Special Issue: Computation and Modeling

Opinion

Why Batch Effects Matter in Omics Data, and How to Avoid Them

Wilson Wen Bin Goh,[1,2,*] Wei Wang,[1] and Limsoon Wong[2,3,*]

Effective integration and analysis of new high-throughput data, especially gene-expression and proteomic-profiling data, are expected to deliver novel clinical insights and therapeutic options. Unfortunately, technical heterogeneity or batch effects (different experiment times, handlers, reagent lots, etc.) have proven challenging. Although batch effect-correction algorithms (BECAs) exist, we know little about effective batch-effect mitigation: even now, new batch effect-associated problems are emerging. These include false effects due to misapplying BECAs and positive bias during model evaluations. Depending on the choice of algorithm and experimental set-up, biological heterogeneity can be mistaken for batch effects and wrongfully removed. Here, we examine these emerging batch effect-associated problems, propose a series of best practices, and discuss some of the challenges that lie ahead.

Trends

Effectively dealing with batch effects will be the next frontier in large-scale biological data analysis, particularly involving the integration of different data sets.

Given how batch-effect correction exaggerates cross-validation outcomes, cross-validation is becoming considered a less authoritative form of evaluation.

Batch effect-resistant methods will become important in the future, alongside existing batch effect-correction methods.

When Nonbiological Effects Confound

Identifying **differential** (see Glossary) effects between **sample classes** (e.g., normal versus disease) in high-throughput biological data is essential for developing gene signatures for diagnostics and prognostics and for providing potential drug targets. These effects are typically identified via statistical testing for **relevant** biological features (such as genes or proteins), which requires identifying real signals against a complex backdrop of inter- and intrasample variation (Figure 1).

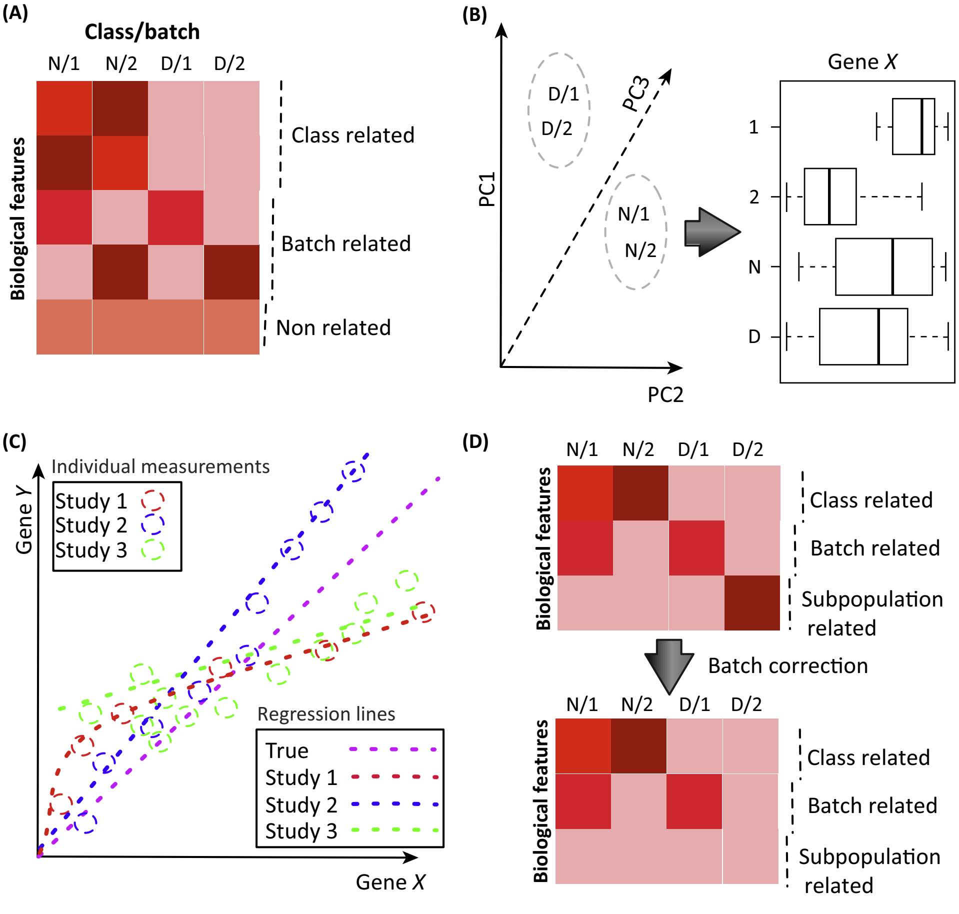
However, some sources of variation are unrelated to inter- and intrasample class differences. These are summarily termed '**batch effects**' and arise from, for example, different experiment times, handlers, reagents, and instruments [1]. Batch effects are pervasive and pertinent to all types of not only high-throughput biological platform, but also low-throughput methods, such as PCR and western blots. Although high-throughput technologies provide higher scalability and coverage, obtaining sufficiently large data sets suitable for clinical analysis requires multiple handlers, reagent batches, machines, segregated running times, and so on. Moreover, the burden of data generation is usually borne among several collaborating laboratories. In practice, batch effects are almost inevitable.

Circumventing batch effects via **meta-analysis** is not foolproof: the relative amount of data in each batch is smaller, implying low statistical **power**; that is, not all relevant features are detectable, especially those with small **effect sizes**. Extrapolating this idea suggests that the pooled result of several independent smaller analyses is not equivalent to that of a single large study. Integrating several smaller data sets theoretically boosts power and better reflects the underlying population. However, proper integration requires resolving technical heterogeneity,

[1]School of Pharmaceutical Science and Technology, Tianjin University, Tianjin 300072, P.R. China
[2]Department of Computer Science, National University of Singapore, Singapore 117417, Republic of Singapore
[3]Department of Pathology, National University of Singapore, Singapore 119074, Republic of Singapore

*Correspondence:
wilson.goh@tju.edu.cn,
goh.informatics@gmail.com
(W.W.B. Goh),
wongls@comp.nus.edu.sg (L. Wong).

http://dx.doi.org/10.1016/j.tibtech.2017.02.012



Trends in Biotechnology

Figure 1. Identifying Real Signals against a Complex Backdrop of Inter- and Intrasample Variations. (a) Biological features can be broadly categorized as those associated with class effects (class-related), batch effects (batch-related), or completely neutral (non-related). (b) Despite the ostensible resolution of batch effects, as shown in the principal-component scatterplot (i.e., same-class samples aggregate together), individual features may still exhibit strong batch correlations. In this case, a top-ranked feature has stronger correlation with batch (1 and 2) than with class (N and D). (c) Meta-analyses do not always agree with each other and can obscure the true picture. Here, we show the correlation between two features, genes *X* and *Y*, across studies 1–3. The dotted colored lines correspond to the regression for each study, which suggests different relationships between *X* and *Y*. The purple dotted line is the true relationship, observable only by combining all three studies. (d) Subpopulation effects, if not properly protected, can be lost via batch-effect correction. Information on disease subpopulations can be critical; for example, in breast cancer, patients are allotted different treatments based on whether they express the estrogen receptor. Since certain drugs target the estrogen receptor, patients who do not express the receptor will not benefit from the drug, yet have to suffer its potential adverse effects.

including batch effects. Given the large amount of data being generated daily (and the need to consider it synergistically instead of independently), we foresee that batch-effect mitigation will become the next big challenge in mega (biological) data analysis.

However, how do batch effects confound? In statistical **feature selection**, batch-correlated variation reduces power, with concomitant non-**reproducibility** (similar experiments conducted on similar samples yield dissimilar batch effects). Furthermore, any **classifier** built using batch-correlated features does not **generalize** and fails subsequent independent evaluations (and as a potential diagnostic indicator). Proper removal of batch-correlated variation can remedy this, as demonstrated in both genomics [2] and proteomics [3]. Note that, unless specified, our arguments are derived mostly from gene expression-based studies.

Glossary

Batch-effect correction: a data-cleaning approach where batch effects are estimated and removed from the data.
Batch effects: technical sources of variation, such as different processing times or different handlers, which may confound the discovery of real explanatory variables from data.
Biological network: a complex system comprising functional relationships among various biological entities. Almost all networks are abstractions, and commonly studied ones include protein–protein interaction, signaling, and metabolic networks.
Class effects: biological sources of variation that provide discrimination between two phenotypes or classes, such as normal versus disease.
Classifier: a trained system that has established a set of classification rules derived from a set of predictor variables. Given observed values of the predictor variables for a sample, the classifier uses its classification rules to determine a class assignment for the sample.
Confounded: mixed together; difficult to disambiguate. Possible to express quantitatively, for example, batch and class effects are perfectly confounded if all class A samples are in batch 1 and all class B samples are in batch 2.
Cross-validation: an evaluative technique to infer how the outcome from a statistical analysis in one instance may generalize to another, provided that the instances are sampled subsets derived from one original dataset.
Differential: a continuous measurement indicating the extent of difference between one variable in one class relative to one in another class.
Effect size: the magnitude of difference for a given variable between two classes.
False negative: a test result that declares a feature as insignificant when in fact it is significant.
False positive: a test result that declares a feature as significant when in fact it is not.
Feature selection: a statistical technique for simplifying models by identifying and including only the most relevant variables among all others being measured.

The most common way to deal with batch effects is to remove them via BECAs. Although developed and benchmarked on microarrays in 2008, ComBat remains the most popular BECA due to its perceived high performance [4], and it is applied across a variety of platforms, including next-generation sequencing and proteomics, sometimes disregarding its limitations [2,5]. Currently, BECAs are considered platform independent, but they do have operating and sample-size requirements (see below).

Unfortunately, we overestimate our ability to mitigate batch effects, and new problems are emerging (Box 1): incomplete removal of batch effects or botched removal attempts given uneven sample-class representation across batches produce adverse effects [6]. New evidence suggests that batch effects specifically inflate **cross-validation** accuracy, which is important because cross-validation is the *de facto* standard for classifier evaluation in the absence of **independent validation** data sets.

Finally and critically, batch effects may obscure and/or confound biologically important **subpopulation** effects [7]. Subpopulations are typically under-represented in data and hard to isolate or identify. However, they represent differential etiologies and, therefore, warrant different treatment strategies [5,8]. Subpopulation effects (which are clinically valuable) may resemble batch effects, and **batch-effect correction** can unintentionally remove the valuable subpopulations effects.

Do Not Trust Batch-Effect Correction When Batch-Group Design Is Unbalanced

It is well appreciated that batch effects, as unwanted variation, reduce statistical power [6]. However, it is less well known that, if the batch-group design is unbalanced (i.e., sample

Box 1. How Do Batch Effects Confound?

Without Batch-Effect Correction

False positives and false negatives: the presence of batch-correlated variation skews analysis in two ways. By increasing ambient levels of uncertainty, the magnitude of any test statistic is reduced, making it harder for a real biological feature to be reported as differential (**false negative**). Alternatively, increased variation may also sporadically introduce false effects, such that an irrelevant biological feature is wrongly reported as differential (false positive).

With Batch-Effect Correction

When batch-group design is unbalanced: depending on the batch effect-correction method used, the false-negative rate is increased if batch effects are underestimated, and the false-positive rate is increased if the class effect is overestimated [6,9].

When incompletely removed: this usually occurs when batch and class effects are highly confounded. Removing batch-correlated variation inadvertently increases false negatives. The presence of remnant batch effects also leads to erroneous estimations of effect size.

When inappropriately removed: when a batch-effect correction method makes incorrect assumptions about the structure of the data or when the operating conditions of the algorithm are not met (e.g., using two-way ANOVA on small data sets), then both false positives and false negatives are produced.

Losing important information: both biological and technical heterogeneity can be removed via batch-effect correction. Technical heterogeneity leads to nonbiological artifacts, aka batch effects. By contrast, biological heterogeneity stems from natural subpopulations or subtypes. Removing this heterogeneity may bias or overemphasize interclass differences, creating less-robust classifiers. Functional studies of batch effect-corrected data sets can potentially mislead, because overemphasis of interclass differences can inflate the estimated effect size or significance (i.e., the relative importance) of a given set of biological features.

Overoptimistic cross-validation accuracy: uncorrected batch effects, even if present at ambient levels, are sufficient to drive optimism bias in cross-validation evaluation.

Generalize: make universally applicable. Here, 'generalizable' means that the findings in one study also apply to another.
Independent validation: an evaluative technique used to infer how the outcome from a statistical analysis in one instance may generalize to another, provided that the instances are not derived from one original dataset but are independently obtained.
Meta-analysis: an analytical technique where data from each laboratory or each batch is analyzed independently, with the expectation that they lead towards mutually supportive findings.
Positive bias: the tendency to overestimate the likelihood of an outcome.
Power: the tendency for a statistical test to detect a true effect.
Relevant: a qualitative variable indicating whether a feature is pertinent towards differentiating one class from another.
Reproducibility: the tendency to produce the same findings given independent resamplings from the same reference populations.
Sample classes: groups of phenotypically equivalent samples, such as normal or disease.
Subpopulation: genetic variants of the same disease (subtypes); also referred to as biological heterogeneity.

classes are unevenly represented across batches), attempts to remove batch effects can mislead [6,9].

Consider a simple approach that calculates and subtracts the mean from all measurements in the same batch (zero-centering). If the batch-group design is balanced, this approach removes most variation attributable to the batch effect, increasing statistical power. However, if unbalanced, the batch variation is underestimated, and corrected data still retain batch variation, reducing the statistical power.

A potential mitigation strategy is when both class and batch effects are estimated, and only the batch effects are removed. In two-way ANOVA, **class effects** are estimated within batches and applied to the batch effect-corrected data. When the batch-group design is balanced, class and batch effects are independent. When unbalanced, class-effect estimations depend on the intrabatch proportions and are variable between batches. In other words, the batch-effect estimation error now also depends on the class proportions within each batch. This phenomenon has myriad consequences: if the batch effect is stronger than the estimation variability, then correcting for batch effects is still better than no correction at all. However, if the batch effect is subtle, then this approach leads to exaggerated intersample class effects.

This issue appears to be independent of sample size and dependent solely on batch-group design imbalances [6]. However, there is a silver lining: the problem is not expected to be severe if the data set is only moderately imbalanced. It should further be noted that the severity also depends on post-hoc corrections, such as certain multiple-testing corrections, which are susceptible to strong outlier effects.

Batch Effects Cause Strong Bias in Cross-Validations

Cross-validation is an established evaluation approach for determining the performance of a classifier in the absence of an independent validation dataset. It involves repeatedly subsplitting the original data set into training and validation data sets. The split does not need to be equal; for example, in a fivefold cross-validation, 80% of the data is used for training and 20% for validation in each split.

In each split, statistical feature selection produces a list of statistically differential features, which are used to train the classifier until it learns how to distinguish sample classes in the corresponding training data set. It is then evaluated with the corresponding validation data set to generate an accuracy score that indicates how well it can distinguish sample classes given that validation data set. Following *n* iterations, *n* accuracy scores are generated. A centrality measure, such as the arithmetic mean, can be used as a point estimate of the overall accuracy of the classifier that would be produced when the same feature-selection and classifier-training methods were applied to the entire data set.

Soneson et al. demonstrated that established tools, such as **cross-validation**, exhibit **positive bias** when batch effects are present and **confounded** with class effects. This bias can not be mitigated by batch-effect correction [10]. Moreover, when batch and class effects are confounded, many of the highly ranked features are correlated only with the batches and are irrelevant to the disease population. Obviously, any classifier built using mostly batch-correlated features will not generalize.

The approach of Soneson et al. is based on simulated data (for both class and batch effects), and it is possible that the outcome is perhaps simply an artifact of batch-effect simulation. However, Qin et al. designed a data set with real batch effects. Specifically, this was a pair of data sets generated from the same set of tumor samples, where the first and second data sets were collected with and without uniform handling, respectively [11]. In the presence of batch effects, cross-validation always tended towards overly optimistic estimates of the error rate, irrespective of batch-effect correction [11]. This result agrees with the simulation experiments and further suggests that even subtle batch effects are sufficient to generate misleading cross-validation results.

Both the studies involving simulated and real data support the notion that cross-validations exhibit positive bias and cannot replace independent validation. In addition, although cross-validation is an established and frequently used evaluation method, caution should be exercised when interpreting cross-validation results.

Batch Effects and Hidden Subpopulations

Batch effects are often confounded with hidden biological subpopulation heterogeneity. Suppose that a doctor requests an analyst to build a classifier to distinguish sample classes A and B. The doctor may not know in advance or may fail to indicate that class A has subtypes A_1, A_2, . . . A_x. This may cause difficulty in training the classifier, or worse, may mistrain the classifier. In this regard, the hidden subpopulation heterogeneity is similar to a batch effect (where each subpopulation could be seen as akin to a batch).

However, while batch effects are not biologically meaningful, subpopulation effects are important (i.e., different subpopulations may have different genetic profiles, warranting different therapeutic interventions) and should be preserved. Current BECAs may not always make this distinction. In addition, analysts may not realize that they are losing vital information by removing subpopulation effects with the use of popular tools, such as surrogate variable analysis (SVA), where the estimated surrogate variables (or batch factors) may include subpopulation information [8]. Moreover, removing subpopulation effects (when they are not equally represented between different data sets or batches) can also exacerbate false effects in a manner not unlike unbalanced batch-group design.

How biological heterogeneity confounds BECAs is multifaceted. Mostly, this depends on the assumptions of the algorithm (Box 2). In SVA [7], both batch and class factors are specifiable, and the algorithm has proven effective at protecting specified class factors (emphasizing the protection of class-differentiating variation), to the extent that most other variation is lost [8]. This methodology has several consequences: first, as intra-class variability is reduced, SVA increases the statistical power, but, second, this comes at the cost of losing many individual sample-specific idiosyncrasies (including subpopulation information). It is also possible to specify both batch and class factors to ComBat, with a concomitant increase in power [8]. However, in ComBat, specifying biological factors is not an absolute requirement and, without it, post-correction *P* values will become generally lower across the board, suggesting that **false positives** become more likely.

It is possible to use existing BECAs and preserve subpopulation effects. If subpopulation factors (e.g., disease subtypes) are known *a priori*, they can be specified as factors in addition to the class factor. If unknown, they can be inferred through an exploratory analysis of the data, a literature search, or the surrogate variable estimation of SVA. It is critical that the analyst understands that the consequences of batch-effect correction depend largely on the specified class and batch factors. Different specified classes, subpopulations, and batch factors in turn provide different insights. Analysis following batch-effect correction is not a one-off process and may require building several different batch effect-corrected data sets with different factor specifications to understand the relative impact and contribution of each factor. Given that power is artificially increased (as a consequence of reduction of intrasample variation), Jaffe et al. importantly pointed out that any differential analysis of batch effect-corrected data is

Box 2. A Short Introduction to Batch Effect-Correction Methods and How to Choose Them

Different Classes of Batch Effect-Correction Algorithms

Simple linear models: a biological feature is modeled as a linear combination of class and batch effects. Methods such as mean-scaling and zero-centering fall under this class.

Empirical Bayes: batch effects of all biological features in a batch are estimated via Bayesian inference. The popular ComBat [4] method uses this approach.

Factor-based analysis: the above two approaches require prespecifying all known batch factors (e.g., the time of experiment); otherwise, batch-effect estimation is unreliable. The full data matrix may be used for estimating batch-related against class-related variation. Several different methods exist in this class; two are described below.

SVA, another widely used method, first requires specifying the class factor and assumes that consistent sources of variation not associated with the class factor are likely associated with some unknown batch factor. Having identified the part of the data associated with batch variation, this method can then be used to estimate the batch effect via singular value decomposition [7]. The batch effect can then be removed from data via regression.

Removed unwanted variation (RUV) is similar to SVA, but it incorporates information about biological invariants [19]. These invariants are taken as housekeeping genes, which are expected to be unaffected by class effects. Thus, they can be used to estimate batch effects. However, specifying housekeeping genes can be controversial [8] and, in some cases, so-called ‘housekeeping genes’ are directly related to disease [20].

How to Choose a BECA

Moderate to large data sets with limited feature space; limited biological heterogeneity; batch and class factors are known; goal is simple exploratory analysis → Two-way ANOVA [6].

Small data sets with limited feature space; limited biological heterogeneity; batch and class factors are known; goal is batch-effect removal → ComBat [4].

Moderate to large data sets with large feature space; existent biological heterogeneity; class factors are known (batch factors not necessarily known); goals are batch-effect removal and batch-factor determination → SVA [7] and RUV [19].

Moderate to large data sets with large feature space; existent biological heterogeneity; batch and class factors not necessarily known; goals are batch-effect removal and batch-factor determination but not to characterize class effect (e.g., simple data integration) → unsupervised methods (e.g., principal component analysis; PCA) can be used if batch variation is saturated in the top *n* principal components [5], RUV can also be used [15].

inflated (in terms of the estimated effect sizes and their associated *P* values) and should be treated with caution [8].

There are also approaches for automatically differentiating class effects and batch effects from data. For example, permuted SVA (pSVA) iteratively refines sample clusters after modeling the contribution from a known batch effect [12]. This algorithm removes technical artifacts in replicate samples while retaining biological heterogeneity in samples. However, this algorithm has not yet achieved widespread adoption.

Given the issues above, we may yet conclude that the traditional two-stage batch effect-correction approach, where batch effects are first estimated and removed from the data, followed by routine analysis (e.g., feature selection and model building)l is cumbersome and difficult to incorporate in routine analysis without a highly qualified analyst. Therefore, work-arounds are needed, particularly approaches that are naturally resistant to batch variation, do not alter data integrity, and preserve biological heterogeneity.

Network-Based Feature-Selection Methods May Be Useful

In practice, it is difficult to completely eradicate batch-effect thinking that all batch factors are known in advance. Although it is possible to estimate batch factors computationally, biological

Box 3. How Gene Fuzzy Scoring Works

In gene fuzzy scoring (GFS), an expression matrix (organizing biological features by sample) is transformed by weighting the individual features per sample based on sample-wise expression ranks. Here, $r(g_i, p_j)$ is the rank of a biological feature g_i in patient p_j, and $q(p_j, \theta)$ is the rank corresponding to the upper θth level of feature ranks in p_j. The GFS score s (g_i, p_j) assigned to feature g_i for patient p_j is determined by Equation [I]:

$$s(g_i,p_j)=\begin{cases} 1 & ,\textit{if } q(p_j,\theta_2)<r(g_i,p_j) \\ \dfrac{r(g_i,p_j)-q(p_j,\theta_2)}{q(p_j,\theta_1)-q(p_j,\theta_2)} & ,\textit{if if } q(p_j,\theta_1)>r(g_i,p_j)\geq \mathrm{q}(p_j,\theta_2) \\ 0 & ,\textit{otherwise} \end{cases} \qquad [\mathrm{I}]$$

GFS has the benefit of eliminating a large amount of non-useful variation while boosting high-confidence signal. Features are ranked and assigned new values based on which category they fall under: those above θ_1 (high confidence), between θ_1 and θ_2 (moderate confidence), and below θ_2 (low confidence). High-confidence features are assigned a weight of 1, while low-confidence features are considered noisy and penalized by a weight assignment of 0. Moderate-confidence features can be assigned a value between 0 and 1 via linear interpolation.

As arbitrarily defined thresholds, θ_1 and θ_2 can take on different values, for example, the default setting in GFS sets θ_1 to 5% and θ_2 to 15%, but altering these thresholds does not impact analysis significantly. Belorkar and Wong compared GFS against other normalization techniques, such as mean scaling, z- and quantile-normalization, and found that GFS boosts class effect even when sampling at small sample sizes, and importantly, is robust against batch effects.

and technical heterogeneity cannot be easily differentiated, and it is not possible to determine whether batch-effect correction inadvertently compromises data integrity.

There is new evidence that powerful data normalization approaches (e.g., Gene Fuzzy Scoring or GFS [13]; Box 3) coupled with network-based analysis techniques (e.g., FSNET or PFSNET [14]) are robust against batch effects without having to explicitly estimate and remove them from the data. Thus, the data integrity is not compromised.

GFS is a signal-boosting technique that eliminates noise and confounding variability by penalizing low-intensity features and equalizing high-intensity features. Coupled with **biological network**-based methods, such as FSNET and PFSNET [14], that incorporate biological constraints by using highly coordinated subnets (e.g., protein complexes), these methods exhibit high power with high reproducibility (even with very small sample sizes). Given that a significant subnet feature may comprise both high- and low-abundance proteins, these methods can also recover, via guilt-by-association, low-abundance proteins lost by GFS. However, it is not well known that these methods are also resistant to batch effects. As far as proteomics is concerned, when tested for correlations with batch and class effects, top network-based features were strongly associated with class but not batch effects, while individual protein features selected by parametric statistical approaches were strongly correlated with batch effects [5]. There is no reason to think that the results will not hold true in a gene-expression context. It is also important to note that not every network-based approach is batch-effect resistant; the scoring and normalization approaches matter. For example, the hypergeometric enrichment approach, which evaluates subnets for enrichment of differential proteins given the Fisher's exact test, does not work because its test statistic remains highly correlated with batch effects [5].

Challenges and Best Practices

Better BECAs may not be the answer: batch effects (and current solutions) are multifaceted (see [15] and Box 2), and it is difficult to predict their severity or attributes. In practical applications, it is impossible to truly know whether a batch effect has been properly negated and whether the data integrity has been preserved without introducing false effects. Where both class and batch effects are deeply confounded, it is difficult to remove batch effects without lowering statistical power.

Thus, we assert that proper experimental design is crucial: batch effects cannot be easily dealt with analytically when batch sizes are unequal and/or sample classes are unevenly distributed across batches. When study groups are not evenly distributed across batches, actual group differences may induce apparent batch differences; in such cases, batch-effect correction may positively bias class differences [16]. Blocking designs, which minimize the impact of known confounding factors (e.g., age distribution of both sample classes are kept similar), should always be incorporated where possible [17]. Platform-specific technical bias, which can be modeled reliably, should also be used where possible (e.g., mixed-effect modeling for multiplexed bead-based flow cytometric immunoassays [18]).

If an experiment is performed with an unbalanced design, it may be possible to identify and account for unresolved batch effects statistically. If this statistical analysis is not feasible, the outcome of any batch-effect correction should be regarded with suspicion and treated as purely exploratory, bearing in mind that it is probably unreliable. Alternatively, one may apply analytical methods that are batch-effect resistant. It also may be possible to estimate the severity of the problem indirectly via logical deduction. For example, the experimenter can

Box 4. A Guide to Best Practices for Dealing with Batch Effects

Use visualization to check for batch effects, but be mindful. PCA-based scatterplots and hierarchical clustering (HCL)-based dendrograms or heatmaps are regularly used to visualize or check the presence or absence of batch effects. Note that the apparent lack of a batch effect based on macroscopic visual cues does not imply that one is absent: Leek et al. make further checks at the level of a group of known genes susceptible to batch effects [1], while we advocate the use of side-by-side univariate scatterplots of principal components derived from selected or individual biological features [5].

Normalize data properly before using batch-effect correction, such as by performing quantile normalization [21] or GFS [13] on data first followed by ComBat [4]. However, this approach is known not to work well if sample classes are not properly balanced across batches.

Proper experimental design is paramount. Distribute sample classes across batches evenly or risk erroneous batch-effect estimation [6].

When sample sizes are small, use empirical Bayes (e.g., ComBat [4]) instead of ANOVA-based methods [6].

When the feature space is small, avoid methods that are highly reliant on multivariate information. Methods such as SVA use variation from a large amount of measured features to estimate batch effects. Therefore, their performance may suffer on platforms where only a limited number of features can be measured, such as target-dependent acquisition proteomics or multiplexed RT-PCR.

Check sources of unexplained variation but do not treat them all as artifacts. It is unlikely that all batch effects are explainable by experimental procedures. In practice, there are likely many sources of batch variation, and what we think is the most pertinent source of batch effects may not prove to be so. Factor-based methods (e.g., SVA) do provide a means of estimating various sources of batch effects, such as surrogate variables. However note that not all surrogate variables are artifacts: some are biologically relevant (e.g., disease subtypes). In other words, do not simply use all estimated surrogate variables and remove them as if they are all batch-correlated because some of them may be biologically meaningful.

Do not naively trust batch-effect correction. Following batch-effect correction, although samples may segregate by classes, ambient batch effects may still remain, resulting in false positives if differential analysis is performed [1]. Alternatively, it is important to remember that batch-effect correction is dependent on the specified batch and class factors, which can strongly affect downstream analysis. For example, if all associated batch variations are removed, class effects are amplified, resulting in misleading perception of the effect size for each variable while subpopulation information is lost.

Understand the limits of your batch effect-correction algorithm (or at least know how it roughly works). There are many different types of batch effect-correction algorithm; it is important to understand how they work and determine which is suitable given the experimental question being asked (see also Box 2, main text).

check the variation and effect sizes of known genes associated with the phenotype as positive checks as well as those of established non-related genes, such as certain housekeepers, as negative checks. Finally, it may be possible to take advantage of ontologies and construct a profile of functional terms associated with the phenotype using gold-standard relevant genes. If the constructed functional profile from differential features deviates strongly, then the selected features may be correlated with non-class factors.

Feature-selection algorithms identify meaningful variables associated with disease among all of the other irrelevant variables (e.g., selecting differential proteins from all of the measured proteins). This procedure is important for developing drug targets and diagnostic signatures. When developing new feature-selection algorithms, it is good to also test their resilience against induced ‘batch’ components, which are likely present in real data [5].

Given that batch effects induce positive bias in cross-validations, we recommend avoiding the use of cross-validation where possible. If cross-validation is unavoidable (e.g., if an independent validation data set is not available), then it is crucial to avoid overoptimistic interpretations (Box 4).

Concluding Remarks and Future Perspectives

The disruptive nature of batch effects on clinical and/or translational applications has historically been underappreciated. Although BECAs exist, they are imperfect solutions with specific requirements and cannot replace good experimental design and analysis techniques.

Even a small batch effect is sufficient to skew cross-validation in the overly optimistic direction. Given that it is impossible to truly be sure that all batch variations have been eliminated, cross-validation cannot replace independent-data validation as a standard for classifier evaluation and feature-selection evaluation.

Estimated sources of batch effects should not be removed carelessly. In real data, subpopulation effects may be confused or confounded with batch effects, and both may be multifactorial. Removing subpopulation effects will affect data integrity (especially when not balanced between data batches) or mislead clinical diagnosis or analysis. However, a thorough exploration and investigation into the downstream effects of incorporating various biological and batch factors can be tedious and complex. In this regard, batch effect-resistant approaches may be appropriate. While BECAs are unlikely to be superseded by batch effect-resistant methods in the near future, it is important nonetheless to understand the idiosyncratic behaviors and limitations of these correction methods.

Effective batch-effect management will become a key obstacle in big biological data analysis (see Outstanding Questions). Without effective management, it is impossible to take advantage of the large amount of information already available and use it synergistically for building better classifiers, performing better functional analysis, and producing clinically useful outcomes.

Outstanding Questions

How can batch and subpopulation effects be distinguished, especially when they are strongly confounded?

Given the multitude of errors and/or biases introduced via batch-effect correction, does it make sense to explicitly try removing batch effects in the first place?

References

1. Leek, J.T. *et al.* (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* 11, 733–739

2. Kupfer, P. *et al.* (2012) Batch correction of microarray data substantially improves the identification of genes differentially expressed in rheumatoid arthritis and osteoarthritis. *BMC Med. Genomics* 5, 23

3. Gregori, J. *et al.* (2012) Batch effects correction improves the sensitivity of significance tests in spectral counting-based comparative discovery proteomics. *J. Proteomics* 75, 3938–3951

4. Johnson, W.E. *et al.* (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8, 118–127

5. Goh, W.W. and Wong, L. Protein complex-based analysis is resistant to the obfuscating consequences of batch effects - a case study in clinical proteomics. BMC Genomics (in press)

6. Nygaard, V. *et al.* (2016) Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* 17, 29–39

7. Leek, J.T. and Storey, J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* 3, 1724–1735
8. Jaffe, A.E. *et al.* (2015) Practical impacts of genomic data 'cleaning' on biological discovery using surrogate variable analysis. *BMC Bioinf.* 16, 372
9. Buhule, O.D. *et al.* (2014) Stratified randomization controls better for batch effects in 450K methylation analysis: a cautionary tale. *Front. Genet.* 5, 354
10. Soneson, C. *et al.* (2014) Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PLoS One* 9, e100335
11. Qin, L.X. *et al.* (2016) Cautionary note on using cross-validation for molecular classification. *J. Clin. Oncol.* 34, 3931–3938
12. Parker, H.S. *et al.* (2014) Preserving biological heterogeneity with a permuted surrogate variable analysis for genomics batch correction. *Bioinformatics* 30, 2757–2763
13. Belorkar, A. and Wong, L. (2016) GFS: Fuzzy preprocessing for effective gene expression analysis. *BMC Bioinf.* 17 (Suppl. 17), 540
14. Goh, W.W.B. and Wong, L. (2016) Advancing clinical proteomics via analysis based on biological complexes: A tale of five paradigms. *J. Proteome Res.* 15, 3167–3179
15. Jacob, L. *et al.* (2016) Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics* 17, 16–28
16. Nygaard, V. *et al.* (2016) Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* 17, 29–39
17. Qin, L.X. *et al.* (2014) Blocking and randomization to improve molecular biomarker discovery. *Clin. Cancer Res.* 20, 3371–3378
18. Clarke, D.C. *et al.* (2013) Normalization and statistical analysis of multiplexed bead-based immunoassay data using mixed-effects modeling. *Mol. Cell. Proteomics* 12, 245–262
19. Gagnon-Bartsch, J.A. and Speed, T.P. (2012) Using control genes to correct for unwanted variation in microarray data. *Biostatistics* 13, 539–552
20. Venet, D. *et al.* (2011) Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.* 7, e1002240
21. Muller, C. *et al.* (2016) Removing batch effects from longitudinal gene expression – quantile normalization plus ComBat as best approach for microarray transcriptome data. *PLoS One* 11, e0156594