36] Blanchet L, Smolinska A, Attali A, et al. Fusion of metabolomics and proteomics data for biomarkers discovery: case study on the experimental autoimmune encephalomyelitis. *BMC Bioinformatics* 2011;12:254.
- [37] Morine MJ, McMonagle J, Toomey S, et al. Bi-directional gene set enrichment and canonical correlation analysis identify key diet-sensitive pathways and biomarkers of metabolic syndrome. *BMC Bioinformatics* 2010;11:499.
- [38] Nakaya HI, Li S, Pulendran B. Systems vaccinology: learning to compute the behavior of vaccine induced immunity. *Wiley Interdiscip Rev Syst Biol Med* 2012 Mar;4(2):193-205.
- [39] Maertzdorf J, Ota M, Repsilber D, et al. Functional correlations of pathogenesis-driven gene expression signatures in tuberculosis. *PLoS One* 2011;6(10):e26938.

Figures

TBA
NK cell surface
signature

TBA

enriched in
dendritic cells
TBA
transmembrane
transport (I)

regulation of transcription,
transcription factors

heme
biosynthesis (I)
cell cycle,
ATP binding

TBA

TBA

enriched in
NK cells (II)

20

0

-20

-40

enriched in B cells (I)
activated
dendritic cells
DC surface signature
enriched in monocytes (IV)
TBA

type I interferon response
enriched in neutrophils (I)
TBA

immune activation
- generic cluster

myeloid cell enriched receptors and transporters
enriched in B cells (II)
complement and other receptors in DCs

antiviral IFN signature
TLR and inflammatory signaling

innate antiviral response
Monocyte surface signature
enriched in activated dendritic cells (II)
platelet activation - actin binding
enriched in myeloid cells and monocytes
complement activation (I)
enriched in monocytes (II)