

Review

# Find the Needle in the Haystack, Then Find It Again: Replication and Validation in the ‘Omics Era

Wei Perng<sup>1,2,\*</sup>  and Stella Aslibekyan<sup>3</sup>

<sup>1</sup> Department of Epidemiology, Colorado School of Public Health, University of Colorado Denver Anschutz Medical Campus, Aurora, CO 80045, USA

<sup>2</sup> Lifecourse Epidemiology of Adiposity and Diabetes (LEAD) Center, University of Colorado Denver Anschutz Medical Campus, Aurora, CO 80045, USA

<sup>3</sup> Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL 35294, USA; saslibek@uab.edu

\* Correspondence: wei.perng@cuanschutz.edu

Received: 29 May 2020; Accepted: 10 July 2020; Published: 12 July 2020



**Abstract:** Advancements in high-throughput technologies have made it feasible to study thousands of biological pathways simultaneously for a holistic assessment of health and disease risk via ‘omics platforms. A major challenge in ‘omics research revolves around the reproducibility of findings—a feat that hinges upon balancing false-positive associations with generalizability. Given the foundational role of reproducibility in scientific inference, replication and validation of ‘omics findings are cornerstones of this effort. In this narrative review, we define key terms relevant to replication and validation, present issues surrounding each concept with historical and contemporary examples from genomics (the most well-established and upstream ‘omics), discuss special issues and unique considerations for replication and validation in metabolomics (an emerging field and most downstream ‘omics for which best practices remain yet to be established), and make suggestions for future research leveraging multiple ‘omics datasets.

**Keywords:** replication; validation; ‘omics; genomics; metabolomics; integrative ‘omics

---

“The natural scientist is concerned with a particular kind of phenomenon . . . he has to confine himself to that which is reproducible . . . I do not claim that the reproducible by itself is more important than the unique. But I do claim that the unique exceeds the treatment by scientific method.” Wolfgang Pauli

Advances in high-throughput technologies have ushered in an era of “big data”, wherein researchers are no longer relegated to investigating single biological pathways. Rather, scientists grapple with a newfound ability to study thousands of pathways simultaneously for a more holistic understanding of (patho)physiology. The analysis and interpretation of large, high-dimensional biological data is spearheaded by the ‘omics fields comprising genomics, epigenomics, transcriptomics, proteomics, and metabolomics. The oldest and best known ‘omics is genomics, or the comprehensive study of DNA within an organism’s genome. Genomics is complemented by epigenomics, the study of the complete set of potentially reversible epigenetic modifications made to the DNA sequence. The next ‘omics layer is transcriptomics, the systematic study of RNA transcripts and gene expression. Following transcriptomics is proteomics, the study of proteins and their role in carrying out basic life processes. At the end of the ‘omics cascade is metabolomics, the study of low-molecular-weight compounds in biological tissues and fluids.

While much of the data generated by the ‘omics fields in the last two decades have been qualitative or semi-quantitative in nature, rigorous quantitative analysis is key to formulating and testing novel

hypotheses with the greatest potential to yield actionable next steps in the health sciences. Yet, a major challenge of quantitative ‘omics research revolves around the reproducibility of findings—a feat that lies in balancing false-positive findings (i.e., type 1 error) with generalizability (i.e., results that are too specific to a given study sample due to systematic bias [1] and/or model overfitting).

In this narrative review, we discuss the concepts of validation and replication in the context of genomics and metabolomics. Genomics is highly relevant to reproducibility given that many accepted approaches to ‘omics analyses were established in the field of genomics. At the other end of the spectrum, metabolomics is the newest ‘omics field, and the most downstream component of the ‘omics cascade. As such, the metabolome is closest to phenotype (and thus, perhaps most useful as a biomarker of disease risk or prognosis) but also exhibits the most inter- and intra-individual variability, thereby making replication and validation more challenging. We begin by defining key terminology, followed by a presentation and discussion of commonly used approaches, summarize guidelines (Table 1) for rigorous follow-up analyses of initial discoveries, and make recommendations for ways in which the methodology discussed herein may be applied to metabolomics research.

**Table 1.** Best practices for replication and validation of ‘omics findings.

<b>Both</b>
<ul style="list-style-type: none"> <li>• A priori calculations of the sample size required to detect a realistic effect.</li> <li>• Use of publicly available ‘omics datasets and/or pursuit of collaborations with other cohorts/consortia to maximize statistical power.</li> <li>• Stringent and appropriate corrections for multiple testing.</li> <li>• Transparent reporting of all relevant methods (from the laboratory work to bioinformatics pipelines, to data cleaning, to formal data analysis), features, and results (including those that failed to establish reproducibility).</li> </ul>
<b>Validation</b>
<ul style="list-style-type: none"> <li>• Harmonizing data across the discovery and validation stages to reduce the likelihood of non-reproducible findings due to systemic differences.</li> <li>• Inclusion of diverse datasets in validation efforts and the use of appropriate statistical methods to account for the resulting heterogeneity.</li> <li>• Judicious incorporation of functional annotations and effect sizes (in addition to statistical significance) when selecting features for validation and interpreting findings.</li> <li>• Distinguishing between reproducibility, functional relevance, and predictive validity, and using the appropriate metrics for each.</li> </ul>
<b>Replication</b>
<ul style="list-style-type: none"> <li>• Original (discovery) and confirmatory (replication) populations should be similar in terms of sex, age, and race/ethnic distributions.</li> <li>• Use of identical laboratory procedures, data processing pipelines, and analytical approaches.</li> </ul>

## 1. Terminology

While the practice of replication and validation is considered to be *de rigueur* in genetic epidemiology and genomics, there remains confusion regarding the two commonly used terms. Confirmation studies often seek to “replicate” or “validate” results without defining what this attempt comprises. As pointed out by Igl et al. [2], investigators have used the two terms interchangeably to describe when the same disease-biomarker associations are detected in two independent study populations [3], while other studies exclusively use the term “validation” [4] or “replication” [5]. Cavalier use of terminology is further compounded by varying degrees of similarity between the original sample (i.e., the discovery sample) vs. The replication/validation sample. For example, epidemiologists may conduct validation/replication studies among study populations of different racial backgrounds to interrogate the utility of a biomarker across the full spectrum of human diversity. Alternatively, laboratory scientists may be interested in using different technologies to measure the same biological relationships in the same sample to corroborate legitimacy of a new technique vs. a gold standard. These inconsistencies make it such that discrepancies in results between the discovery and replication/validation samples may not only be explained by random variation, but also by systematic differences. Such issues complicate the interpretation of results – especially when confirmation of the

initial findings fails. Thus, in this section, we define replication and validation and provide examples of each in the field of metabolomics.

### 1.1. Replication

Replication is the more straightforward of the two types of confirmatory studies. When replication is the goal, the original and confirmatory study samples should be independent but similar (i.e., similar sex, age, and race/ethnic distribution among the two samples) to reduce systematic differences, and the laboratory method used should be identical. Under these conditions, discrepancies in results are presumably due to random variation. Clarke et al. [6] describe the gold standard for replication as a test of the association of interest under identical circumstances using an independent sample. In the case of 'omics analyses that include both untargeted (hypothesis-generating) and targeted (hypothesis-driven) approaches, replication entails using the same biological and statistical approaches in the discovery and replication samples (e.g., the choice of analytical platform, timing of measurements, quality control procedures, statistical models, and other design variables), and success is defined by the reproducibility of findings across the two populations. In practice, however, achieving such parity of conditions presents a nontrivial challenge, especially when multiple cohorts, investigators, and measurement periods are involved.

Replication is particularly challenging for metabolomics, given that circulating metabolites are subject to metabolic flux (up-regulation or down-regulation of biochemical pathways), even in the absence of alterations to upstream 'omics that influence metabolism. This phenomenon contributes not only to inter-individual variability (which is relatively easy to control or account for in the study design or statistical analysis), but also intra-individual variability in compounds due to factors for which researchers do not typically collect information, such as time of day of data collection [7], season [8], and circadian or hormonal cycles [9]. Therefore, most efforts to confirm 'omics findings fall under the umbrella of validation (discussed in the next section) rather than true replication. We note that in the absence of true replication, there are specific examples in the literature where certain metabolite profiles have emerged as being relevant to specific biological conditions across different study populations and age groups, and using different laboratory platforms and statistical techniques. One such example is that of metabolites involved in branched chain amino acid and acylcarnitine metabolic pathways which have been identified as correlates [10–14] and prospective risk factors for cardiometabolic biomarkers [15], as well as biomarkers that are evident up to a decade prior to development of obesity-related chronic disease like type 2 diabetes [16,17]. While these studies were conducted without the a priori intent of replication, consistency in findings within the broader literature provide confidence that such findings are not spurious.

### 1.2. Validation

The term “validation” is less straightforward, and refers to a few distinct concepts. The most established meaning of the term predates the era of 'omics discovery, and refers to comparisons of specific laboratory techniques against a gold standard and is usually conducted on the same population. A hypothesis-free association study may yield several associations of interest (“hits”) in the discovery sample. The practice to date has been to validate such hits using the more expensive and time-consuming methods. In genomics, this approach has historically involved Sanger sequencing to confirm the nucleotide sequence. In metabolomics, this process involves implementation of targeted metabolomics assays to measure absolute concentrations of metabolites on a specific biochemical pathway following an initial associational discovery analysis of data generated on an untargeted metabolomics platform. In such a scenario, the discovery and validation samples are from the same population so that the only variation is introduced by the laboratory technique rather than systematic differences between populations. This form of validation is best suited for novel laboratory techniques or platforms and is less applicable to large-scale population-based studies.

The second distinction is internal vs. external validation. Internal validation is used in the development of predictive models that seek to identify features most strongly associated with the disease or biological condition of interest. Building the predictive model requires a training dataset to identify a set or a group of features that account for maximal variability in the biological condition. After training, the model is tested on an independent dataset or multiple resampled datasets to assess model validity. Techniques like bootstrapping and cross-validation optimize the model and check for overfitting. Examples of such approaches include random forests [18,19], a classification method that operates by constructing a multitude of decision trees with the goal of reducing variance; partial least squares-discriminant analysis [20], which relates the independent variable (e.g., metabolite concentration) to the response vector (the biological condition of interest) with a regression model; and least absolute shrinkage and selection operator (LASSO) [21]. A critical limitation of predictive modeling is the tendency to overfit data; thus, internal validation in an independent sample is of paramount importance. On the other hand, external validation refers to a study where the confirmation sample is a different population from that of the discovery sample. Differences between the two samples may be with respect to ethnic background or other sociodemographic characteristics, recruitment or sampling strategy, ascertainment of key variables, or time-point of investigation. In such scenarios, differences in findings may be explained by both random and systematic variability. Thus, large-scale, agnostic hypothesis-generating studies that achieve reproducibility using this approach indicate a robust association of the 'omics feature with the biological condition and supports an argument for causality. In contrast, failure to reproduce findings from the discovery sample in a sufficiently different population does not necessarily imply false positivity of the initial finding, but, rather, a need to attempt external validation again. This strategy is currently the most prevalent in the field due to the diversity of existing cohorts; therefore, we will focus our discussion on this approach to validation.

## 2. A Deeper Dive into Validation

### 2.1. Selecting Features for Validation

#### 2.1.1. Correcting for Multiple Comparisons

With the advent of genome-wide association studies (GWAS) and GWAS consortia like the U.K. Biobank [22] or industry cohorts (e.g., 23andMe), the sample size within which scientists may explore single nucleotide polymorphism (SNP)–trait associations increased upwards of a thousand-fold, thereby providing greater statistical power to detect true associations. Additionally, these consortia served as large, diverse populations for validation of cohort-specific findings. As a consequence, the trend of irreproducible SNP–trait associations that plagued early genomics studies has been sharply reversed [23]. In metabolomics, researchers have taken approaches that mirror those of GWAS by examining associations of individual metabolites with a predictor or outcome and correcting for multiple comparisons in metabolome-wide association studies (MWAS). Yet, because metabolomics data exhibit high collinearity among compounds on similar biochemical pathways, there is greater likelihood for biological and statistical correlation among metabolites than SNPs, and conventional approaches to multiple comparisons correction are likely too stringent. Arguably, the most popular approach is to control the family wise error rate, which is the probability of one or more significant results under the null hypothesis of no association. For a high-dimensional dataset, the simplest way to do this is to control for the number of molecular variables using the Bonferroni correction. However, these methods assume independence among the molecular variables of interest. Less conservative approaches include the Benjamini–Hochberg false discovery rate (FDR) procedure [24], which sets a threshold for statistical significance based on the distribution of  $p$ -values, given an allowable error rate. The FDR procedure has been successfully applied to datasets with a relatively large number of true positive associations, such as gene expression data. However, the use of FDR procedures for MWAS is compounded by the correlated nature of individual metabolic pathways (and super- and sub-pathways, for that matter), and lack of available annotations for much of the human metabolome. Thus, assignment

of molecular formulas to spectral peaks during the data processing step of metabolomics pipelines does not yield comprehensive structural identifications that are critical to biological understanding [25]. Suffice it to say, identification of suitable multiple testing thresholds in MWAS is a Herculean task and an area of active investigation that involves crosstalk among laboratory scientists, bioinformaticians, and statisticians. These efforts, in conjunction with metabolomics consortia—so far, the primary one being the COntortium of METabolomics Studies (COMETS)—will provide valuable resources that may be used to validate initial cohort-specific findings and to test/compare the utility of appropriate multiple comparisons correction procedures.

Around the time when genetic consortia rose to prominence, it also became commonplace to reduce the dimensionality of big data by aggregating the molecular variables into pathways to first identify pathways of interest prior to assessment of individual features within the pathways. Dimension reduction is especially attractive and biologically intuitive for metabolomics analyses. Widely-used multivariate procedures include unsupervised principal components analysis [10,13,26] and weighted correlation network analysis [27] to consolidate metabolites into clusters prior to regression analysis. However, it must be cautioned that such procedures can lead to higher rates of false-positive findings [28], since investigators typically implement multiple testing corrections at the step of identifying metabolite clusters of interest, but not when examining associations with respect to individual metabolites. Additionally, because dimension reduction procedures often require complete data, the investigator must accept heroic assumptions about missing values. The assumptions are made following decisions regarding whether missing values are biologically meaningful (e.g., indicative of lower exposure to an environmental factor and thus, lower exposure of a metabolite affected by that factor) or whether missingness is a result of errors in laboratory procedures. If investigators suspect the former to be the case, then imputation procedures are implemented prior to use of dimension reduction algorithms. Common algorithms include setting missing values to  $\frac{1}{2}$  the lowest detected value when the percent of missing values for a compound is below an arbitrary threshold (typically < 20%); or more sophisticated approaches that yield a range of values, like K-nearest neighbor algorithm, when the proportion of missingness is higher [29]. Alternatively, if metabolite missingness is suspected to be due to issues with the laboratory procedure, then the investigator may opt to omit that particular compound from the analysis altogether. The advantages, disadvantages, and appropriateness of these approaches are discussed in detail elsewhere [30]. The key point here is that given the interconnectedness of metabolic pathways, such decisions are not trivial as their consequences have reverberations for subsequent multivariate analyses and interpretation of results.

### 2.1.2. Metrics for Selection of Findings to Follow-Up Studies

The traditional approach to validating 'omics findings begins with identifying a set of features that are associated with biological trait of interest at a predefined threshold of statistical significance, then further exploring the feature of interest in follow-up analyses. As discussed in the previous section, conservative corrections for multiple comparisons improve the reproducibility of 'omics findings, but because metabolomics features are not independent, such correction is overly conservative. Moreover, corrections based on the number of aggregate pathways rather than single features have been shown to be insufficient [28], leaving room for development of other approaches. Thus, an ideal approach—particularly for validating discovery analyses—is to first implement a dimension reduction procedure to identify aggregate pathways of interest while controlling type 1 error using an appropriate method for the multivariate procedure implemented (e.g., Bonferroni correction is acceptable for procedures that derive uncorrelated latent variables like principal components analysis, whereas the more lenient FDR correction may be more appropriate for procedures like WGCNA where metabolite networks may be correlated), then explore associations with respect to individual metabolites. On balance, acceptable false discovery rates depend on the research question and the relative potential harm caused by false-positive (irreproducible) vs. false-negative (missed) findings.

Such considerations will require investigators to weigh the costs vs. benefits of false-negative vs. false-positive findings, which will depend on the biological outcome of interest.

Of note, the most widely-used approach to selecting signals for further follow up (whether replication or validation) relies solely on  $p$ -values, leaving the analysis vulnerable to false-positive findings even after adjustment for multiple comparisons [31]. One remedy is to consider not only the  $p$ -value, but also the magnitude of association (i.e., effect size). This notion is fundamental to Sir Bradford Hill's first guideline for evaluating causality: the strength of an association provides support for a causal effect of an exposure on disease risk [32]. This is due, in part, to the fact that large effect sizes require a large amount of bias to attenuate to the null and therefore are robust to both random and systematic error (the interested reader may peruse the article by Fedek et al. [33] for additional insights into application of Hill's causal guidelines in the age of high-dimensional data).

In practice, there are a number of statistical techniques that focus on identifying a set of predictors from a high-dimensional dataset on the basis of magnitude of association, in addition to statistical significance. For example, the simulation extrapolation (SIMEX) algorithm has been successfully used for metabolomics data [34] estimates regression coefficients for the strongest set of predictors via a simulation step that generates multiple datasets with varying degrees of random error in the original data that plague laboratory procedures (e.g., residual batch effects; uncontrolled and/or unmeasured variation in technical methods; noise), followed by an extrapolation step that corrects for the additive measurement error via a covariate representing the impact of unknown sources of error based on the simulated data. Another option involves use of regularized regression techniques like LASSO [21], which impose a tuning parameter on model coefficients in order to shrink weak or null  $\beta$  estimates to 0. This "shrinkage" step effectively removes predictors that are statistically significant but only weakly associated with the outcome. Such approaches have several advantages for both inference and prediction, especially when paired with cross-validation and resampling procedures (e.g., bootstrap LASSO [35]), including a reduction in false-positive findings and protecting against model-overfitting.

Alternative strategies, such as Bayesian analyses providing a priori estimates that the association is falsely positive [36], are occasionally discussed in literature but rarely implemented, likely due to their complexity and strong assumptions. An emerging body of work suggests that augmenting  $p$ -value information with functional annotation when selecting hits for validation may improve reproducibility. These characteristics, summarized by Gorlov et al. [37] into a cogent "reproducibility index" used in genomics, range from practical (e.g., allele frequency) to biological (e.g., whether the protein encoded by the gene containing the variant of interest is a receptor or a transcription factor). Such criteria may be easily adopted for metabolomics research. However, a major challenge for metabolomics investigations lies in prioritizing features with known mechanisms of action at the expense of truly novel discoveries. More generally, the utility of incorporating functional data depends on the strength of prior evidence [38], and is expected to increase with future molecular and biochemical advances.

### 2.1.3. Validation beyond Single Features

The increasing affordability of next-generation technologies has prompted re-evaluation of validation practices beyond single feature associations. That is to say, by prioritizing whole genes in genomics, differentially methylated regions of the genome in epigenomics, clusters of RNA transcripts in transcriptomics, groups of metabolites in metabolomics, and more recently, networks for multiple 'omics variables. Validation of such integrative analyses requires internal cross-validation approaches to ensure stability and consistency in the 'omics clusters. Many of these approaches follow those implemented in the field of genomics, but have been successfully applied to metabolomics or multiple-'omics datasets. These include stability analyses to evaluate the optimal number of clusters [39], sample bootstrapping to ensure the stability of clusters [40,41], as well as qualitative assessment of whether key clusters are preserved across the validation attempts prior to external validation. We note that these analyses often involve exploratory unsupervised learning algorithms,

for which there are no traditional power calculations. As this area expands, additional research is required to fill this chasm.

## 2.2. Selecting a Population for Validation

Once findings from the discovery stage have been prioritized for validation, considerations of the appropriate population for follow-up tests become paramount. The ideal choice of population lies between the Scylla of split-sample analysis (i.e., a priori splitting the original cohort into discovery and validation subsets) and the Charybdis of choosing a follow-up cohort that is so systematically different from the discovery one that any chances for successful validation are diminished.

In the first scenario, dividing the original sample into two subsamples and requiring statistical significance at a given level  $\alpha$  in each of the subsets yields lower statistical power than analyzing the entire dataset at the  $\alpha^2$  level [31]. This approach is usually undertaken when choosing a truly independent external cohort is not feasible, which occurs in studies of rare phenotypes or unique populations. In such cases, iterative resampling approaches, such as one by Kang et al., successfully implemented in GWAS studies [42]; or K-fold cross-validation, which has proven utility for metabolomics [43] may be employed to optimize statistical power. However, the performance of such methods is contingent on the proper choice of parameters, such as the ratio of the discovery to validation subsets.

In the second scenario of selecting an independent validation population, all efforts must be undertaken to harmonize the phenotype definitions and other experimental conditions across the discovery and validation samples. To reduce systematic variation, statistical models used in the validation stage should also mirror those used in the discovery stage as closely as possible. Unaccounted sources of variation between samples, whether due to discrepant trait measurements, covariate adjustment, or population characteristics, can lead to falsely negative results in the validation stage. Differences between discovery and validation populations need to be documented and, if possible, accounted for statistically. Importantly, criteria for reconciling phenotypic definitions must be established prior to analysis to avoid cherry-picking only the definitions that yield statistical significance in the validation stage.

Special caution has been applied to validation efforts across different ethnic groups. This is intuitive and imperative for genomics studies. For metabolomics, the importance of similarity of race/ethnic distributions in discovery and validation populations has been debated. On the one hand, differences in genetic architecture undoubtedly affect multiple aspects of physiology and disease etiology, and accounting for such differences will identify metabolomics profiles associated with a biological outcome that are independent of race/ethnicity. In this scenario, researchers may be interested in identifying a biomarker that performs well in a specific racial or ethnic group, such as the example of ethnicity-specific serum biomarkers of coronary artery disease [44], or the investigator may be interested in understanding specific mechanisms underlying a certain biological condition, which may differ by racial or ethnic background whether it be due to underlying psychosocial or genetic pathways. On the other hand, many metabolomics analyses seek to identify population-level biomarkers of disease risk or prognosis used in clinical settings. In these instances, the utility of a valid biomarker should be apparent across racial and ethnic subgroups and thus, a validation population need not be similar to the discovery population in terms of race/ethnicity. The decision of matching on or accounting for race/ethnic distribution for validation studies hinges on the aims of the original study, which should be clearly articulated up front.

## 2.3. Interpreting Findings

What constitutes a validated finding? As the data landscape expands and statistical methods evolve, this question is critical to the success of 'omics science. Generally, a validated finding refers to an association between some feature and a phenotypic trait, where sufficient evidence exists to support a causal effect of the former on the latter. Sufficient causal evidence, in turn, most

commonly refers to reproducibility as defined using established guidelines [45]. We propose additionally considering evidence within the context of Hill's causal inference principles [5]. For instance, it is becoming commonplace to include data from bioinformatics databases to support biological plausibility, or use animal models and/or cell cultures to generate experimental evidence. The latter may be especially helpful in situations where establishing reproducibility proves challenging due to a lack of suitable validation cohorts, such as in some studies of exposure to unique pharmaceutical agents. (Nota bene, validation afforded by these functionally-driven approaches have been summarized by Gallagher et al. [46] and have potential to augment the lessons learned from statistical tests of the same feature-trait association across discovery and validation cohorts.)

In addition to the biological plausibility and experimental evidence guidelines discussed above, several other Hill's postulates can be used to contemplate the quality of causal evidence supporting a given association. Consistency is directly related to reproducibility: are the same findings repeatedly observed in different places, people, circumstances and times? Strength of the association refers to its magnitude, and distinguishes between mere statistical significance (easy to obtain given large enough sample sizes) and a meaningful effect, the latter of which will require longer-term follow-up studies in the field of metabolomics to understand how differences in a given metabolite concentration related to future health outcomes. Coherence situates the new finding within the existing body of evidence, and finally, considering the criterion of specificity can elucidate the etiologic architecture of a trait. While Hill himself warned against using his criteria as a checklist for causality, they nonetheless provide a useful framework for evaluating evidence generated by discovery and validation studies.

With the growing relevance of direct-to-consumer profiling [47] in the wake of precision nutrition/precision medicine [48], another meaning of validation has come into sharper focus: how well does a feature predict trait development and/or how well does it discriminate between affected and unaffected individuals? While the numerous ethical considerations pertaining to such efforts are beyond the scope of this review, it is important to note that predictive validity extends beyond mere reproducibility or functional validation, and should be assessed using appropriate metrics that would be relevant to clinical prediction (area under the receiver operating characteristic curve, prediction  $r^2$ , and others). Currently, the predictive power of 'omics findings lags far behind reproducibility, likely due to insufficient sample sizes to detect all true effects as well as other relevant factors that explain trait variance (e.g., biological interaction). It is important to remember that successful validation – as defined by establishing reproducibility and, if possible, functional relevance – represents a necessary but not sufficient step towards using 'omics findings to predict disease risk and, ultimately, serve as a way to identify viable avenues to improve health outcomes.

#### 2.4. Metadata

In a survey conducted by *Nature* in 2016, 90% of scientists responded that there was either a “slight” or “significant” crisis of reproducibility in science [49]. This narrative has risen to prominence in recent years [50], emphasizing the importance of independently verifiable scientific discoveries. Metabolomics findings are fraught with a unique set of challenges given technical differences in analytical instrumentation (nuclear magnetic resonance spectroscopy vs. mass spectrometry, the latter of which is paired with either gas or liquid chromatography or both) and biospecimen matrix (even among studies using blood, harmonizing data from serum vs. plasma is challenging due to the type of anticoagulant used for plasma samples), as well as differences that stem from the type of platform (untargeted vs. targeted) driven by the research question. These issues, in conjunction with the high level of variability in the metabolome, make it challenging to compare and contrast metabolomics data across studies. Thus, the bare minimum scientists can do, it seems, is to be completely transparent by publishing metadata, or data about the data (“annotation”), alongside scientific findings.

In 2007, the Metabolomics Standards Initiative (MSI) put forth a set of recommendations on minimal reporting standards for the laboratory chemical analysis, as well as data analysis [51]. These standards are used as guidelines for annotation of metabolomics metadata, and continue to

evolve. At present, key annotations comprise sample preparation, experimental analysis, quality control, compound identification, and data pre-processing, and focus on mass spectrometry and chromatography-based nuclear magnetic resonance spectroscopy due to the popularity of these techniques in metabolomics [52]. As additional instrumental (e.g., use of capillary electrophoresis based mass spectrometry), bioinformatics, and statistical techniques emerge, proper annotation of the scientific data to ensure metadata quality will become ever more important. As discussed by Ferreira et al. [53], quality control of metadata is no small task and will likely require automated mechanisms [54] to assess the quality of metadata. Metadata may include annotations of specific biological pathways that are relevant to the study findings, and/or a description of processes by which the data were collected (e.g., how the data was collected, by whom and when, storage procedures, biospecimen processing techniques, and software used) and processed (e.g., bioinformatics pipelines and decisions made regarding artefacts of the data). Ultimately, such semantics allow for the harmonization of data from different sources, facilitate reproducible research, and allow for more confident interpretation and dissemination of findings.

### 3. Summary of Considerations for Metabolomics

For metabolomics specifically, replication and validation is critical and challenging given the higher degree of variability in metabolites, and the vulnerability of metabolite concentrations to lifestyle and environmental factors that are of markedly less concern in other 'omics fields, notably genomics. Regardless, several topics discussed in this review are relevant to metabolomics analyses, including the need for confirmation via replication in independent study samples, the importance of internal and external validation (the latter of which requires selection of an appropriate study population to reduce systematic variability), selection of appropriate methods to correct for multiple comparisons, and use of criteria beyond *P*-values to identify "hits" – namely, consideration of the magnitude of effects. Additionally, while not formally discussed herein, researchers should be aware of statistical phenomena that may impact the magnitude of effects detected in initial discovery studies (e.g., "Winner's curse"), and employ *a priori* considerations of statistical power.

### 4. Future Directions: Integration of Multiple 'Omics

The use of multiple 'omics datasets is becoming increasingly popular in all facets of life science. In the biomedical sciences, genomics-metabolomics analyses using undirected network-based approaches have demonstrated capacity to recreate true biochemical pathways [55] and home in on those involved in disease etiology [56]. While such studies require a more complex set of strategies to deal with replication and validation, probing 'omics profiles from both ends of the central dogma of molecular biology—genetics on the one side, and metabolomics on the other side—to provide valuable information on modifiability of the mechanisms represented by the networks. That is to say, in a network where a SNP explains substantial variation in the cluster of molecular variables, we can infer that pathway of interest stems from genetic variation, and that metabolites are mediators of the process. Given the non-modifiability of genetics, identifying determinants of networks comprised predominantly of SNPs is less relevant to preventive action but may be informative for surveillance. On the other hand, in a network without any genetic associations, metabolic perturbations are likely to originate from external factors and thus, may represent modifiable mechanisms underlying disease progression. Accordingly, linking environmental and/or lifestyle characteristics to these "modifiable networks" will point toward avenues for prevention. As an example, Krumsiek et al. [57] analyzed untargeted metabolomics data from 1756 men and women in the KORA study to identify metabolites that exhibit marked differences in serum concentrations. These compounds were on steroid metabolism, fatty acids (a large fraction of which were related to amino acid metabolism), oxidative phosphorylation, and purine metabolism pathways. A network-based clustering of metabolites of interest with a gender-stratified GWAS revealed gender-specific differences in metabolic networks that were predominantly explained by variability in the metabolites, as opposed to the SNPs, suggesting

that gender differences in these metabolic pathways are not caused by genetic dimorphisms, but, rather, polymorphisms of the metabolome that may be influenced by the environment and/or behaviors. Future studies identifying lifestyle and behavioral correlates of such metabolic networks will shed light on specific preventive interventions that have potential to mitigate risk of chronic diseases known to be influenced by the above-mentioned biochemical pathways (e.g., type 2 diabetes, cardiovascular disease).

The above concepts are the embodiment of innovative studies that capitalize on multiple ‘omics datasets to unveil findings with meaningful implications for actionable next steps. The rapid growth of ‘omics technologies and statistical packages to analyze such data necessitates the need to establish best practices for bioinformatics and analytical workflows within each individual ‘omics field, as well as for multiple ‘omics. Given that reproducible research is the foundation upon which scientific inference is made, statistical validation and replication serve as pillars that flank the journey from hypothesis testing to informing next steps for preventive or therapeutic action.

**Author Contributions:** W.P. and S.A. conceived the idea for this review. Both authors drafted the initial manuscript and gave final approval of the content. All authors have read and agreed to the published version of the manuscript.

**Funding:** Perng’s salary is supported by KL2-TR002534.

**Conflicts of Interest:** Aslibekyan is an employee of 23andMe.

## References

1. Colhoun, H.M.; McKeigue, P.M.; Davey Smith, G. Problems of reporting genetic associations with complex outcomes. *Lancet (London, England)* **2003**, *361*, 865–872. [[CrossRef](#)]
2. Igl, B.W.; Konig, I.R.; Ziegler, A. What do we mean by ‘replication’ and ‘validation’ in genome-wide association studies? *Hum. Hered.* **2009**, *67*, 66–68. [[CrossRef](#)] [[PubMed](#)]
3. Greenwood, C.M.; Rangrej, J.; Sun, L. Optimal selection of markers for validation or replication from genome-wide association studies. *Genet. Epidemiol.* **2007**, *31*, 396–407. [[CrossRef](#)] [[PubMed](#)]
4. Todd, J.A.; Walker, N.M.; Cooper, J.D.; Smyth, D.J.; Downes, K.; Plagnol, V.; Bailey, R.; Nejentsev, S.; Field, S.F.; Payne, F.; et al. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat. Genet.* **2007**, *39*, 857–864. [[CrossRef](#)]
5. Gudbjartsson, D.F.; Arnar, D.O.; Helgadóttir, A.; Gretarsdóttir, S.; Holm, H.; Sigurdsson, A.; Jonasdóttir, A.; Baker, A.; Thorleifsson, G.; Kristjánsson, K.; et al. Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature* **2007**, *448*, 353–357. [[CrossRef](#)]
6. Clarke, G.M.; Carter, K.W.; Palmer, L.J.; Morris, A.P.; Cardon, L.R. Fine mapping versus replication in whole-genome association studies. *Am. J. Hum. Genet.* **2007**, *81*, 995–1005. [[CrossRef](#)]
7. Ang, J.E.; Revell, V.; Mann, A.; Mäntele, S.; Otway, D.T.; Johnston, J.D.; Thumser, A.E.; Skene, D.J.; Raynaud, F. Identification of human plasma metabolites exhibiting time-of-day variation using an untargeted liquid chromatography-mass spectrometry metabolomic approach. *Chronobiol. Int.* **2012**, *29*, 868–881. [[CrossRef](#)]
8. Kim Nk Park, H.M.; Lee, J.; Ku, K.-M.; Lee, C.H. Seasonal Variations of Metabolome and Tyrosinase Inhibitory Activity of *Lespedeza maximowiczii* during Growth Periods. *J. Agric. Food Chem.* **2015**, *63*, 8631–8639. [[CrossRef](#)]
9. Wallace, M.; Hashim, Y.Z.; Wingfield, M.; Culliton, M.; McAuliffe, F.; Gibney, M.J.; Brennan, L. Effects of menstrual cycle phase on metabolomic profiles in premenopausal women. *Hum. Reprod. (Oxford, England)* **2010**, *25*, 949–956. [[CrossRef](#)]
10. Perng, W.; Rifas-Shiman, S.L.; Sordillo, J.; Hivert, M.F.; Oken, E. Metabolomic Profiles of Overweight/Obesity Phenotypes During Adolescence: A Cross-Sectional Study in Project Viva. *Obesity (Silver Spring, Md)* **2020**, *28*, 379–387. [[CrossRef](#)]
11. Perng, W.; Hector, E.C.; Song, P.X.K.; Tellez Rojo, M.M.; Raskind, S.; Kachman, M.; Cantoral, A.; Burant, C.F.; Peterson, K.E. Metabolomic Determinants of Metabolic Risk in Mexican Adolescents. *Obesity (Silver Spring, Md)* **2017**, *25*, 1594–1602. [[CrossRef](#)] [[PubMed](#)]
12. Perng, W.; Oken, E.; Roumeliotaki, T.; Sood, D.; Siskos, A.P.; Chalkiadaki, G.; Dermitzaki, E.; Vafeiadi, M.; Kyrtopoulos, S.; Kogevas, M.; et al. Leptin, acylcarnitine metabolites and development of adiposity in the Rhea mother-child cohort in Crete, Greece. *Obes. Sci. Pract.* **2016**, *2*, 471–476. [[CrossRef](#)] [[PubMed](#)]

13. Perng, W.; Gillman, M.W.; Fleisch, A.F.; Michalek, R.D.; Watkins, S.M.; Isganaitis, E.; Patti, M.E.; Oken, E. Metabolomic profiles and childhood obesity. *Obesity (Silver Spring, Md)* **2014**, *22*, 2570–2578. [[CrossRef](#)]
14. Butte, N.F.; Liu, Y.; Zakeri, I.F.; Mohny, R.P.; Mehta, N.; Voruganti, V.S.; Göring, H.; Cole, S.A.; Comuzzie, A.G. Global metabolomic profiling targeting childhood obesity in the Hispanic population. *Am. J. Clin. Nutr.* **2015**, *102*, 256–267. [[CrossRef](#)] [[PubMed](#)]
15. Perng, W.; Rifas-Shiman, S.L.; Hivert, M.F.; Chavarro, J.E.; Oken, E. Branched Chain Amino Acids, Androgen Hormones, and Metabolic Risk Across Early Adolescence: A Prospective Study in Project Viva. *Obesity (Silver Spring, Md)* **2018**, *26*, 916–926. [[CrossRef](#)]
16. Wang, T.J.; Larson, M.G.; Vasan, R.S.; Cheng, S.; Rhee, E.P.; McCabe, E.; Lewis, G.D.; Fox, C.S.; Jacques, P.F.; Fernandez, C.; et al. Metabolite profiles and the risk of developing diabetes. *Nat. Med.* **2011**, *17*, 448–453. [[CrossRef](#)]
17. Flores-Guerrero, J.L.; Osté, M.C.J.; Kieneker, L.M.; Gruppen, E.G.; Wolak-Dinsmore, J.; Otvos, J.D.; Connelly, M.A.; Bakker, S.J.L.; Dullaart, R.P.F. Plasma Branched-Chain Amino Acids and Risk of Incident Type 2 Diabetes: Results from the PREVEND Prospective Cohort Study. *J. Clin. Med.* **2018**, *7*, 513. [[CrossRef](#)]
18. Chen, T.; Cao, Y.; Zhang, Y.; Liu, J.; Bao, Y.; Wang, C.; Jia, W.; Zhao, A. Random Forest in Clinical Metabolomics for Phenotypic Discrimination and Biomarker Selection. *Evid.-Based Complement. Altern. Med.* **2013**, *2013*, 298183. [[CrossRef](#)]
19. Kokla, M.; Virtanen, J.; Kolehmainen, M.; Paananen, J.; Hanhineva, K. Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: A comparative study. *BMC Bioinform.* **2019**, *20*, 492. [[CrossRef](#)]
20. Fonville, J.M.; Richards, S.E.; Barton, R.H.; Boulange, C.L.; Ebbels, T.M.D.; Nicholson, J.K.; Holmes, E.; Dumas, M.-E. The evolution of partial least squares models and related chemometric approaches in metabolomics and metabolic phenotyping. *J. Chemom.* **2010**, *24*, 636–649. [[CrossRef](#)]
21. Robert, T. Regression shrinkage and selection via the lasso: A retrospective. *J. R. Stat. Soc. Ser. B (Statistical Methodology)* **2011**, *73*, 273–282. [[CrossRef](#)]
22. Sudlow, C.; Gallacher, J.; Allen, N.; Beral, V.; Burton, P.; Danesh, J.; Downey, P.; Elliott, P.; Green, J.; Landray, M.; et al. UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **2015**, *12*, e1001779. [[CrossRef](#)] [[PubMed](#)]
23. Marigorta, U.M.; Rodríguez, J.A.; Gibson, G.; Navarro, A. Replicability and Prediction: Lessons and Challenges from GWAS. *Trends Genet. TIG* **2018**, *34*, 504–517. [[CrossRef](#)] [[PubMed](#)]
24. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B (Methodological)* **1995**, *57*, 289–300. [[CrossRef](#)]
25. Benjamini, Y.; Hochberg, Y. False discovery rate estimation for metabolomics. *Nat. Methods* **2018**, *15*, 15. [[CrossRef](#)]
26. Newgard, C.B.; An, J.; Bain, J.R.; Muehlbauer, M.J.; Stevens, R.D.; Lien, L.F.; Haqq, A.M.; Shah, S.H.; Arlotto, M.; Slentz, C.A.; et al. A branched-chain amino acid-related metabolic signature that differentiates obese and lean humans and contributes to insulin resistance. *Cell Metab.* **2009**, *9*, 311–326. [[CrossRef](#)]
27. Langfelder, P.; Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* **2008**, *9*, 559. [[CrossRef](#)]
28. Johnson, R.C.; Nelson, G.W.; Troyer, J.L.; Lautenberger, J.A.; Kessing, B.D.; Winkler, C.A.; O'Brien, S.J. Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics* **2010**, *11*, 724. [[CrossRef](#)]
29. Hastie, T.; Tibshirani, R.; Sherlock, G.; Eisen, M.; Brown, P.; Botstein, D. *Imputing Missing Data for Gene Expression Arrays*; Online; Department SUS, Stanford University: Palo Alto, CA, USA, 1999.
30. Wei, R.; Wang, J.; Su, M.; Jia, E.; Chen, S.; Chen, T.; Ni, Y. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Sci. Rep.* **2018**, *8*, 663. [[CrossRef](#)]
31. Thomas, D.C.; Casey, G.; Conti, D.V.; Haile, R.W.; Lewinger, J.P.; Stram, D.O. Methodological Issues in Multistage Genome-wide Association Studies. Statistical science: A review. *J. Inst. Math. Stat.* **2009**, *24*, 414–429. [[CrossRef](#)]
32. Hill, A.B. The Environment and Disease: Association or Causation? *Proc. R. Soc. Med.* **1965**, *58*, 295–300. [[CrossRef](#)] [[PubMed](#)]

33. Fedak, K.M.; Bernal, A.; Capshaw, Z.A.; Gross, S. Applying the Bradford Hill criteria in the 21st century: How data integration has changed causal inference in molecular epidemiology. *Emerg. Themes Epidemiol.* **2015**, *12*, 14. [[CrossRef](#)] [[PubMed](#)]
34. Cook, J.R.; Stefanski, L.A. Simulation-Extrapolation Estimation in Parametric Measurement Error Models. *J. Am. Stat. Assoc.* **1994**, *89*, 1314–1328. [[CrossRef](#)]
35. Bach, F.R. Bolasso: Model consistent lasso estimation through bootstrap. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–8 July 2008; pp. 33–40.
36. Hunter, D.J.; Kraft, P. Drinking from the fire hose—statistical issues in genomewide association studies. *N. Engl. J. Med.* **2007**, *357*, 436–439. [[CrossRef](#)] [[PubMed](#)]
37. Gorlov, I.P.; Moore, J.H.; Peng, B.; Jin, J.L.; Gorlova, O.Y.; Amos, C.I. SNP characteristics predict replication success in association studies. *Hum. Genet.* **2014**, *133*, 1477–1486. [[CrossRef](#)]
38. Lewinger, J.P.; Conti, D.V.; Baurley, J.W.; Triche, T.J.; Thomas, D.C. Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet. Epidemiol.* **2007**, *31*, 871–882. [[CrossRef](#)]
39. Lovmar, L.; Ahlford, A.; Jonsson, M.; Syvänen, A.C. Silhouette scores for assessment of SNP genotype clusters. *BMC Genomics* **2005**, *6*, 35. [[CrossRef](#)]
40. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Anna. Stat.* **1979**, *7*, 1–26. [[CrossRef](#)]
41. Shannon, C.P.; Chen, V.; Takhar, M.; Hollander, Z.; Balshaw, R.; McManus, B.M.; Tebbutt, S.J.; Sin, D.D.; Ng, R.T. SABRE: A method for assessing the stability of gene modules in complex tissues and subject populations. *BMC Bioinform.* **2016**, *17*, 460. [[CrossRef](#)]
42. Kang, G.; Liu, W.; Cheng, C.; Wilson, C.L.; Neale, G.; Yang, J.J.; Ness, K.K.; Robison, L.L.; Hudson, M.M.; Srivastava, D.K. Evaluation of a two-step iterative resampling procedure for internal validation of genome-wide association studies. *J. Hum. Genet.* **2015**, *60*, 729–738. [[CrossRef](#)]
43. Triba, M.N.; Le Moyec, L.; Amathieu, R.; Goossens, C.; Bouchemal, N.; Nahon, P.; Rutledge, D.N.; Savarin, P. PLS/OPLS models in metabolomics: The impact of permutation of dataset rows on the K-fold cross-validation quality parameters. *Mol. BioSyst.* **2015**, *11*, 13–19. [[CrossRef](#)]
44. Gijssberts, C.M.; Seneviratna, A.; Bank, I.E.M.; Den Ruijter, H.M.; Asselbergs, F.W.; Agostoni, P.; Remijn, J.A.; Pasterkamp, G.; Kiat, H.C.; Roest, M.; et al. The ethnicity-specific association of biomarkers with the angiographic severity of coronary artery disease. *Neth. Heart J.* **2016**, *24*, 188–198. [[CrossRef](#)] [[PubMed](#)]
45. Chanock, S.J.; Manolio, T.; Boehnke, M.; Boerwinkle, E.; Hunter, D.J.; Thomas, G.; Hirschhorn, J.N.; Abecasis, G.; Altshuler, D.; Bailey-Wilson, J.E.; et al. Studies N-NWGoRiA. Replicating genotype–phenotype associations. *Nature* **2007**, *447*, 655–660. [[CrossRef](#)] [[PubMed](#)]
46. Gallagher, M.D.; Chen-Plotkin, A.S. The Post-GWAS Era: From Association to Function. *Am. J. Hum. Genet.* **2018**, *102*, 717–730. [[CrossRef](#)] [[PubMed](#)]
47. Allyse, M.A.; Robinson, D.H.; Ferber, M.J.; Sharp, R.R. Direct-to-Consumer Testing 2.0: Emerging Models of Direct-to-Consumer Genetic Testing. *Mayo Clin. Proc.* **2018**, *93*, 113–120. [[CrossRef](#)]
48. Wu, Y.; Perng, W.; Peterson, K.E. Precision nutrition and childhood obesity: A scoping review. *Metabolites* **2020**, *10*, 235. [[CrossRef](#)]
49. Baker, M. 1500 scientists lift the lid on reproducibility. *Nature* **2016**, *533*, 452–454. [[CrossRef](#)]
50. Fanelli, D. Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 2628–2631. [[CrossRef](#)]
51. Fiehn, O.; Robertson, D.; Griffin, J.; Van der Werf, M.; Nikolau, B.; Morrison, N.; Sumner, L.W.; Goodacre, R.; Hardy, N.W.; Taylor, C.; et al. The metabolomics standards initiative (MSI). *Metab. Off. J. Metab. Soc.* **2007**, *3*, 175–178. [[CrossRef](#)]
52. Sumner, L.W.; Amberg, A.; Barrett, D.; Beale, M.H.; Beger, R.; Daykin, C.A.; Fan, T.W.; Fiehn, O.; Goodacre, R.; Griffin, J.L.; et al. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metab. Off. J. Metab. Soc.* **2007**, *3*, 211–221. [[CrossRef](#)]
53. Ferreira, J.D.; Inácio, B.; Salek, R.M.; Couto, F.M. Assessing Public Metabolomics Metadata, Towards Improving Quality. *J. Integr. Bioinform.* **2017**, *14*. [[CrossRef](#)]
54. Inácio, B.; Ferreira, J.D.; Couto, F.M. (Eds.) Metadata analyser: Measuring metadata quality. In Proceedings of the 11th International Conference on Practical Applications of Computational Biology & Bioinformatics, Porto, Portugal, 21–23 June 2017; Springer International Publishing: Cham, Switzerland, 2017.

55. Shin, S.Y.; Fauman, E.B.; Petersen, A.K.; Krumsiek, J.; Santos, R.; Huang, J.; Arnold, M.; Erte, I.; Forgetta, V.; Yang, T.P.; et al. An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **2014**, *46*, 543–550. [[CrossRef](#)] [[PubMed](#)]
56. Krumsiek, J.; Suhre, K.; Evans, A.M.; Mitchell, M.W.; Mohny, R.P.; Milburn, M.V.; Wägele, B.; Römisch-Margl, W.; Illig, T.; Adamski, J.; et al. Mining the unknown: A systems approach to metabolite identification combining genetic and metabolic information. *PLoS Genet.* **2012**, *8*, e1003005. [[CrossRef](#)] [[PubMed](#)]
57. Krumsiek, J.; Mittelstrass, K.; Do, K.T.; Stückler, F.; Ried, J.; Adamski, J.; Peters, A.; Illig, T.; Kronenberg, F.; Friedrich, N.; et al. Gender-specific pathway differences in the human serum metabolome. *Metab. Off. J. Metab. Soc.* **2015**, *11*, 1815–1833. [[CrossRef](#)] [[PubMed](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).