DA SOL

Smarter Studies

DATA ANALYSIS SOLUTIONS
DA-SOL GmbH

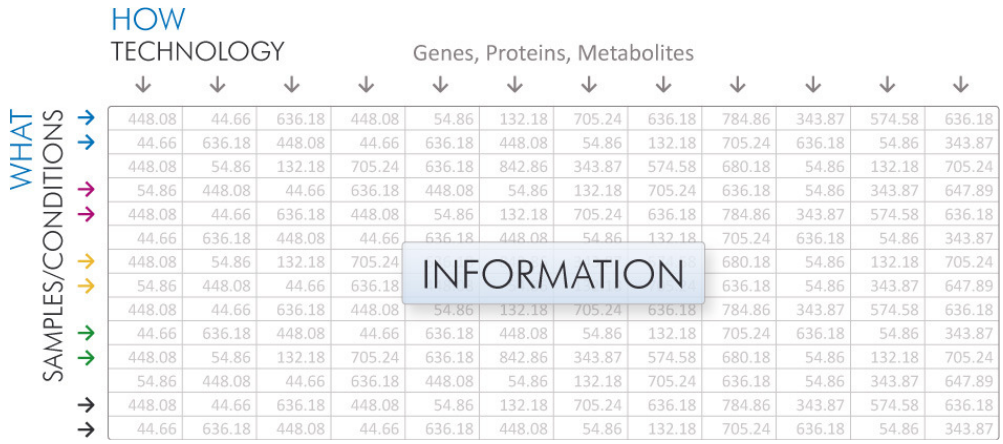
Dr. Juergen von Frese
jvf@da-sol.com

SMARTER STUDIES

The Way to Success: Designing Rigorous Omics Studies

Modern omics measurement technology offers an unprecedented power for unraveling systems biology, developing molecular diagnostics and personalized medicine as well as leveraging drug development. But one crucial aspect for harvesting its enormous power is often neglected: Omics measurements are inherently comparative. - We compare healthy to diseased, responders to non-responders, good versus poor prognosis. The focus is usually on the ***how*** of this comparison (technology), but it matters just as much ***what*** we compare (samples, conditions).

Whereas it is generally very difficult and costly to improve the measurement performance, it is often rather simple to achieve a quantum leap in result quality and validity by an improved study design. Moreover, any upfront mistake in the study design will seriously limit the overall results, but can usually not be corrected anymore after the experiments have been performed.



The information from omics experiments is inherently comparative and actually arises from an interplay of ***how*** we compare (technology) and ***what*** we compare (samples, conditions). Do you place sufficient emphasis on the latter?

Comparing Apples and Oranges

Omics technologies "see everything", so they will pick up all differences between the groups we want to compare - regardless if they reflect the true biology of interest or not. Therefore, some of the biggest dangers which have to be dealt with are skewed comparisons (bias) and additional misleading differences between groups (confounding factors). Both would lead to spurious results and unreliable predictors.

Examples:

- An imbalance between female and male patients in the compared groups might erroneously show sex specific genes (e.g. XIST, RPS4Y) to be differentially expressed.
- A landmark study compared two forms of leukemia, ALL and AML. But as the former is the most common form of leukemia in children, whereas the latter is for adults, they actually compared childhood ALL with adult AML. Thus, for that particular data it becomes impossible to attribute any of the found markers clearly to either the ALL/AML or childhood/adult difference.

Fortune Telling

It might sound trivial - but it is violated very often nevertheless: If data does not contain information about the endpoint (diagnostic question) of interest, nothing valid can be extracted from that data. Bioinformatics and data mining will always return something and thus capitalize on chance differences in this case (i.e. artifacts, bias or confounding factors).

Examples:

- It has been shown that the quality of surgery influences the prognosis for early stage colorectal cancer. Likewise positive margins would indicate residual tumor for a breast cancer patient. No omics classifier is able to predict suboptimal surgery. Thus, the choice of inclusion/exclusion criteria becomes very important (e.g. excluding local recurrences).
- Drug response is governed by both pharmacokinetics and pharmacodynamics. If a drug does not reach a therapeutic dose, no omics technology can correctly predict the response for what has effectively been an untreated sample. A point in case is the breast cancer drug tamoxifen. It is actually a prodrug which has to be metabolized in the liver to the active compounds. Both due to genetic polymorphisms in the metabolizing P450 enzymes as well as due to inhibition from concurrent medication the effective concentration of the active compounds can be decreased significantly.

Small Cause - Big Effects

We are all aware that small faults can render complex systems or machines non-functional. Likewise gross outliers can lead any analysis on an otherwise perfect dataset astray. It is very important to realize that outliers cannot just arise from poor sample or measurement quality, but also because of the biological complexity, e.g. from implicit molecular subgroups in the data, breaking the pattern observed for the majority of samples. Thus, precise inclusion/exclusion criteria and a careful definition of groups can have a big effect.

Examples:

- Tumor biology is complex and for diagnostic and prognostic purposes a larger variety of known clinical subtypes are often subsumed into one group although they might follow a different biological mechanism (e.g. small cell/neuroendocrine lung tumors, microsatellite instable colorectal cancers or comparing prognosis/treatment response across ER or Her2 positive and negative breast cancers).
- Younger cancer patients might include an elevated fraction of early-onset hereditary tumors.

Buckle Up!

Artifacts are an unavoidable hurdle any real-world project has to cope with. If undetected they might lead to spurious findings. Randomization of samples and measurements is your number one safeguard against it. Hopefully it will not come into effect, but if it does, it might make all the difference. - And it basically comes for free. There is no better bargain in risk mitigation to be had!

Example:

- Several omics studies on ovarian cancer were publicly challenged as their results could not be reproduced. Clinical groups of interest were measured one after the other and it could be shown that the results were dominated by batch effects and instrumental artifacts.

Tricks of the Trade

The above are only obvious examples of why a careful study design is a key to successful omics studies. Behind the scenes a lot more statistical and biomedical knowledge is required to address such questions as:

- Precise definition of outcome measures/endpoints
- Sample size planning for model building and validation
- Definition of the required clinical information
- Inclusion/exclusion criteria
- Ensuring representative data
- Balancing
- Randomization
- Replication
- References (e.g. housekeeping genes for qPCR)
- Analytical performance, in particular between laboratories and over time
- Measurement controls and standards
- Design of time course experiments
- Response surface designs (DoE)
- Efficient designs for companion diagnostics
- Evaluation of individual drugs in combination therapies
- Representation of less abundant clinical groups
- Satisfying regulatory requirements

Conclusion

Study design means drafting the future analyses. Study designs means asking questions. - Valid answers will only be obtained for rigorously planned studies.

In some way or the other samples will be selected for your studies. They will invariably determine what can be discovered from the data and how reliable the results will be. - Why not choose the smarter study?