Practical Considerations for Omics Experiments in Biomedical Sciences

Marc Vaudel[1,], Harald Barsnes[1,2], Rolf Bjerkvig[3], Andreas Bikfalvi[4], Frode Selheim[1], Frode S. Berven[1] and Thomas Daubon[3,4]*

[1] Proteomics Unit, Department of Biomedicine, University of Bergen, Bergen, Norway
[2] Department of Clinical Science, University of Bergen, Bergen, Norway
[3] NorLux Neuro-Oncology, Department of Biomedicine, University of Bergen, Bergen, Norway
[4] INSERM U1029, University of Bordeaux, France

[*] Correspondence: Marc Vaudel, Proteomics Unit, Department of Biomedicine, University of Bergen, Norway.
Jones Liesvei 91
N-5009 Bergen
Norway
Email: marc.vaudel@uib.no
Tel: +47 55 58 63 78
Fax: +47 55 58 63 60

Words:
6,445

Abstract

Modern analytical techniques provide an unprecedented insight to biomedical samples, allowing an in depth characterization of cells or body fluids, to the level of genes, transcripts, peptides, proteins, metabolites, or metallic ions. The fine grained picture provided by such approaches holds the promise for a better understanding of complex pathologies, and consequently the personalization of diagnosis, prognosis and treatment procedures. In practice however, technical limitations restrict the resolution of the acquired data, and thus of downstream biomedical inference. As a result, the study of complex diseases like leukemia and other types of cancer is impaired by the high heterogeneity of pathologies as well as patient profiles. In this review, we propose an introduction to the general approach of characterizing samples and inferring biomedical results. We highlight the main limitations of the technique with regards to complex and heterogeneous pathologies, and provide ways to overcome these by improving the ability of experiments in discriminating samples.

Keywords: Omics, Systems biology, personalized medicine, high resolution medicine, biomedical data interpretation

Introduction

The establishment of methods for sequencing entire genomes have opened the era of large scale systemic analyses of biomedical samples[1]. Since the sequences of proteins, the main actors of biological systems, find their origin in genetic information[2], the completion of the Human Genome Project[1a], the "sequencing of the book of life", revolutionized the "diagnosis, prevention and treatment of most, if not all, human diseases" as foreseen by U.S. president Clinton (www.genome.gov/10001356). Genes and their transcripts present the advantage of being replicable, allowing the detection of low abundant signals by amplification, ultimately enabling a comprehensive characterization of genes and transcripts in a sample, respectively termed genomics and transcriptomics. Modern sequencing techniques hence make it affordable to sequence entire genomes and transcriptomes within a reasonable timeframe[3].

However, cellular processes involve the post-translationally processed, mature form of proteins, and complex interactions with internal and external compounds, e.g., ions, metabolites (including sugars and lipids). As a consequence, genetic, transcriptional information is not sufficient to accurately describe disease mechanisms[4]. In order to characterize samples in a more comprehensive manner, other *omics* approaches have been established, as described Figure 1A: proteomics for the study of proteins in their different forms and modification statuses, metabolomics for metabolites, metallomics for metals and their isotopic distributions. As a result, researchers now benefit from a vast panel of analytical techniques allowing the characterization of biomedical samples at increasing depth and level of detail. The systemic analysis of these samples, may provide more comprehensive and precise results, in turn allowing a more accurate inference of underlying biological processes, and may thus provide better understanding of the various pathologies.

This increased performance in biomedical sample characterization holds the promise for a better understanding of complex pathologies. Indeed, many cancers types show an extensive intratumor as well as an intertumor heterogeneity which is reflected at the molecular and clinical level. In particular, omics experiments represent important tools for the classification of patients into risk stratification groups, or tailoring the treatment according to the patient pathological profile[5]. In this context, a direct benefit of an increased coverage and higher precision of omics techniques is a higher resolution when attempting to discriminate patients, and thus move towards a more personalized medicine. This twenty-first century medicine paradigm is notably known through US President Obama's "Precision Medicine Initiative", aiming at accounting for "individual variability in genes, environment, and lifestyle for each person" (www.nih.gov/precisionmedicine).

In this review, we highlight different factors limiting the resolution power of omics approaches in biomedical sciences. We describe potential solutions available to critically evaluate and interpret the vast datasets produced by such techniques. Finally, we suggest ways to improve the resolution of biomedical experiments and identify the limiting factors impairing high resolution medicine.

Limitations in Omics Approaches

Using systemic characterization of samples, biomedical researchers aim at identifying the mechanism underlying the pathologic state of an organism. This characterization can be done in a discovery mode, by attempting to characterize samples as comprehensively as possible, or in a targeted fashion, by accurately monitoring putative compounds. The performance of the analysis will be evaluated in terms of sensitivity, *i.e.* the ability to comprehensively characterize the sample, and in terms of specificity, *i.e.* the propensity to produce correct results. Generally, discovery and targeted studies respectively prioritize sensitivity or specificity.

The sensitivity is limited by the initial sample amount in relationship to the detection limit of the workflow, while to a lower extend for analytical approaches benefiting from signal amplification techniques like the polymerase chain reaction (PCR)[6]. This is notably the case when characterizing low abundant biological compounds like post-translationally modified proteins[7], where a complete profiling requires large amounts of starting material. Obtaining such amounts can be challenging in time and financially when culturing cells, and especially when working with cell sub-populations; it is often impossible for limited patient material. [6]

Ideally, omics techniques should profile cells individually[8], providing subcellular localization of the identified compounds, and combine the results in a systemic overview. However, due to sensitivity limits in the analytical technique, populations of cells are generally studied together, and, despite being of paramount biological importance, little or no information is available on the localization of the compounds in the cells. As a result, the characterization is an average of distinct sub-populations, hence limiting the resolution of the biomedical study. This is notably a critical aspect when studying samples obtained from cell populations with heterogeneous cytogenetic profiles, as encountered in cancer samples, or with heterogeneous cell functions, as encountered in targeted differentiation of induced pluripotent stem cells. However, it is sometimes possible to separate the cells prior to analysis. The efficiency of the categorization is then a crucial parameter, and implies working with less material, which might in turn impair the detection of low abundant compounds as discussed above. The improvement of workflows and instrumentation, allowing to work with minute amounts of starting material, is a promising way toward characterization of heterogeneous samples[9]. Similarly methods have been developed to assess the localization of proteins in cell sub-compartments in proteomic analyses, paving the

way towards spatially resolved proteome characterization[10]. On a tissue scale, MALDI imaging is certainly the most promising technique for spatial identification and quantification of metabolites and proteins, including post-translational modifications[11]. Spatially resolved omics are however still at their infancy, and current experiments generally couple omics studies with techniques of high spatial resolution, but of much less coverage, typically using antibodies[12].

With the problem of sample heterogeneity, comes often problems of sample purity and preparation. By design, sample extraction introduces a bias in the observation of biological systems, and it is important to verify potential consequences of sample handling on downstream analysis, and importantly their compatibility with the analytical method used. This is particularly the case for cell culture techniques involving the use of feeder cells or media which might interfere with the systems biology characterization. The question of sample purity is crucial to avoid biological signals from sample preparation, which might increase variability or interfere with the sample compounds. The same considerations apply to contaminants, including viral contaminations[13], multiple species samples[14], and contamination by the experimentalist when preparing the sample as commonly seen with the presence of foreign compounds and famously illustrated by the case of the Phantom of Heilbronn. The sample uptake will also influence downstream analysis, notably at the post-translational level, depending on the inhibition of immune system and degradation processes.

In order to account for the loss of specificity introduced during sample preparation or the analytical workflow, and whenever conclusions must be drawn on populations, biomedical results must be replicated and evaluated statistically. For this, observed differences are compared to the variability of populations. As illustrated Figure 1B, the most encountered statistical evaluation in biomedical sciences is the test of a significant difference between two conditions, e.g. sick and healthy, using a Student's t-test[15] and modelling population observations, e.g. patient and control observations, using normal distributions. The significance of a regulation will then depend on the populations' variability and on the number of samples considered[16]. When designing an experiment, the number of samples is thus a key parameter to consider. As detailed Figure 1C, it is possible to link the smallest significant difference to other parameters of the model used to evaluate the significance of the results, and thus, in the case of normal distributions and a Student's t-test, to the population variability and the number of samples. Interestingly, the smallest significant difference relative to the population variability can be used to illustrate the resolution of the experiment. As illustrated Figure 1D, it is hence possible to directly link the maximal resolution of the experiment to the number of samples, and thus verify that it allows detecting the expected changes. When working with limited patient materials, it is recommended to estimate proactively the resolution of the experiment, and consider the relevance of the approach.

Processing Experimental Data

The first step in evaluating scientific data is to conduct a visual inspection. This task may be challenging due to the size of the dataset, yet it is crucial to assess the parameters submitted to the statistical analysis, for example, by simply plotting data in a scatter plot, or visualizing the standard deviation against the average. This upstream visual quality control of the data allows avoiding numerous issues, and helps the scientist familiarizing with the dataset. By such an approach, it may be possible to detect trends in the data, and thus adapt the downstream data interpretation procedure. As illustrated by Anscombe's quartet[17] Figure 2, various datasets and their statistical characteristics can be interpreted as noise, and can remain undetected unless specifically looked for. A simple visualization of the data will allow detecting these and improve the data interpretation strategy.

Biomedical studies often focus on fold change variations of compounds between populations, where one is used as a reference in order to estimate a ratio. Here it is important to apply a logarithmic transformation of the ratio before further processing. The main objective is to move the observations to a space where standard mathematical models can be applied. For example, ratios distribute around 1, with all down-regulations in the range between 0 and 1, and all up-regulations in the range between 1 and infinity. As a result, the variability of down regulated compounds will not be comparable to the variability of up regulated compounds. After logarithmic transformation most standard operators will be applicable to ratios. In general, it is vital to verify that the observations abide by the hypotheses of the descriptive model used, and otherwise transform them, e.g. if the model assumes normally distributed values, makes sure to confirm that the values are normally distributed before applying the model.

The first step in data processing often consists in normalizing the results in order to correct for systematic errors induced by the experimental workflow. As detailed in the previous section, the minimal significant change depends on the variance in the population. An example is illustrated in Figure 3A**,** where proteomic data obtained from the cerebrospinal fluid expression of proteins unaffected by blood flow[18] among different patients, courtesy of Dr. J.A. Opsahl, the normalization of the results allows reducing the variability between patients, and thus allows to more reliably detect low abundant changes. Different normalization procedures exist, based on statistical estimators like the mode, median, mean, or the sum of observed values. In such strategies, the normalization corrects according to a hypothesis about the sample content. For example, the median assumes that the majority of observations have same abundance between samples, while the sum assumes that the global abundance of compounds is the same between samples. For the example shown in Figure 3A, the protein expression levels were normalized according to the mean and median across patients, the latter providing the lowest variability.

Alternatively, when population based estimators are inaccurate, it is possible to normalize on sub-populations of compounds known to be stable, or to use internal standards. It should be emphasized that while this step allows identifying systemic perturbations of low abundance compounds, the transfer of such results to the clinics can only be achieved if the normalization procedures are applicable. This is notably challenging when moving from discovery to targeted strategies.

As detailed in the previous section, the analytical methods used in omics experiments all present sensitivity limits, and some compounds will thus show missing values. Missing values represent a challenge in the downstream processing of the data interpretation. A common approach is to assume that these compounds are present in low quantity, below the detection limit of the analytical characterization. They are thus assigned an arbitrarily low value, referred to as imputation. In order not to skew the distribution of the actual observations, it is possible to infer randomized values distributed in the low range of the observed values, as illustrated Figure 3B with imputation from the Perseus software in the MaxQuant suite[19]. Another approach consists in assigning the missing values an arbitrary non regulated ratio, to ensure that the outcome of the study will not be based on imputated values. After missing values imputation, the same number of numerical values will be available for every patient, allowing the conduction of standard downstream analyses, e.g. statistical tests or clustering. Ultimately, it is important to verify the influence of imputated values on the final list of regulated compounds, and eventually conduct follow-up experiments to verify the candidates selected with a strong prevalence of missing values.

Statistical Evaluation

After initial processing, the data can be subjected to statistical evaluation, where the validity of the scientific hypothesis is evaluated in light of the experimental data and translated into a probability that such or more extreme results could be observed by chance, i.e. a p-value. While the p-value was originally conceived as a way to select the most promising results, it has evolved towards a self-sufficient validation procedure in biomedical sciences[20]. However, it is not because a result is not statistically significant that it is wrong – the test might simply fail due to the low resolution of the experiment – and it is not because a result is statistically significant that it is correct. On the contrary, by design, the results are validated with an arbitrary minimal probability of being wrong. In biomedical studies, due to the often low amount of samples, systematic errors, or inaccurate modelling, this probability is moreover generally under evaluated[21]. When producing multiple results at a given error probability, the chances of having a false result accumulate and the overall probability of the result set to contain errors increases. This effect has a strong prevalence in omics studies where large numbers of compounds are

characterized and statistically tested. This phenomenon is described as the *multiple comparisons* or *multiple testing* problem, and can be corrected for, generally by increasing the stringency of the threshold until reaching a desired false discovery rate (FDR)[22]. The rationale behind multiple hypothesis correction is that incorrect results occur randomly and thus distribute evenly. By selecting the most extreme results, one thus avoids the prevalence of false positives. As illustrated Figure 3C, in practice, multiple hypothesis correction can be achieved by creating volcano plots, where the results related to significance is plotted against its magnitude, and by selecting the points the furthest from the origin. This can be achieved by calibrating the p-values[23], or by thresholding the candidates based on their initial p-value and fold-change until reaching a desired FDR[24].

While statistical evaluation has become a *sine qua non* condition for publication, it is crucial to underline that the role of this procedure is to select the most confident results in a given context, and not to provide a universal validation of the results. One of the dangers of emphasizing the importance of statistical evaluation is the modification of the context of the data, voluntary or involuntary, in order to pass a given p-value threshold; a problem termed *data dredging* or *p-hacking*[25]. Another prejudicial consequence is the assumption that results passing the statistical evaluation do not require additional validation. Due to inadequateness between the model and the observations, dependence between observations, or systematic errors, false results can artificially reach extreme p-values, but be irreproducible using other techniques. This is notably the case for observations with skewed distributions, or incorrectly inferred missing values. To tackle these issues, it is recommended to validate and reproduce the results, preferably with an orthogonal analytical technique, preferably in a targeted manner in order to gain precision.

Finally, statistically significant differences between samples do not necessarily imply separation of populations. The performance of the separation is generally evaluated using a Receiver Operating Characteristic (ROC), as illustrated Figure 3D. In this way the specificity and sensitivity can be visually evaluated, and the separation efficiency is measured by taking the Area Under the Curve (AUC), or the value of the sensitivity at a given specificity threshold. The robustness of this separation can be evaluated by studying the variability of the retained efficiency metric when sequentially removing one patient after the other, the so-called *leave one out* approach.

Strategies to Improve Resolution

Several techniques have been developed to increase the discrimination power of omics techniques, often by improving experimental designs. The first consists in reducing the variability by conducting paired studies. Because patient intra variability is generally lower than inter variability, differences relevant to a change in conditions, e.g. introduced by a treatment,

are less likely to be masked by inter sample variability. Such designs however presume the availability of samples prior to conditional change, a prerequisite not met for patient entering research studies after diagnosis. In such cases it may be possible to find family controls, as for monogenic diseases[26], allowing the design of comparative studies with reduced variability between patients.

The second possibility for increasing the resolution of omics experiments is to increase the comprehensiveness of sample characterization. Enhancing sample coverage increases the chances of identifying discriminating biological parameters between patients, hence allowing their distinction. By this approach, additional knowledge needs to be acquired regarding the sample, for example by combining different analytical techniques that provide extended information on different biological species. As an example, transcriptomic data can be completed by post-translational analyses allowing the differentiation of patients both at the genetic and post-translational level[27]. Modern studies hence propose so-called *multi-omics* datasets, combining genomics, transcriptomics, proteomics, metabolomics, *etc*[28].

A third solution to increase the experimental resolution is to tune the experimental design to reduce the prevalence of false positives, for example introduced by multiple hypothesis testing. The most common approach in this context is the use of time series. As illustrated Figure 4A, when comparing the status of a patient in triplicates before and after treatment, the chances of getting an up regulation randomly is substantially higher than the chances of getting a series of six time points in increasing abundance. As illustrated with the red bands Figure 4A, the chances of having six random values distributed around 0 with a standard deviation of 1, in a band of $y=0.2x\pm0.5$ or $y=0.5x\pm0.5$, is 0.1% and 0.006%, respectively. In comparison, as illustrated with the blue squares, the chances of having three replicates of similar magnitude (over 0.8 or 2) randomly occurring is much higher (1% or 0.01%, respectively). Given that random measurements take their values erratically, the chances of getting a specific time series profile are thus lower. This can be intuitively understood by the fact that achieving a specific pattern in a given order is less likely to be achieved by erratic events than a specific value for three (unordered) replicates. As illustrated Figure 4B, clustering methods thus allow to efficiently extract patterns from the biological and technical noise.

Different dimensions can be used to analyse systems biology data: it is possible to monitor systems dynamics in time , on spatial evolutions, e.g. following biological fluid flow[18], thermal profiling where the sample is heated and characterized at different temperatures, e.g. when monitoring interactions and complexes[29]. In these approaches, the resolution of the analytical procedure on the studied dimension, e.g. the sampling frequency for time series, is a crucial factor. It is necessary that the sampling time is much lower than the time scale of the

biochemical event to monitor. This makes it particularly difficult to monitor metabolite-protein interactions, which occur at very low time scales, and techniques of higher time resolution must be used complementarily to omics techniques.

The monitoring of systems dynamics over time is a key feature in the description of cellular processes. While analytically challenging, such experiments provide a view on the cellular response over time and thus a better understanding of the mechanisms involved. Such approaches were for example applied to study the dynamics of the circadian cycle to the metabolic, transcriptomic and proteomic level[28], or to monitor the phosphorylation events involved during platelet activation[30]. The mathematical and computational modelling of such dynamic events can ultimately allow simulating the effect of perturbations on the biological system, and predict its evolution.

As previously discussed, in order to reduce the incidence of multiple hypothesis problems, it is possible to retain only the most significant and highest regulated components until reaching the desired false discovery rate. Another approach relies on the rationale that false positives will appear in an uncoordinated manner, while true positives are likely to represent specific functions of interest. As illustrated Figure 4C, by mapping results on known biochemical pathways, it is thus possible to discriminate the relevant compounds from the others. While this method is extremely promising for systems biology based biomedical studies, it suffers from the lack of maturity of pathway databases, and from low accuracy when matching experimental and theoretical data, notably at the post-translational level[31].

Different bioinformatic tools allow the conduction of these data interpretation steps. However, they often require advanced computational skills. The software Perseus, from the MaxQuant suite[19], is a good alternative for users lacking advanced computational skills. It allows processing multi-omics datasets seamlessly, implementing most of the data interpretation techniques of the field, and support the implementation of additional plugins. Additionally, its workflow based design allows navigating the individual steps of the analysis and visually inspecting intermediate results. Multiple resources are available for functional and pathway analyses, computing a false discovery rate or p-value per pathway[32], as illustrated Figure 4D.

Conclusion

As outlined in this review, one of the characteristics of omics approaches in biomedical sciences is the complexity and size of modern datasets. As a consequence, browsing and efficiently mining published results requires new ways of accessing and visualizing data[33], and associated training material[34]. Modern system wide studies thus come with advanced data sharing and display solutions allowing scientists to fully benefit from the results[35]. The dependence on the

data interpretation also poses new questions concerning the review and reproducibility of results. The public availability of the raw data and of their interpretation procedures have now become a requirement for many journals, and specialized repositories are available allowing the storage and mining of the knowledge produced globally by the scientific community[32b, 36].

The tremendous advances of analytical techniques for the characterization of biological samples have made it possible for researchers to inspect samples to an unprecedented level of detail. The best established techniques, genomics, transcriptomics, and, to a certain extent, metabolomics and proteomics, are now affordable at reasonable costs from core facilities and companies. Additionally, the characterization of more complex biological compounds, like proteins carrying complex modifications or mutated proteins, is a rapidly evolving field[37], and will soon be accessible to all[38]. However, while our ability to analyse samples in detail increases, the functional interpretation of the large datasets produced remains challenging due to the lack of knowledge on how biological compounds interact. As a result, while modern analytical techniques allow ever improving characterizing of samples, it is not always possible to make sense out of the data. Systems biology approaches based on pathway and network analyses aim at solving this need. It is for example already possible to compare ones results against so-called connectivity maps generated by characterizing the reaction of cells to known drugs[39]. There is no doubt that such approaches will greatly benefit from the evolution towards spatial, dynamic, multi-omics analyses.

The standardization in biomedical sample characterization simplifies the possibility to set up international across-lab studies, as illustrated by the setup of ambitious projects like the 100,000 genomes project (http://www.genomicsengland.co.uk). It also opens the possibility for data reprocessing and crowd sourcing studies, where publicly available data is reprocessed and repurposed to answer new scientific questions[40]. Finally, recent developments have moved towards setting up big data strategies for the reprocessing of data produced globally, providing insight on biomedical sciences at large[35a, 41]. In the future, it is thus realistic to envision the use of *in silico* controls to compare the results of a disease study against large amounts of data from the healthy population, or patients affected by other pathologies. The field is already witnessing the move towards *in silico* experiments for the prediction of drug effects[42], and it can be anticipated that the design of *in silico* omics experiments will play a major role in the 21st century biomedical sciences, dramatically reducing the need for animal testing, reducing the experimental costs, and increasing the ability of the field to rapidly tackle new issues.

Acknowledgements

The authors acknowledge Dr. J.A. Opsahl for the cerebrospinal fluid data used for illustrative purposes in Figure 3A, H. Vethe and Prof. Dr. H. Ræder for the stem cell data used for illustrative purposes in Figure 3C.
FS and FSB acknowledge the support of the Norwegian Cancer Society.

References

1. (a) International Human Genome Sequencing, C., Finishing the euchromatic sequence of the human genome. *Nature* **2004,** *431* (7011), 931-45; (b) Mouse Genome Sequencing, C.; Waterston, R. H.; Lindblad-Toh, K.; Birney, E.; Rogers, J.; Abril, J. F.; Agarwal, P.; Agarwala, R.; Ainscough, R.; Alexandersson, M.; An, P.; Antonarakis, S. E.; Attwood, J.; Baertsch, R.; Bailey, J.; Barlow, K.; Beck, S.; Berry, E.; Birren, B.; Bloom, T.; Bork, P.; Botcherby, M.; Bray, N.; Brent, M. R.; Brown, D. G.; Brown, S. D.; Bult, C.; Burton, J.; Butler, J.; Campbell, R. D.; Carninci, P.; Cawley, S.; Chiaromonte, F.; Chinwalla, A. T.; Church, D. M.; Clamp, M.; Clee, C.; Collins, F. S.; Cook, L. L.; Copley, R. R.; Coulson, A.; Couronne, O.; Cuff, J.; Curwen, V.; Cutts, T.; Daly, M.; David, R.; Davies, J.; Delehaunty, K. D.; Deri, J.; Dermitzakis, E. T.; Dewey, C.; Dickens, N. J.; Diekhans, M.; Dodge, S.; Dubchak, I.; Dunn, D. M.; Eddy, S. R.; Elnitski, L.; Emes, R. D.; Eswara, P.; Eyras, E.; Felsenfeld, A.; Fewell, G. A.; Flicek, P.; Foley, K.; Frankel, W. N.; Fulton, L. A.; Fulton, R. S.; Furey, T. S.; Gage, D.; Gibbs, R. A.; Glusman, G.; Gnerre, S.; Goldman, N.; Goodstadt, L.; Grafham, D.; Graves, T. A.; Green, E. D.; Gregory, S.; Guigo, R.; Guyer, M.; Hardison, R. C.; Haussler, D.; Hayashizaki, Y.; Hillier, L. W.; Hinrichs, A.; Hlavina, W.; Holzer, T.; Hsu, F.; Hua, A.; Hubbard, T.; Hunt, A.; Jackson, I.; Jaffe, D. B.; Johnson, L. S.; Jones, M.; Jones, T. A.; Joy, A.; Kamal, M.; Karlsson, E. K.; Karolchik, D.; Kasprzyk, A.; Kawai, J.; Keibler, E.; Kells, C.; Kent, W. J.; Kirby, A.; Kolbe, D. L.; Korf, I.; Kucherlapati, R. S.; Kulbokas, E. J.; Kulp, D.; Landers, T.; Leger, J. P.; Leonard, S.; Letunic, I.; Levine, R.; Li, J.; Li, M.; Lloyd, C.; Lucas, S.; Ma, B.; Maglott, D. R.; Mardis, E. R.; Matthews, L.; Mauceli, E.; Mayer, J. H.; McCarthy, M.; McCombie, W. R.; McLaren, S.; McLay, K.; McPherson, J. D.; Meldrim, J.; Meredith, B.; Mesirov, J. P.; Miller, W.; Miner, T. L.; Mongin, E.; Montgomery, K. T.; Morgan, M.; Mott, R.; Mullikin, J. C.; Muzny, D. M.; Nash, W. E.; Nelson, J. O.; Nhan, M. N.; Nicol, R.; Ning, Z.; Nusbaum, C.; O'Connor, M. J.; Okazaki, Y.; Oliver, K.; Overton-Larty, E.; Pachter, L.; Parra, G.; Pepin, K. H.; Peterson, J.; Pevzner, P.; Plumb, R.; Pohl, C. S.; Poliakov, A.; Ponce, T. C.; Ponting, C. P.; Potter, S.; Quail, M.; Reymond, A.; Roe, B. A.; Roskin, K. M.; Rubin, E. M.; Rust, A. G.; Santos, R.; Sapojnikov, V.; Schultz, B.; Schultz, J.; Schwartz, M. S.; Schwartz, S.; Scott, C.; Seaman, S.; Searle, S.; Sharpe, T.; Sheridan, A.; Shownkeen, R.; Sims, S.; Singer, J. B.; Slater, G.; Smit, A.; Smith, D. R.; Spencer, B.; Stabenau, A.; Stange-Thomann, N.; Sugnet, C.; Suyama, M.; Tesler, G.; Thompson, J.; Torrents, D.; Trevaskis, E.; Tromp, J.; Ucla, C.; Ureta-Vidal, A.; Vinson, J. P.; Von Niederhausern, A. C.; Wade, C. M.; Wall, M.; Weber, R. J.; Weiss, R. B.; Wendl, M. C.; West, A. P.; Wetterstrand, K.; Wheeler, R.; Whelan, S.; Wierzbowski, J.; Willey, D.; Williams, S.; Wilson, R. K.; Winter, E.; Worley, K. C.; Wyman, D.; Yang, S.; Yang, S. P.; Zdobnov, E. M.; Zody, M. C.; Lander, E. S., Initial sequencing and comparative analysis of the mouse genome. *Nature* **2002,** *420* (6915), 520-62.
2. Crick, F., Central dogma of molecular biology. *Nature* **1970,** *227* (5258), 561-3.
3. Pettersson, E.; Lundeberg, J.; Ahmadian, A., Generations of sequencing technologies. *Genomics* **2009,** *93* (2), 105-11.
4. Smith, L. M.; Kelleher, N. L.; Consortium for Top Down, P., Proteoform: a single term describing protein complexity. *Nat Methods* **2013,** *10* (3), 186-7.
5. (a) Estey, E. H., Acute myeloid leukemia: 2013 update on risk-stratification and management. *Am J Hematol* **2013,** *88* (4), 318-27; (b) Dohner, H.; Estey, E. H.; Amadori, S.; Appelbaum, F. R.; Buchner, T.; Burnett, A. K.; Dombret, H.; Fenaux, P.; Grimwade, D.; Larson, R. A.; Lo-Coco, F.; Naoe, T.; Niederwieser, D.; Ossenkoppele, G. J.; Sanz, M. A.; Sierra, J.; Tallman, M. S.; Lowenberg, B.; Bloomfield, C. D.; European, L., Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood* **2010,** *115* (3), 453-74; (c) Kumar, C. C., Genetic abnormalities and challenges in the treatment of acute myeloid leukemia. *Genes Cancer* **2011,** *2* (2), 95-107.
6. (a) Shintaku, H.; Nishikii, H.; Marshall, L. A.; Kotera, H.; Santiago, J. G., On-chip separation and analysis of RNA and DNA from single cells. *Analytical chemistry* **2014,** *86* (4), 1953-7; (b) Nawy, T., Single-cell sequencing. *Nat Methods* **2014,** *11* (1), 18.
7. (a) Loroch, S.; Dickhut, C.; Zahedi, R. P.; Sickmann, A., Phosphoproteomics--more than meets the eye. *Electrophoresis* **2013,** *34* (11), 1483-92; (b) Olsen, J. V.; Mann, M., Status of large-scale analysis of post-translational modifications by mass spectrometry. *Molecular & cellular proteomics :*

MCP **2013,** *12* (12), 3444-52; (c) Solari, F. A.; Dell'Aica, M.; Sickmann, A.; Zahedi, R. P., Why phosphoproteomics is still a challenge. *Molecular bioSystems* **2015**.
8. Method of the Year 2013. *Nat Meth* **2014,** *11* (1), 1-1.
9. Dickhut, C.; Radau, S.; Zahedi, R. P., Fast, efficient, and quality-controlled phosphopeptide enrichment from minute sample amounts using titanium dioxide. *Methods in molecular biology* **2014,** *1156*, 417-30.
10. Gatto, L.; Breckels, L. M.; Burger, T.; Nightingale, D. J.; Groen, A. J.; Campbell, C.; Nikolovski, N.; Mulvey, C. M.; Christoforou, A.; Ferro, M.; Lilley, K. S., A foundation for reliable spatial proteomics data analysis. *Molecular & cellular proteomics : MCP* **2014,** *13* (8), 1937-52.
11. McDonnell, L. A.; Heeren, R. M., Imaging mass spectrometry. *Mass spectrometry reviews* **2007,** *26* (4), 606-43.
12. Marchant, D. J.; Bellac, C. L.; Moraes, T. J.; Wadsworth, S. J.; Dufour, A.; Butler, G. S.; Bilawchuk, L. M.; Hendry, R. G.; Robertson, A. G.; Cheung, C. T.; Ng, J.; Ang, L.; Luo, Z.; Heilbron, K.; Norris, M. J.; Duan, W.; Bucyk, T.; Karpov, A.; Devel, L.; Georgiadis, D.; Hegele, R. G.; Luo, H.; Granville, D. J.; Dive, V.; McManus, B. M.; Overall, C. M., A new transcriptional role for matrix metalloproteinase-12 in antiviral immunity. *Nat Med* **2014,** *20* (5), 493-502.
13. Chernobrovkin, A. L.; Zubarev, R. A., Detection of viral proteins in human cells lines by xeno-proteomics: elimination of the last valid excuse for not testing every cellular proteome dataset for viral proteins. *PloS one* **2014,** *9* (3), e91433.
14. Knudsen, G. M.; Chalkley, R. J., The effect of using an inappropriate protein database for proteomic data analysis. *PloS one* **2011,** *6* (6), e20873.
15. STUDENT, THE PROBABLE ERROR OF A MEAN. *Biometrika* **1908,** *6* (1), 1-25.
16. STUDENT, PROBABLE ERROR OF A CORRELATION COEFFICIENT. *Biometrika* **1908,** *6* (2-3), 302-310.
17. Anscombe, F. J., Graphs in Statistical Analysis. *The American Statistician* **1973,** *27* (1), 17-21.
18. Aasebo, E.; Opsahl, J. A.; Bjorlykke, Y.; Myhr, K. M.; Kroksveen, A. C.; Berven, F. S., Effects of blood contamination and the rostro-caudal gradient on the human cerebrospinal fluid proteome. *PloS one* **2014,** *9* (3), e90429.
19. Cox, J.; Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* **2008,** *26* (12), 1367-72.
20. Nuzzo, R., Scientific method: statistical errors. *Nature* **2014,** *506* (7487), 150-2.
21. Vaudel, M.; Sickmann, A.; Martens, L., Introduction to opportunities and pitfalls in functional mass spectrometry based proteomics. *Biochimica et biophysica acta* **2014,** *1844* (1 Pt A), 12-20.
22. Benjamini, Y.; Hochberg, Y., Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **1995,** *57* (1), 289-300.
23. Schweder, T.; Spjotvoll, E., Plots of P-Values to Evaluate Many Tests Simultaneously. *Biometrika* **1982,** *69* (3), 493-502.
24. Tusher, V. G.; Tibshirani, R.; Chu, G., Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **2001,** *98* (9), 5116-21.
25. (a) Simmons, J. P.; Nelson, L. D.; Simonsohn, U., False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* **2011,** *22* (11), 1359-66; (b) Smith, G. D.; Ebrahim, S., Data dredging, bias, or confounding. *BMJ* **2002,** *325* (7378), 1437-8; (c) Gadbury, G. L.; Allison, D. B., Inappropriate fiddling with statistical analyses to obtain a desirable p-value: tests to detect its presence in published literature. *PloS one* **2012,** *7* (10), e46363.
26. (a) Raeder, H.; Johansson, S.; Holm, P. I.; Haldorsen, I. S.; Mas, E.; Sbarra, V.; Nermoen, I.; Eide, S. A.; Grevle, L.; Bjorkhaug, L.; Sagen, J. V.; Aksnes, L.; Sovik, O.; Lombardo, D.; Molven, A.; Njolstad, P. R., Mutations in the CEL VNTR cause a syndrome of diabetes and pancreatic exocrine dysfunction. *Nature genetics* **2006,** *38* (1), 54-62; (b) Raeder, H.; McAllister, F. E.; Tjora, E.; Bhatt, S.; Haldorsen, I.; Hu, J.; Willems, S. M.; Vesterhus, M.; El Ouaamari, A.; Liu, M.; Raeder, M. B.; Immervoll,

H.; Hoem, D.; Dimcevski, G.; Njolstad, P. R.; Molven, A.; Gygi, S. P.; Kulkarni, R. N., Carboxyl-ester lipase maturity-onset diabetes of the young is associated with development of pancreatic cysts and upregulated MAPK signaling in secretin-stimulated duodenal fluid. *Diabetes* **2014,** *63* (1), 259-69.
27. Cox, J.; Mann, M., 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC bioinformatics* **2012,** *13 Suppl 16*, S12.
28. Robles, M. S.; Cox, J.; Mann, M., In-vivo quantitative proteomics reveals a key contribution of post-transcriptional mechanisms to the circadian regulation of liver metabolism. *PLoS Genet* **2014,** *10* (1), e1004047.
29. Savitski, M. M.; Reinhard, F. B.; Franken, H.; Werner, T.; Savitski, M. F.; Eberhard, D.; Martinez Molina, D.; Jafari, R.; Dovega, R. B.; Klaeger, S.; Kuster, B.; Nordlund, P.; Bantscheff, M.; Drewes, G., Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science* **2014,** *346* (6205), 1255784.
30. Beck, F.; Geiger, J.; Gambaryan, S.; Veit, J.; Vaudel, M.; Nollau, P.; Kohlbacher, O.; Martens, L.; Walter, U.; Sickmann, A.; Zahedi, R. P., Time-resolved characterization of cAMP/PKA-dependent signaling reveals that platelet inhibition is a concerted process involving multiple signaling pathways. *Blood* **2014,** *123* (5), e1-e10.
31. (a) Khatri, P.; Sirota, M.; Butte, A. J., Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology* **2012,** *8* (2), e1002375; (b) Muller, T.; Schrotter, A.; Loosse, C.; Helling, S.; Stephan, C.; Ahrens, M.; Uszkoreit, J.; Eisenacher, M.; Meyer, H. E.; Marcus, K., Sense and nonsense of pathway analysis software in proteomics. *Journal of proteome research* **2011,** *10* (12), 5398-408; (c) Takami, H.; Taniguchi, T.; Moriya, Y.; Kuwahara, T.; Kanehisa, M.; Goto, S., Evaluation method for the potential functionome harbored in the genome and metagenome. *BMC genomics* **2012,** *13*, 699.
32. (a) Croft, D.; O'Kelly, G.; Wu, G.; Haw, R.; Gillespie, M.; Matthews, L.; Caudy, M.; Garapati, P.; Gopinath, G.; Jassal, B.; Jupe, S.; Kalatskaya, I.; Mahajan, S.; May, B.; Ndegwa, N.; Schmidt, E.; Shamovsky, V.; Yung, C.; Birney, E.; Hermjakob, H.; D'Eustachio, P.; Stein, L., Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research* **2011,** *39* (Database issue), D691-7; (b) Ogata, H.; Goto, S.; Sato, K.; Fujibuchi, W.; Bono, H.; Kanehisa, M., KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research* **1999,** *27* (1), 29-34; (c) Pico, A. R.; Kelder, T.; van Iersel, M. P.; Hanspers, K.; Conklin, B. R.; Evelo, C., WikiPathways: pathway editing for the people. *PLoS biology* **2008,** *6* (7), e184; (d) Vizcaino, J. A.; Mueller, M.; Hermjakob, H.; Martens, L., Charting online OMICS resources: A navigational chart for clinical researchers. *Proteomics. Clinical applications* **2009,** *3* (1), 18-29.
33. (a) Oveland, E.; Muth, T.; Rapp, E.; Martens, L.; Berven, F. S.; Barsnes, H., Viewing the proteome: How to visualize proteomics data? *Proteomics* **2015,** *15* (8), 1341-55; (b) Streit, M.; Lex, A.; Kalkusch, M.; Zatloukal, K.; Schmalstieg, D., Caleydo: connecting pathways and gene expression. *Bioinformatics* **2009,** *25* (20), 2760-1.
34. Vaudel, M.; Venne, A. S.; Berven, F. S.; Zahedi, R. P.; Martens, L.; Barsnes, H., Shedding light on black boxes in protein identification. *Proteomics* **2014,** *14* (9), 1001-5.
35. (a) Wilhelm, M.; Schlegl, J.; Hahne, H.; Moghaddas Gholami, A.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; Mathieson, T.; Lemeer, S.; Schnatbaum, K.; Reimer, U.; Wenschuh, H.; Mollenhauer, M.; Slotta-Huspenina, J.; Boese, J. H.; Bantscheff, M.; Gerstmair, A.; Faerber, F.; Kuster, B., Mass-spectrometry-based draft of the human proteome. *Nature* **2014,** *509* (7502), 582-7; (b) Wang, R.; Fabregat, A.; Rios, D.; Ovelleiro, D.; Foster, J. M.; Cote, R. G.; Griss, J.; Csordas, A.; Perez-Riverol, Y.; Reisinger, F.; Hermjakob, H.; Martens, L.; Vizcaino, J. A., PRIDE Inspector: a tool to visualize and validate MS proteomics data. *Nature biotechnology* **2012,** *30* (2), 135-7.
36. (a) Kanehisa, M., A database for post-genome analysis. *Trends in genetics : TIG* **1997,** *13* (9), 375-6; (b) Sussman, J. L.; Lin, D.; Jiang, J.; Manning, N. O.; Prilusky, J.; Ritter, O.; Abola, E. E., Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta crystallographica. Section D, Biological crystallography* **1998,** *54* (Pt 6 Pt 1), 1078-84; (c) Apweiler, R.; Bairoch, A.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.;

Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L. S., UniProt: the Universal Protein knowledgebase. *Nucleic acids research* **2004,** *32* (Database issue), D115-9; (d) Flicek, P.; Amode, M. R.; Barrell, D.; Beal, K.; Billis, K.; Brent, S.; Carvalho-Silva, D.; Clapham, P.; Coates, G.; Fitzgerald, S.; Gil, L.; Giron, C. G.; Gordon, L.; Hourlier, T.; Hunt, S.; Johnson, N.; Juettemann, T.; Kahari, A. K.; Keenan, S.; Kulesha, E.; Martin, F. J.; Maurel, T.; McLaren, W. M.; Murphy, D. N.; Nag, R.; Overduin, B.; Pignatelli, M.; Pritchard, B.; Pritchard, E.; Riat, H. S.; Ruffier, M.; Sheppard, D.; Taylor, K.; Thormann, A.; Trevanion, S. J.; Vullo, A.; Wilder, S. P.; Wilson, M.; Zadissa, A.; Aken, B. L.; Birney, E.; Cunningham, F.; Harrow, J.; Herrero, J.; Hubbard, T. J.; Kinsella, R.; Muffato, M.; Parker, A.; Spudich, G.; Yates, A.; Zerbino, D. R.; Searle, S. M., Ensembl 2014. *Nucleic acids research* **2014,** *42* (Database issue), D749-55; (e) Uhlen, M.; Fagerberg, L.; Hallstrom, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, A.; Kampf, C.; Sjostedt, E.; Asplund, A.; Olsson, I.; Edlund, K.; Lundberg, E.; Navani, S.; Szigyarto, C. A.; Odeberg, J.; Djureinovic, D.; Takanen, J. O.; Hober, S.; Alm, T.; Edqvist, P. H.; Berling, H.; Tegel, H.; Mulder, J.; Rockberg, J.; Nilsson, P.; Schwenk, J. M.; Hamsten, M.; von Feilitzen, K.; Forsberg, M.; Persson, L.; Johansson, F.; Zwahlen, M.; von Heijne, G.; Nielsen, J.; Ponten, F., Proteomics. Tissue-based map of the human proteome. *Science* **2015,** *347* (6220), 1260419.

37. Venne, A. S.; Kollipara, L.; Zahedi, R. P., The next level of complexity: crosstalk of posttranslational modifications. *Proteomics* **2014,** *14* (4-5), 513-24.

38. Altelaar, A. F.; Munoz, J.; Heck, A. J., Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature reviews. Genetics* **2013,** *14* (1), 35-48.

39. Lamb, J.; Crawford, E. D.; Peck, D.; Modell, J. W.; Blat, I. C.; Wrobel, M. J.; Lerner, J.; Brunet, J. P.; Subramanian, A.; Ross, K. N.; Reich, M.; Hieronymus, H.; Wei, G.; Armstrong, S. A.; Haggarty, S. J.; Clemons, P. A.; Wei, R.; Carr, S. A.; Lander, E. S.; Golub, T. R., The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **2006,** *313* (5795), 1929-35.

40. (a) Barsnes, H.; Martens, L., Crowdsourcing in proteomics: public resources lead to better experiments. *Amino acids* **2013,** *44* (4), 1129-37; (b) Matic, I.; Ahel, I.; Hay, R. T., Reanalysis of phosphoproteomics data uncovers ADP-ribosylation sites. *Nat Methods* **2012,** *9* (8), 771-2; (c) Hahne, H.; Moghaddas Gholami, A.; Kuster, B., Discovery of O-GlcNAc-modified proteins in published large-scale proteome data. *Molecular & cellular proteomics : MCP* **2012,** *11* (10), 843-50; (d) Vaudel, M.; Burkhart, J. M.; Zahedi, R. P.; Oveland, E.; Berven, F. S.; Sickmann, A.; Martens, L.; Barsnes, H., PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nature biotechnology* **2015,** *33* (1), 22-4.

41. Volders, P. J.; Verheggen, K.; Menschaert, G.; Vandepoele, K.; Martens, L.; Vandesompele, J.; Mestdagh, P., An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic acids research* **2015,** *43* (Database issue), D174-80.

42. Cronin, M. T.; Bajot, F.; Enoch, S. J.; Madden, J. C.; Roberts, D. W.; Schwobel, J., The in chemico-in silico interface: challenges for integrating experimental and computational chemistry to identify toxicity. *Altern Lab Anim* **2009,** *37* (5), 513-21.

43. Aasebo, E.; Vaudel, M.; Mjaavatten, O.; Gausdal, G.; Van der Burgh, A.; Gjertsen, B. T.; Doskeland, S. O.; Bruserud, O.; Berven, F. S.; Selheim, F., Performance of super-SILAC based quantitative proteomics for comparison of different acute myeloid leukemia (AML) cell lines. *Proteomics* **2014**.

44. Bjorlykke, Y.; Vethe, H.; Vaudel, M.; Barsnes, H.; Berven, F. S.; Tjora, E.; Raeder, H., Carboxyl-Ester Lipase Maturity-Onset Diabetes of the Young Disease Protein Biomarkers in Secretin-Stimulated Duodenal Juice. *Journal of proteome research* **2014**.

Figure Legends

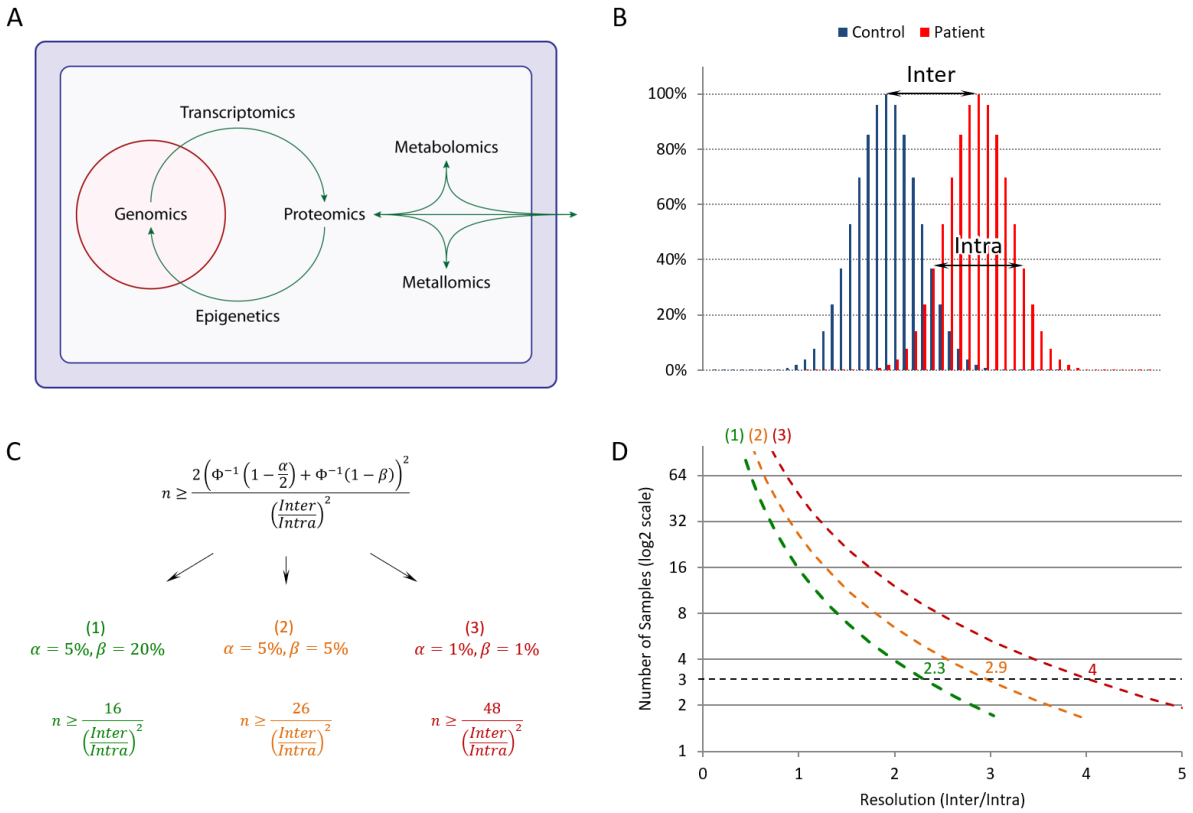

Figure 1

A. Representation of the different omics techniques. Genomics and transcriptomics are used to characterize the gene transcriptional activity and splice variation of the genes transcribed. Proteomics then allows the characterization of the proteins in their different forms and modification statuses. In turn, the regulation of gene expression operated by a portion of the proteome is characterized using epigenetics. The interaction of the proteins with metabolites and metals is characterized using metabolomics and metallomics, respectively.

B. Omics techniques can be used to compare populations, e.g. patients and controls, of different conditions, e.g. disease and healthy. The separation power of discriminating factors is then evaluated by studying the significance of the inter variability compared to the intra variability, as illustrated here by two artificial Gaussian distributions. The inter-intra variation ratio can be used to illustrate the resolution of the experiment.

C. The properties of the distributions and the test allow linking the inter-intra variation ratio, the resolution of the experiment, to the properties of the statistical model. As illustrated here with a normal distribution and a Student's t-test, the number of samples n can be directly linked to the resolution of the experiment $\frac{Inter}{Intra}$, and the stringency of the testing α and β, where α represents the probability of assuming that the populations differ while they do not, a false positive result,

and β the probability of assuming that the populations do not differ while they do, a false negative result. Φ represents the cumulative probability of the normal distribution. The numerator can easily be calculated as illustrated here for different stringencies after rounding with: (1) $\alpha = 5\%,\ \beta = 20\%$ in green, (2) $\alpha = 5\%,\ \beta = 5\%$ in orange, and (3) $\alpha = 1\%,\ \beta = 1\%$ in red.

D. Using the values of C, the number of samples necessary to achieve a given resolution is plotted for the three different cases, note the log scale for the number of samples. Reciprocally, it is possible to see the maximal resolution achievable with a given number of samples, as illustrated here by the resolution which can be achieved with three replicates at the different levels of stringency: (1) 2.3 for $\alpha = 5\%,\ \beta = 20\%$, in green, (2) 2.9 for $\alpha = 5\%,\ \beta = 5\%$, in orange, and (3) 4 for $\alpha = 1\%,\ \beta = 1\%$, in red.

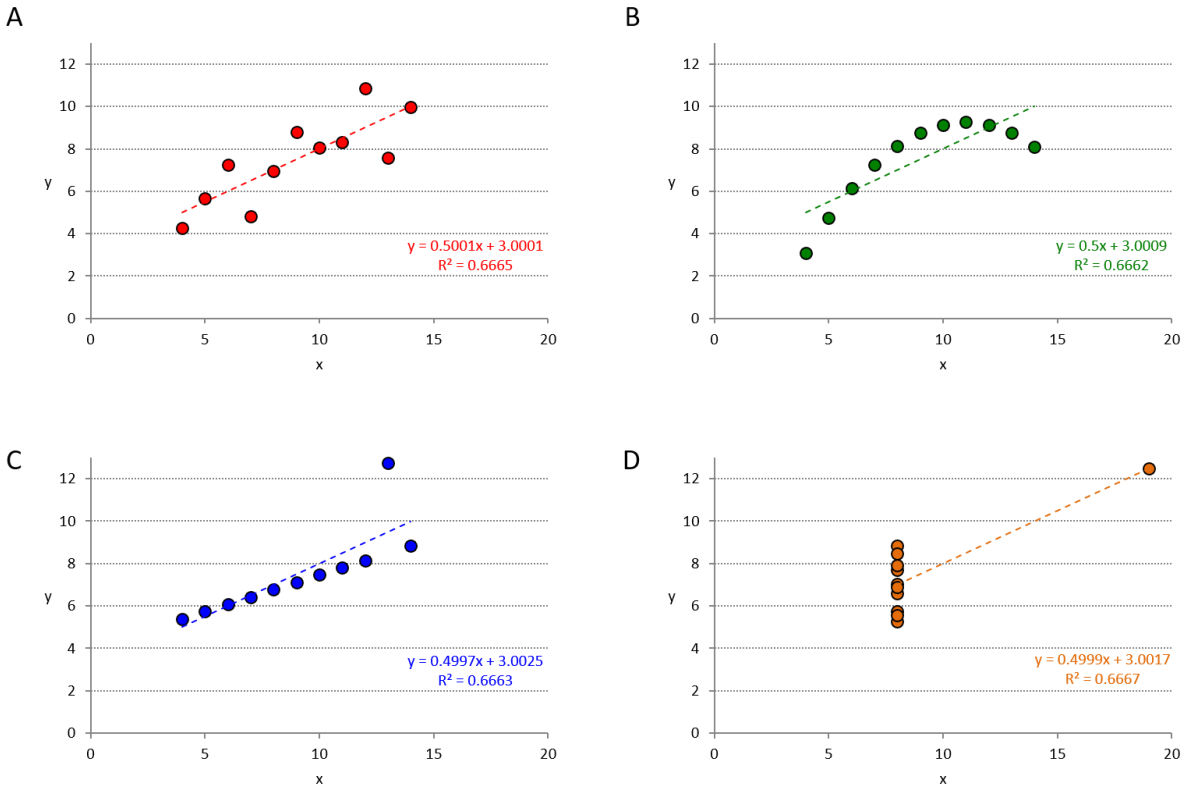

Figure 2

Representation of Anscombe's quartet[17] illustrating the importance of visually inspecting results and not simply relying on descriptive statistics. Here four series of doublets are plotted, which, even though barely distinguishable by linear regression, clearly have completely different shapes when visually inspected.

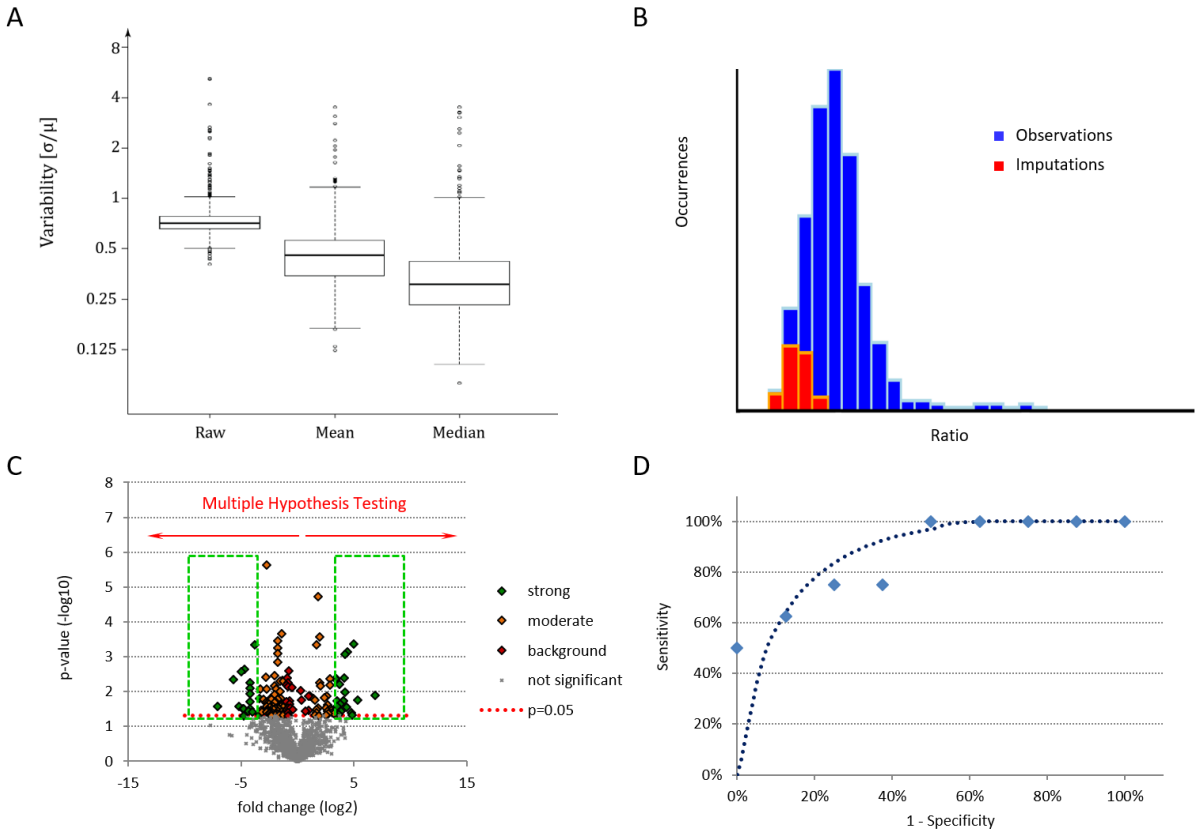

Figure 3

A. Variability of the cerebrospinal fluid expression of proteins unaffected by blood flow among different patients, courtesy of Dr. J.A. Opsahl. The variability was evaluated by the standard deviation divided by the mean of the protein abundance between patients. A box plot of the variability for the different proteins is displayed before normalization (Raw), and after normalization by the mean and median of the protein expression level for every patient.

B. Histograms of the protein expression ratio between Acute Myeloid Leukemia cell lines[43], with the observations in blue, and the missing values imputation in red using Perseus.

C. Example of a volcano plot obtained from the protein expression of induced pluripotent stem cells before and after reprogramming, courtesy of H. Vethe and Prof. Dr. H. Ræder. The significance of the testing is plotted against the fold change of the protein expression between replicates, note the logarithmic scales. Proteins were clustered in four categories after thresholding at a p-value of 0.05: (1) strong regulation, *i.e.* ≥10, and significant p-value in green, (2) moderate regulation, *i.e.* <10 and ≥2, and significant p-value in orange, (3) background regulation, *i.e.* <2, and significant p-value in red, and (4) not significant p-value in grey. Green rectangles indicate the most promising candidates, presenting strong and significant regulations.

D. Example of a Receiver Operating Characteristic (ROC) obtained from the monitoring of a protein from the MAPK pathway in the duodenal juice of diabetic patients and controls[44], with

the sensitivity of the patient separation plotted against the specificity, when going from the highest regulation to the lowest (blue diamonds), and the interpolation using non symmetrical normal distributions calibrated on the quartiles of the protein expression as a dotted line.

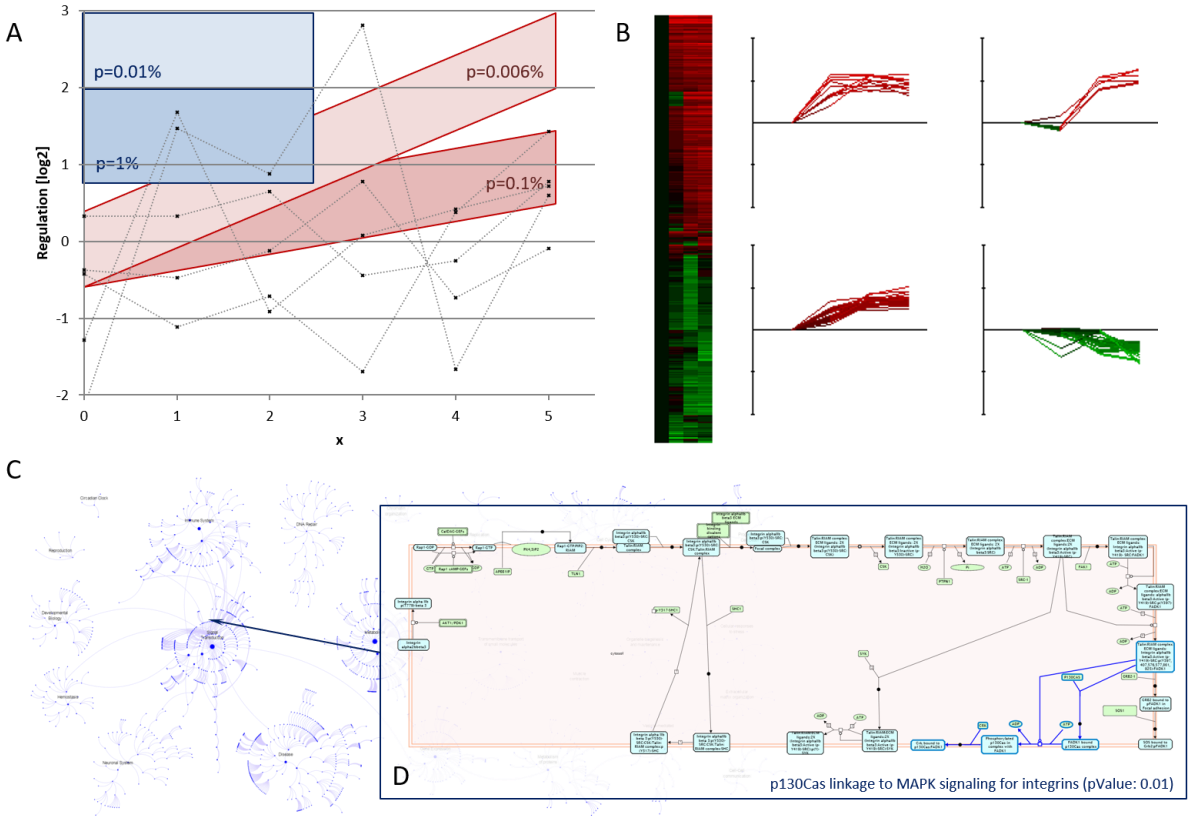

Figure 4

A. Illustration of the interest of series compared to on-off situations. Randomly generated series of six regulations with a standard deviation of 1 are plotted representing a simulated regulation ($\log_2$ scale) against six conditions, e.g. time points. To the left, two blue squares represent the levels to be achieved by three replicates to have a regulation of at least 1.7 or 4 (0.8 and 2 in base 2 logarithm, respectively). As indicated in the squares, the probability of three replicates to achieve such a regulation by chance is 1% and 0.01%, respectively. Two red bands illustrate the pattern a series would have to follow to achieve this magnitude of regulation at $x = 4$, with a tolerance of 0.5 standard deviation, i.e. $y = 0.2x \pm 0.5$ and $y = 0.5x \pm 0.5$, respectively. As indicated in the red bands, the probability of having a series of random values following these patterns is 0.1% and 0.006%, respectively. Here x can be time, distance, or any dimension of interest, see main text for details.

B. Heat map and illustrative example of clusters obtained from a time series study of phosphorylation events during platelet activation obtained using EPCLUST (http://www.bioinf.ebc.ee/EP/EP/EPCLUST) on data from Beck *et al.*[30].

C. Illustration of functional networks as displayed in Reactome[32a] (www.reactome.org).

D. Illustration of the *p130Cas linkage to MAPK signaling for integrins* pathway as displayed in Reactome[32a] (www.reactome.org). This pathway was covered by the protein expression significantly different between Acute Myeloid Leukemia cell lines[43], as highlighted in blue, and was given a p-value of 0.01.