





# Harmonization of quality metrics and power calculation in multi-omic studies

Sonia Tarazona<sup>1</sup>, Leandro Balzano-Nogueira<sup>2</sup>, David Gómez-Cabrero<sup>3,4,5,6</sup>, Andreas Schmidt<sup>7</sup>, Axel Imhof <sup>7,8</sup>, Thomas Hankemeier<sup>9</sup>, Jesper Tegnér <sup>3,4,10</sup>, Johan A. Westerhuis <sup>11,12</sup> & Ana Conesa <sup>2,13</sup>✉

Multi-omic studies combine measurements at different molecular levels to build comprehensive models of cellular systems. The success of a multi-omic data analysis strategy depends largely on the adoption of adequate experimental designs, and on the quality of the measurements provided by the different omic platforms. However, the field lacks a comparative description of performance parameters across omic technologies and a formulation for experimental design in multi-omic data scenarios. Here, we propose a set of harmonized Figures of Merit (FoM) as quality descriptors applicable to different omic data types. Employing this information, we formulate the MultiPower method to estimate and assess the optimal sample size in a multi-omics experiment. MultiPower supports different experimental settings, data types and sample sizes, and includes graphical for experimental design decision-making. MultiPower is complemented with MultiML, an algorithm to estimate sample size for machine learning classification problems based on multi-omic data.

<sup>1</sup>Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València, Valencia, Spain. <sup>2</sup>Microbiology and Cell Science Department, Institute for Food and Agricultural Research, University of Florida, Gainesville, FL, USA. <sup>3</sup>Unit of Computational Medicine, Department of Medicine, Solna, Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden. <sup>4</sup>Science for Life Laboratory, Solna, Sweden. <sup>5</sup>Mucosal & Salivary Biology Division, King's College London Dental Institute, London, UK. <sup>6</sup>Navarrabiomed, Complejo Hospitalario de Navarra (CHN), Universidad Pública de Navarra (UPNA), IdiSNA, Pamplona, Spain. <sup>7</sup>Protein Analysis Unit, Biomedical Center, Faculty of Medicine, LMU Munich, Planegg-Martinsried, Germany. <sup>8</sup>Munich Center of Integrated Protein Science LMU Munich, Planegg-Martinsried, Germany. <sup>9</sup>Division Analytical Biosciences, Leiden/Amsterdam Center for Drug Research, Leiden, The Netherlands. <sup>10</sup>Biological and Environmental Sciences and Engineering Division, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. <sup>11</sup>Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands. <sup>12</sup>Department of Statistics, Faculty of Natural Sciences, North-West University (Potchefstroom Campus), Potchefstroom, South Africa. <sup>13</sup>Genetics Institute, University of Florida, Gainesville, FL, USA. ✉email: [aconesa@ufl.edu](mailto:aconesa@ufl.edu)

The genomics research community has been increasingly proposing the parallel measurement of diverse molecular layers profiled by different omic assays as a strategy to obtain comprehensive insights into biological systems<sup>1–4</sup>. Encouraged by constant cost reduction, data-sharing initiatives, and availability of data (pre)processing methods<sup>5–13</sup>, the so-called multi-platform or multi-omics studies are becoming popular. However, the success of a multi-omic project in revealing complex molecular interconnections strongly depends on the quality of the omic measurement and on the synergy between a carefully designed experimental setup and a suitable data integration strategy. For example, multi-omic measurements should derive from the same samples, observations should be many, and variance distributions similar if the planned approach for data integration relies on correlation networks. Frequently, these issues are overlooked, and analysis expectations are frustrated by underpowered experimental design, noisy measurements, and the lack of a realistic integration method.

A thorough understanding of individual omic platform properties and their influence on data integration efforts represents an important, but usually ignored, aspect of multi-omic experiment planning. Several tools are available to assess omic data quality, even cross-platform, such as FastQC for raw sequencing (seq) reads<sup>14</sup>, Qualimap<sup>15,16</sup> and SAMstat<sup>17</sup> for mapping output, and MultiQC<sup>18</sup> that combines many different tools in a single report. However, these tools do not apply to non-sequencing omics (i.e., metabolomics) and are not conceived to compare platform performance nor support multi-omic experimental design choices. Figures of Merit (FoM) are performance metrics typically used in analytical chemistry to describe devices and methods. FoM include accuracy, reproducibility, sensitivity, and dynamic range; descriptors also applied to omic technologies. However, the definition of each FoM acquires a slightly different specification depending on the omic technology considered, and each omic platform possesses different critical FoM. For example, RNA-seq usually provides unbiased, comprehensive coverage of the targeted space (i.e., RNA molecules), while this is not the case for shotgun proteomics, which is strongly biased toward abundant proteins. Importantly, we currently lack both a systematic description of FoM discrepancies across omic assays and a definition of a common performance language to support discussions on the multi-omic experimental design.

FoM are relevant to statistical analyses that aim to detect differential features, due to their impact on the number of replicates required to achieve a given statistical power. The statistical power of an analysis method, which is the ability of the method to detect true changes between experimental groups, is determined by the within-group variability, the size of the effect to be detected, the significance level to be achieved, and the number of replicates (or observations) per experimental group, also known as sample size. All these parameters are highly related to FoM. Estimating power in omic experiments is challenging because many features are assessed simultaneously<sup>19</sup>. These features may have different within-condition variability and the significance level must be adapted to account for the multiple testing scenario. Deciding on the effect size to detect may also prove difficult, especially when the natural dynamic range of the data has changed due to normalization procedures. Moreover, different omic platforms present distinct noise levels and dynamic ranges, and hence analysis methods might not be equally applicable to all of them. As a consequence, independently computing the statistical power for each omic might not represent the best approach for a multi-omic experiment, if the different measurements are to be analyzed in an integrative fashion and a joint power study for all platforms seems more appropriate. Although several methods have been proposed to optimize sample size and evaluate statistical power in

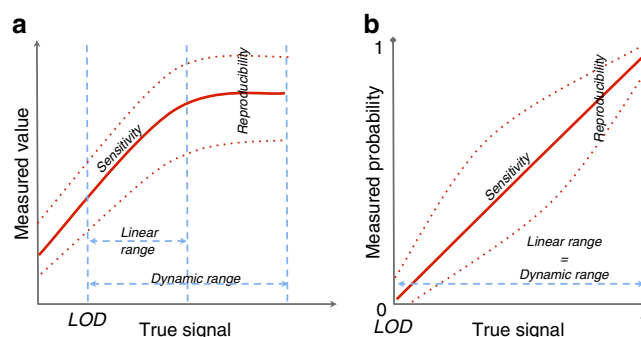
single-omic experiments<sup>19,20</sup>, no such tool exists in the context of multi-omics data. Similarly, multi-omic datasets are increasingly collected to develop sample class predictors applying machine learning (ML) methods. In this case, the classification error rate (ER), rather than the significance value, is used to assess performance. In the field of ML, the estimation of the number of samples required to achieve an established prediction error is still an open question<sup>21,22</sup> and there are not yet methods that answer this question for multi-omics applications.

In summary, the multi-omics field currently lacks a comparative description of performance metrics across omic technologies and methods to estimate the number of samples required for their multiple applications. In this work, we propose a formal definition of FoM applicable across several omics and provide a common language to describe the performance of high-throughput methods frequently combined in multi-omic studies. We leverage this harmonized quality control vocabulary to develop MultiPower, an approach for power calculations in multi-omic experiments applicable to across omics platforms and types of data. Additionally, we present MultiML, an R method to obtain the optimal sample size required by ML approaches to achieve a target classification ER. The FoM definitions, together with the MultiPower and MultiML calculations proposed here constitute a framework for quality control and precision in the design of multi-omic experiments.

## Results

**Comparative descriptions of omic measurement quality.** We selected seven commonly used FoM that cover different quality aspects of molecular high-throughput platforms (see “Methods” section, Fig. 1, Supplementary Table 1). Figure 2 summarizes the comparative analysis for these FoM across seven different types of omic platforms, classified as mass spectrometry (MS) based (proteomics and metabolomics) or sequencing based. In this study, we consider short-read seq-based methods that measure genome variation and dynamic aspects of the genomes, and divided them into feature-based (RNA-seq and miRNA-seq) and region-based methods (DNA-seq, ChIP-seq, Methyl-seq, and ATAC-seq).

Sensitivity is defined as the slope of the calibration line that compares the measured value of an analyte with the true level of that analyte (Fig. 1). For a given platform and feature, sensitivity is the ability of the platform to distinguish small differences in the levels of that feature. Features with low sensitivity suffer from less accurate quantification and are more difficult to be significant by differential analysis methods. In metabolomics platforms, sensitivity primarily depends on instrumental choices, such as the



**Fig. 1 Analytical FoM related to the calibration lines.** **a** Calibration line for omics measuring the levels of the target features (MS platforms and gene-based sequencing platforms). **b** Calibration line for omics not measuring concentrations but finding genomic regions, where a biological event occurs (region-based sequencing platforms).

|   | MS-based  |  | Seq-based   |   |   |
|---|---|--|---|---|---|
|   |   |  | Feature-based   | Region-based  |   |
|   | Proteomics  | Metabolomics   | RNA-seq<br>miRNA-seq  | ChIP-seq<br>Methyl-seq<br>ATAC-seq  | DNA-seq   |
| Sensitivity                                     | Depends on the platform, MS detector, chromatographic column, etc. Usually, sensitivity is higher in targeted approaches                            |  | Depends on the sequencing depth. Features with a high number of reads are more accurately measured and smaller relative changes can be detected | Sensitivity can be described in terms of true positive rate or recall, i.e., the proportion of true sites/regions identified as such, provided a given number of reads in the sequencing output |   |
| Reproducibility                                 | Highly abundant proteins are generally more reproducibly quantified   | Depends mainly on the stability of the platform and can be improved by internal standards and quality control systems. | Better at higher signal levels and improves with sequencing depth   |   | Depends on the read coverage and sequencing error rate. |
| Limit of detection (LOD) and quantitation (LOQ) | Both are compound and platform-specific. Values < LOD are considered missing values.  |  | Depends on the sequencing depth. Higher for shorter features or regions. \ Values < LOD are taken as 0.   |   | Depends on the read coverage                            |
| Linear and dynamic range                        | Dynamic range is usually linear. From 3 (targeted mode) to 4 (untargeted mode) orders of magnitude.   | At 4–5 and 3–4 orders of magnitude for dynamic and linear ranges.  | Dynamic range depends on sequencing depth. Linear range depends on the feature and possible sequencing biases.                                  | Both ranges are between 0 and 1.  |   |
| Selectivity                                     | Better for targeted platforms and improves using selected reaction monitoring.  |  | Improves with sequencing depth. Highly abundant features worsen the selectivity of other features.  |   | Repetitive regions compromise variant detection         |
| Identification                                  | Direct in targeted approaches. For untargeted approaches, identification is critical and depends on comparisons with spectra or compound databases. |  | Usually not critical FoM but hindered by multi-mapping of sequencing reads.   |   |   |
|   | Redundant peptides from protein families are difficult to assign.   | Large number of unknown metabolites is an issue.   |   |   |   |
| Coverage  | Lower for targeted platforms, as feature space is restricted, although the targeted space is most captured.   |  | Depends on the sequencing depth. Complete feature space can potentially be covered at the cost of increasing false positives.                   |   |   |

**Fig. 2 FoM across omic platforms.** The table summarizes critical aspects of each FoM and omic.

chromatographic column type, the mass detector employed, and the application of compound derivatization<sup>23</sup>. Targeted proteomic approaches sample many data points per protein, leading to higher accuracy when compared to untargeted methods<sup>24</sup>. In nuclear magnetic resonance (NMR), no separation takes place

and a low number of nuclei change energy status, leading the detection of only abundant metabolites and lower sensitivity than liquid chromatography (LC)–MS or gas chromatography (GC)–MS. NMR is capable of detecting ~50–75 compounds in human biofluid, with a lower sensitivity limit of 5 μmolar (ref. <sup>25</sup>),

while MS platforms are able to measure hundreds to thousands of metabolites in a single sample. Sensitivity in sequencing platforms depends on the number of reads associated with the feature, a parameter influenced by sequencing depth. Features with an elevated number of reads are more accurately measured, and hence smaller relative changes can be detected. For region-based omics, where the goal is to identify genomic regions where a certain event occurs, the above definition is difficult to apply, and sensitivity is described in terms of true positive rate or recall, i.e., the proportion of true sites or regions identified as such, given the number of reads in the seq output.

Reproducibility measures how well a repeated experiment provides the same level for a specific feature or, when referring to technical replicates, the magnitude of dispersion of measured values for a given true signal. Traditionally, the relative standard deviation (RSD) is used as a measure of reproducibility, with RSD normally differing over the concentration range in high-throughput platforms<sup>26,27</sup>. Generally, reproducibility in sequencing platforms improves at high signal levels, such as highly expressed genes or frequent chromatin-related events. RSD is roughly constant in relation to signal in MS platforms, although in LC-MS the lifetime of the chromatographic column strongly influences reproducibility, leading to small datasets being more reproducible than experiments with many samples. This constraint imposes the utilization of internal standards, quality control samples, and retention time alignment algorithms in these technologies<sup>28</sup>. Moreover, untargeted LC-MS proteomics quantifies protein levels with multiple and randomly detected peptides, each with a different ER, a factor that compromises the reproducibility of this technique. On the contrary, NMR is a highly linear and reproducible technique<sup>29</sup>, and reproducibility issues are associated with slight differences in sample preparation procedures among laboratories. Sequencing methods normally achieve high reproducibility for technical replicates<sup>5</sup> and are further improved as sequencing depth increases. Reproducibility at the library preparation level depends on how reproducible the involved biochemical reactions are. Critical factors affecting reproducibility include RNA stability and purification protocols in RNA-seq<sup>30</sup>, antibody affinity in ChIP-seq<sup>31</sup>, quenching efficiency in metabolomics<sup>31</sup>, and proteolytic digestion and on-line separation of peptides in proteomics<sup>32,33</sup>. Methyl-seq experiments based on the robust *MspI* enzymatic digestion and bisulfite conversion usually are more reproducible data than enrichment-based methods, such as methylated DNA immunoprecipitation (MeDIP)<sup>34</sup>. Finally, reproducibility for DNA variant calling is associated with the balance between read coverage at each genome position and the technology sequencing errors.

The limit of detection (LOD) of a given platform is the lowest detectable true signal level for a specific feature, while the limit of quantitation (LOQ) represents the minimum measurement value considered reliable by predefined standards of accuracy<sup>35</sup>. Both limits affect the final number of detected and quantified features, which in turn impacts the number of tested features and the significance level when correcting for multiple testing. For MS-based methods, LOD and LOQ depend on the platform, can be very different for each compound, and normally require changes of instrument or sample preparation protocol for different chemicals. Additionally, sample complexity strongly affects LOD, as this reduces the chance of detecting low-abundance peptides, while pre-fractionation can reduce this effect at the cost of longer MS analysis time. NMR has usually higher LOD than MS-based methods. Conversely, LOD depends fundamentally on sequencing depth in seq-based technologies, where more features are easily detected by simply increasing the number of reads. However, there also exist differences in LOD across features in sequencing assays. Shorter transcripts and regions usually have

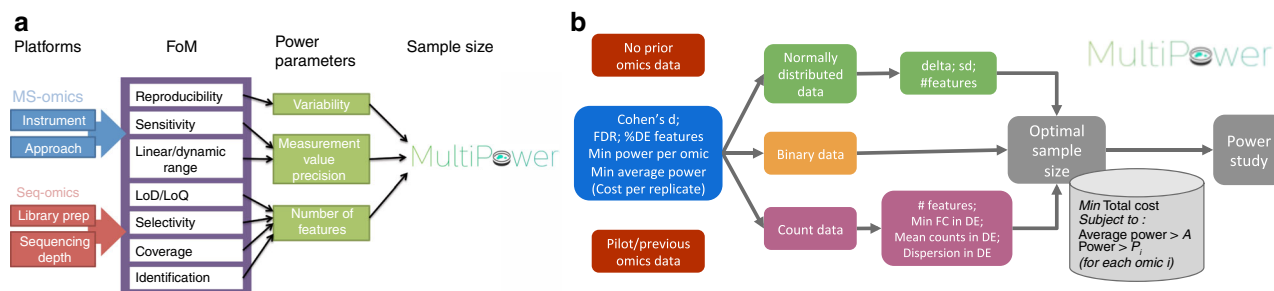
higher LODs and are more affected by sequencing depth choices. For DNA-seq, the ability to detect a genomic variant is strongly dependent on the read coverage. MS-based and seq-based methods also differ in the way features under LOD are typically treated. MS methods either apply imputation to estimate values below the LOD (considered missing values)<sup>12</sup>, or exclude features when repeatedly falling under the LOD. In sequencing methods, LOD is assumed to be zero and data do not contain missing values, although, also in this case, features with few counts in many samples risk exclusion from downstream analyses.

The dynamic range of an omic feature indicates the interval of true signal levels that can be measured by the platform, while the linear range represents the interval of true signal levels with a linear relationship between the measured signal value and the true signal value (Fig. 1). These FoM influence the reliability of the quantification value and, consequently, the differential analysis, as detection of the true effect size depends on the width of these ranges.

In proteomics, molecule fragmentation by data-independent acquisition approaches increases the dynamic range by at least two orders of magnitude. A typical proteomic sample covers protein abundance over 3–4 to four orders of magnitude, a value that increases for targeted approaches<sup>36,37</sup>. In metabolomics, linear ranges usually span 3–4 orders of magnitude, while dynamic ranges increase to 4–5 orders and can be extended using the isotopic peak of the analytes. A combination of analytical methods can increase the dynamic range, as different instruments may better capture either high or low concentration metabolites. NMR has a high dynamic range and can measure highly abundant metabolites with precision, although it is constrained by a high detection limit. For feature-based sequencing platforms, the dynamic range strongly depends on sequencing depth, and values can range from zero counts to up to hundreds of thousands, or even millions (in RNA-seq, for some mitochondrial RNA transcripts). However, due to technical biases, linear range boundaries are difficult to establish and may require the use of calibration RNAs (spike-ins). Moreover, linear ranges may be feature dependent and affected by sequence GC content, which then requires specific normalization. For region-based sequencing platforms, linear and dynamic ranges coincide, and vary from zero to one.

A platform displays good selectivity if the analysis of a feature is not disturbed by the presence of other features. Selectivity influences the number and quantification of features and, consequently, statistical power. For MS platforms, targeted rather than untargeted methods obtain the best selectivity. In metabolomics, selected reaction monitoring, a procedure that couples two sequential MS reactions to obtain unique compound fragments and control quantitation bias using isobaric compounds<sup>38</sup>, is used to improve selectivity, while in proteomics selectivity strongly depends on sample complexity, which determines whether a given feature is detected or not. Similarly, in sequencing experiments, selectivity relates to competition of fragments to undergo sequencing. Highly abundant features or regions (e.g., a highly expressed transcript or a highly accessible chromatin region) may outcompete low-abundant elements. Generally, this problem is difficult to address by means other than increasing the sequencing depth. In the case of transcriptomics, the application of normalized libraries can partially alleviate selectivity problems; however, at the cost of compromising expression level estimations. Particularly for DNA-seq, variant detection is compromised at repetitive regions and can be alleviated with strategies that increase read length.

Identification refers to the relative difficulty faced when determining the identity of the measured feature, a critical issue in proteomics and metabolomics. In principle, identification does



**Fig. 3 The MultiPower approach.** **a** Relationship between omic platforms (MS in blue and sequencing in red), FoM (purple), power parameters (green), and MultiPower. MS- and seq-based omics have different key properties that determine their FoM, which in turn affect differently to power parameters used as input by MultiPower to compute sample size in multi-omics experiments. **b** Overview of the MultiPower method. MultiPower works either with no prior data and with pilot data (red boxes). The blue box contains the parameters to be set by users. Green, orange, and purple boxes represent the data types accepted by MultiPower and the rest of power parameters estimated by MultiPower from prior data when available or given by the user. MultiPower solves an optimization problem to estimate the optimal sample size and to provide a power study of the experiment (gray boxes).

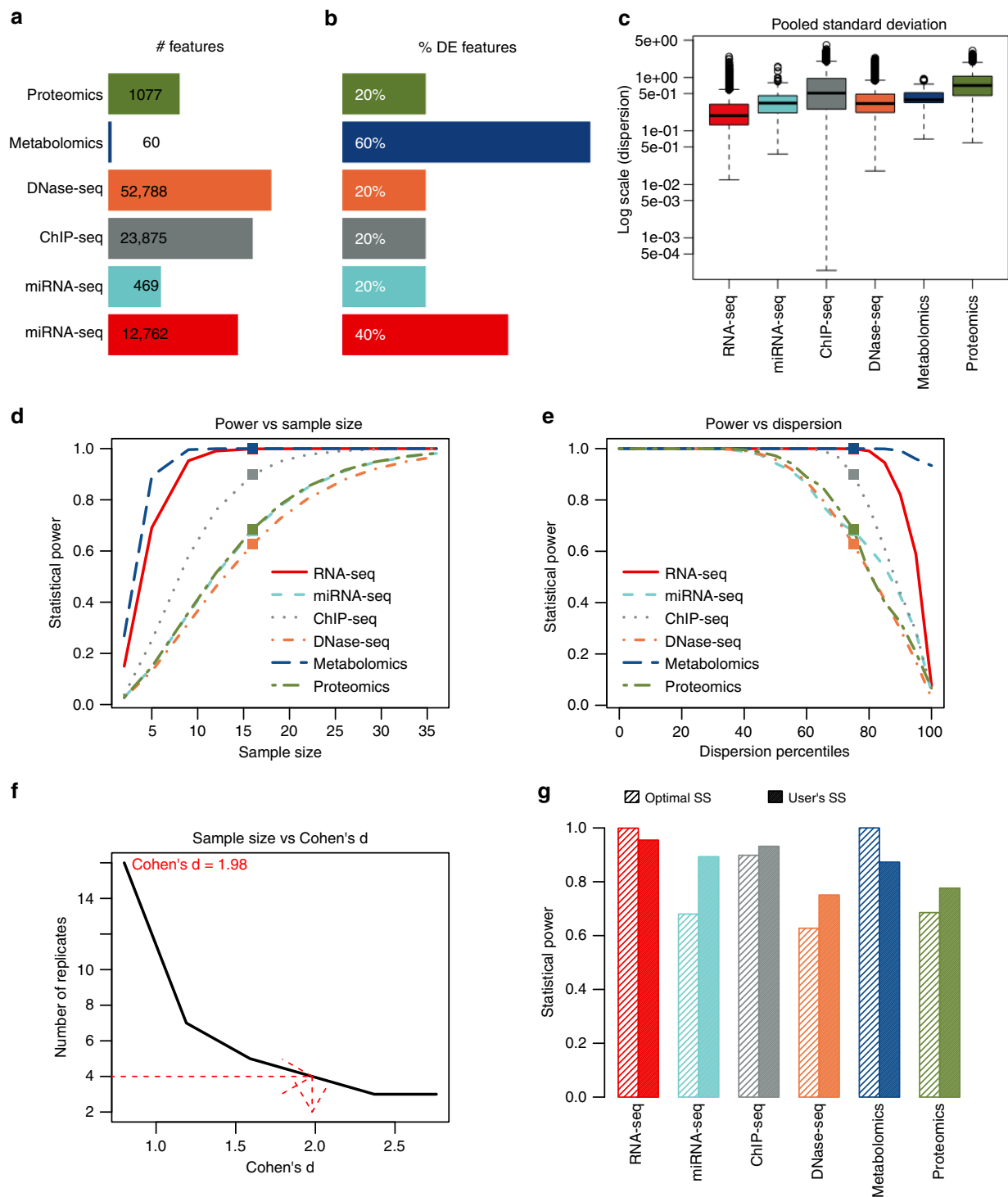
not affect power calculations, but influences downstream integration and interpretation. Metabolite identification in untargeted MS-based methods requires comparisons with databases that collect spectra from known compounds<sup>39–41</sup>. Compound fragments with similar masses and chemical properties often compromise identification, while database incompleteness also contributes to identification failures. A typical identification issue in lipidomics is that many compounds are reported with the same identification label (i.e., sphingomyelins), but slightly different carbon composition and unclear biological significance. Conversely, NMR is highly specific. Each metabolite has a unique pattern in the NMR spectrum, which is also often used for identification of unknown compounds. For untargeted proteomics data, either spectral or sequence databases are used to determine the sequence of the measured peptide. Current identification rates lie at ~40–65% of all acquired spectra with a false discovery rate (FDR) of 1–5%. In targeted approaches, identification is greatly improved by the utilization of isotopically labeled standards. Proteomics suffers from the additional complexity of combining peptides to identify proteins, which is not straightforward. In fact, a recent study highlighted identification as one of the major problems in MS-based proteomics due to differences in search engines and databases<sup>42</sup> and to high false-positive rates<sup>43</sup>. A common identification problem in seq-based methods is the difficulty to allocate reads with sequencing errors or multiple mapping positions, which is addressed either by discarding multi-mapped reads or by estimating correct assignments using advanced statistical tools<sup>44,45</sup>. In ChIP-seq, an identification-related issue is the specificity of the antibody targeting the protein or epigenetic modification. Low antibody specificity may result in reads mapping to nonspecific DNA sequences, leaving true binding regions unidentified. The ENCODE consortium has developed working standards to validate antibodies for different types of ChIP assays<sup>6</sup>.

The coverage of a platform is defined here as the proportion of detected features in the space defined by the type of biomolecule (aka feature space). Targeted MS platforms measure a small subset of compounds with high accuracy; hence the coverage is restricted and lower than for untargeted approaches. As sample complexity is frequently much larger than the sampling capacity of current instruments, even in untargeted methods, identification is limited to compounds with the highest abundances. Repeating sample measurement excluding the features identified in the first run is an efficient strategy to improve coverage, although this requires increased instrument runtime and sample amounts. Coverage in seq-based methods strongly depends on sequencing depth and can potentially reach the complete feature

space associated with each library preparation protocol. In RNA-seq, oligodT-based methods recover polyA RNAs, whereas total RNA requires ribo-depletion, capture of antisense transcripts imposes strand-specific protocols, and microRNAs require specific small RNA protocols. In Methyl-seq, coverage also relates to the applied protocol. Whole-genome bisulfite sequencing has greater genome-wide coverage of CpGs when compared to Reduced Representation by Bisulfite Sequencing (RRBS), while RRBS and MeDIP provide greater coverage at CpG islands. In general, region-based sequencing approaches can cover the whole reference genome, with coverage depending on the efficiency of the protocol employed to enrich the targeted regions.

**From FoM to experimental design.** The FoM analysis across omics revealed that each omic data type possesses different critical performance metrics. In MS, FoM mainly depend on the choice of instrument and approach—targeted vs. untargeted—, while FoM in sequencing methods rely on sequencing depth, library preparation, and eventual bioinformatics post-processing. Although FoM are not a property of data but of the analytical platform, they directly impact data characteristics relevant to experimental design. Overall, FoM strongly relate to the variability in the measurements, which changes in a feature-dependent manner. FoM also determine the final number of measured features and the magnitude of detectable change. Hence, reproducibility influences measurement variability; sensitivity, linear, and dynamic range determine the magnitude and precision of the measurements, and are associated to effect size; the number of features measured by the omic platform is given by the LOD, selectivity, identification, and coverage. These three parameters, variability, effect size, and number of variables, are the key components of power calculations in high-throughput data used by MultiPower to estimate sample size in multi-omic experiments. Figure 3a illustrates the relationship between platform properties, FoM, power parameters, and MultiPower, while Fig. 3b presents an overview of the MultiPower algorithm (see “Methods” section for details).

**MultiPower estimates sample size for a variety of experimental designs.** We applied MultiPower to the STATegra data<sup>46</sup> (Supplementary Note 1) to illustrate power assessment of an existing multi-omic dataset. Figure 4 compares the number of features ( $m$ ), expected percentage of features with a significant signal change ( $p_1$ ), and variability measured as pooled standard deviation (PSD). Note that the number of features measured by each omic platform varies by several orders of magnitude, from 60



**Fig. 4** MultiPower application to STATegra data. **a–c** Pilot data analysis. Each color represents a different omic data type. **a** Number of features per omic. **b** Expected percentage of differentially expressed (DE) features for each omic. **c** Pooled standard deviation (PSD) per gene and per omic for the estimated DE features (pseudo-DE features, see “Methods” section for details). Boxplots represent the median and interquartile range (IQR). Whiskers depict the minimum and maximum of data without outliers, which are the values outside the interval ( $Q1 - 1.5 \times IQR$ ,  $Q3 + 1.5 \times IQR$ ). **d–g** MultiPower results with parameters same sample size in all the omics, minimum power per omic = 0.6, minimum average power = 0.8, FDR = 0.05, initial Cohen’s  $d = 0.8$ . **d** Statistical power curves for each omic for sample sizes between 2 and 35. Squared dots indicate power at the optimal sample size ( $n = 16$ ). **e** Statistical power curves for each omic when considering different percentiles of PSD. Squared dots indicate power for the dispersion used in power calculations for each omic (75th percentile). **f** Curve relating the initial Cohen’s  $d$  to the optimal sample size needed to detect each magnitude of change. For the specified sample size ( $n = 4$ ), the red arrows and text highlight the magnitude of change to be detected (Cohen’s  $d = 1.98$  in this case). **g** Statistical power per omic using the optimal sample size ( $n = 16$ ) with Cohen’s  $d = 0.8$ , and the maximum sample size allowed by the user ( $n = 4$ ) with Cohen’s  $d = 1.98$ .

**Table 1 MultiPower parameters and results from the STATegra pilot data.**

| Omic         | numFeat <sup>b</sup> | DEperc <sup>b</sup> | Delta <sup>a</sup> | Dispersion <sup>a</sup> | minSampleSize | optSampleSize | Power |
|--------------|----------------------|---------------------|--------------------|-------------------------|---------------|---------------|-------|
| RNA-seq      | 12,762               | 0.4                 | 0.61               | 0.32                    | 5             | 16            | 0.999 |
| miRNA-seq    | 469                  | 0.2                 | 0.50               | 0.46                    | 14            | 16            | 0.680 |
| ChIP-seq     | 23,875               | 0.2                 | 1.35               | 0.96                    | 10            | 16            | 0.898 |
| DNase-seq    | 52,788               | 0.2                 | 0.51               | 0.49                    | 16            | 16            | 0.627 |
| Metabolomics | 60                   | 0.6                 | 1.20               | 0.52                    | 4             | 16            | 1.000 |
| Proteomics   | 1077                 | 0.2                 | 1.16               | 1.05                    | 14            | 16            | 0.685 |

MultiPower results were obtained for the same sample size in all technologies, a minimum power per omic of 0.6, a minimum average power of 0.8, and a Cohen's  $d$  of 0.8.

numFeat number of omic features, DEperc expected proportion of DE features, delta difference of means to be detected, dispersion pooled standard deviation, minSampleSize sample size to achieve the minimum power per omic, optSampleSize optimal sample size for the experiment, power power reached with the optimal sample size.

<sup>a</sup>Parameter estimated by MultiPower.

<sup>b</sup>Parameter provided by the user.

metabolites to 52,788 DNase-seq regions (Fig. 4a). This, together with the expected percentage of differentially abundant features (Fig. 4b), affects statistical power when multiple testing correction is applied. Given that PSD is different for each feature (Fig. 4c), users can set the percentile of PSD for power estimations (see “Methods” section). In this example, PSD equals the third quartile, which is a conservative choice. We used MultiPower to calculate the optimal number of replicates for each omic imposing a minimum power of 0.6 per technology, an average power of at least 0.8, an FDR of 0.05, an initial Cohen's  $d$  of 0.8, and same sample size across platforms. MultiPower estimated the optimal number of replicates to be 16 (Fig. 4d, Table 1), which is the number of replicates required by DNase-seq to reach the indicated minimum power. Power estimates were lowest for DNase-seq, followed by proteomics and miRNA-seq, while features with variability below the  $P_{60}$  percentile, DNase-seq, proteomics, and miRNA-seq displayed power values above 0.8 (Fig. 4e). The power plots also indicated that metabolomics and RNA-seq data had the highest power, implying that the detection of differentially expressed (DE) features for these omics is expected to be easier. As costs for generating 16 replicates per omic might be prohibitive, alternatives can be envisioned such as allowing a different number of replicates per omic—at the expense of sacrificing power in some technologies—, accepting a higher FDR, or detecting larger effect sizes. For instance, with four replicates per condition and omic, significant changes were detected at a Cohen's  $d$  of 1.98 (Fig. 4f). Figure 4g depicts a per omic summary of the effective power at the optimal sample size ( $n = 16$ ) with a Cohen's  $d$  of 0.8, and at a sample size ( $n = 4$ ) with a larger Cohen's  $d$ . Results showed that a reduction in the number of replicates can counteract the loss of power if accepting an increase in the magnitude of change to be detected. Finally, MultiPower estimates were further validated by calculating power and Cohen's  $d$  with the published replicate numbers in STATegra ( $n = 3$ ), and verifying the agreement in magnitude and direction of change between RNA-seq and RT-PCR for six B-cell differentiation marker genes<sup>46</sup> (Supplementary Fig. 1).

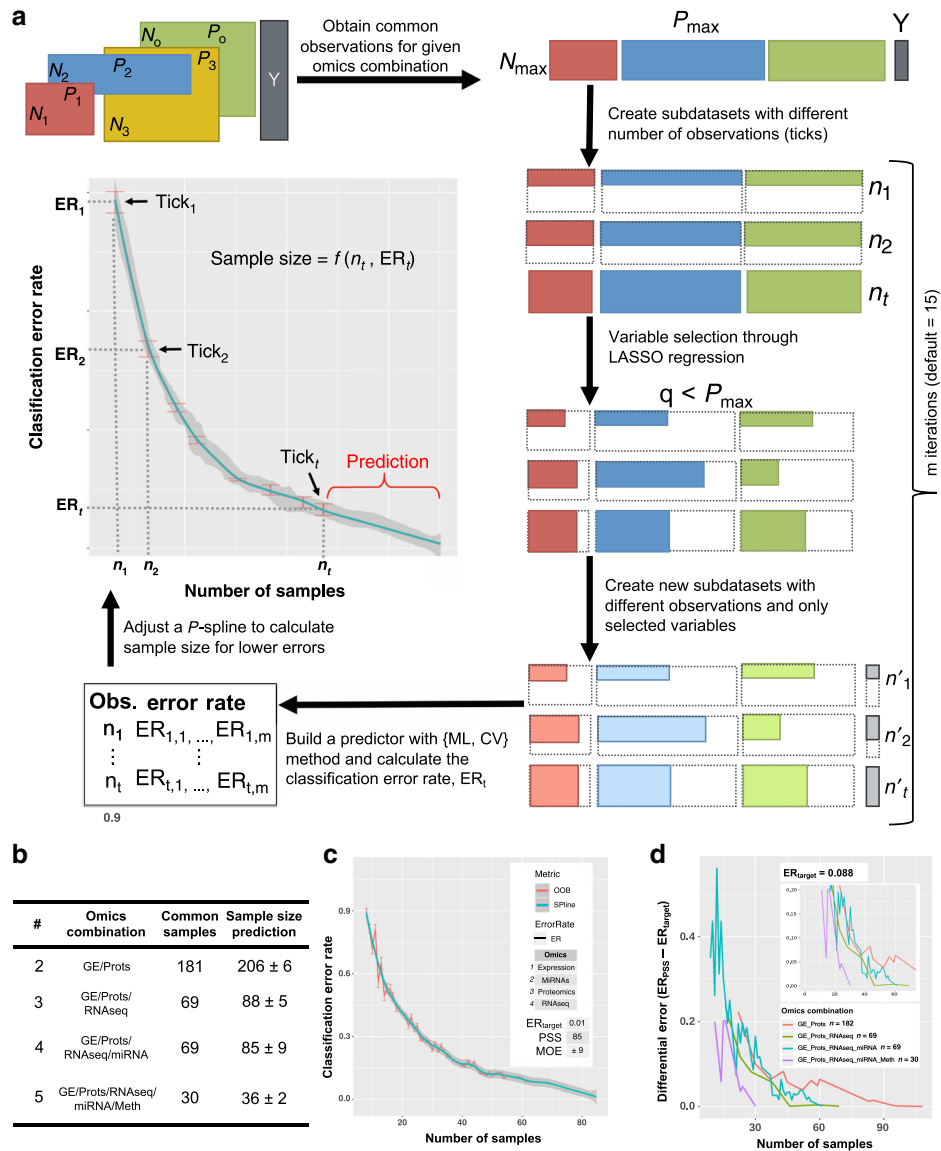
Experimental designs with different sample sizes per omic may limit statistical analysis options, but might be unavoidable or preferred in certain studies. We assessed this possibility with the STATegra dataset assuming the same cost for each technology and keeping the rest of the parameters identical to the previous example. MultiPower analysis revealed that miRNA-seq and DNase-seq required the highest sample size ( $n = 17$ ), while only six and nine replicates per group were required by metabolomics and RNA-seq, respectively (Supplementary Table 2, Supplementary Fig. 2). Power plots revealed that decreasing the sample size for miRNA-seq, DNase-seq, or proteomics results in a strong reduction in power. Again, an alternative to reducing power is to increase the magnitude of change to detect that was initially set to

Cohen's  $d = 0.8$  (Supplementary Fig. 2c). For example, for a sample size not higher than  $n = 5$ , the graph indicates a Cohen's  $d$  of 1.59 for all omics. Additionally, MultiPower can also handle different costs per omic platform and use this information to propose larger sample sizes for inexpensive technologies (Supplementary Tables 3 and 4, Supplementary Fig. 3).

Human multi-omic cohort studies such as The Cancer Genome Atlas database (TCGA) usually collect data from a large number of subjects, where biological variability is naturally higher than in controlled experiments. In such studies, not all subjects may have measurements at all omic platforms and when integrating data decisions should be made to either select individuals profiled by all omic assays—to keep a complete multi-omic design—or to allow a different number of individuals per platform. We illustrate the utility of our method in cohort studies by using MultiPower to estimate power for the integrative analysis of four omic platforms available for the TCGA Glioblastoma dataset<sup>47</sup> (Supplementary Note 1). MultiPower indicated that  $n = 24$  (Supplementary Table 5, Supplementary Fig. 4) is the optimal sample size for complete designs. In this case, Methyl-seq set the required sample size due to the high number of features, low expected percentage of DE features and high variability of this dataset. As the optimal sample size is similar to the number of samples available in the less prevalent omics modality (only 22 samples are available for proneural tumor in methylation data), the joint analysis of current data is not expected to suffer from a major lack of power (Supplementary Fig. 5). However, smaller sample sizes would dramatically impact the number of detected DE features (Supplementary Fig. 5), further validating the results of the MultiPower method. Given that power is affected by the number of omic features (Fig. 3a), an alternative to adjusting power here is the exclusion of methylation features with low between-group variance, as this reduces the magnitude of the multiple testing correction effect on the loss of power. MultiPower helps to assess these options. For example, keeping only methylation sites with an absolute log<sub>2</sub> fold change  $>0.05$  for the power analysis resulted in a reduction of the optimal sample size from 24 to 22 (Supplementary Table 6).

### MultiML predicts sample size for multi-omic based predictors.

Multi-omics datasets may be used in cohort studies to classify biological samples into, for instance, disease subtypes or to predict drug response. In these cases, the analysis goal is not to detect a size effect but to achieve a specified prediction accuracy. MultiML computes the optimal sample size for this type of problem. Briefly, MultiML uses a pilot multi-omics dataset to estimate the relationship between sample size and prediction error, which is then used to infer the sample size required to reach a target classification ER (Fig. 5a, “Methods” section). MultiML is a



**Fig. 5 The MultiML algorithm.** **a** Given a multi-omic dataset with  $O$  different omic types and  $N_O$  samples per omic and a classification vector  $Y$ , MultiML obtains the maximum number of common observations  $N_{max}$  ( $N_{max} \leq N_O$ ) for the combination of omic types specified by the user. MultiML then creates subdatasets (ticks) having different number of observations ( $n_t$ ) from the available  $N_{max}$  and obtains the variables that best explain each tick through LASSO regression. Taking provided machine learning algorithm ML and cross-validation method CV, the selected variables at each tick are used to predict the classification error rate (ER) on a different random subdatasets of size  $n_t$ , resulting in  $ER_t$ . This variable selection-ER prediction process is repeated through  $m$  iterations to obtain an average ER and confidence interval for each tick. On these,  $n_1$  to  $n_t$  and  $ER_1$  to  $ER_t$  values, a first-order-smoothed penalized P-spline regression is adjusted to estimate the learning curve that will allow the prediction of classification ERs for sample sizes larger than  $N_{max}$ . **b-d** MultiML results on the TCGA Glioblastoma data. **b** Predicted sample size for different omics combinations using a target ER of 0.01. Margin of error is given as number of samples. GE expression microarrays, Prots proteomics, miRNA-seq microRNA microarrays, Meth methylation microarrays. **c** Example of MultiML graphical output with the predictive P-spline for a combination of four omics types. OOB out-of-bag classification ER,  $ER_{target}$  target error rate, PSS predicted sample size, MOE margin of error. **d** MultiML evaluation. An increasing number of observations and omic types were run as input data in MultiML having as  $ER_{target}$  the ER achieved with complete data for four omics (69 observations). For the PSS, their ER are recovered from the actual data and their deviation from  $ER_{target}$  is calculated as  $\Delta Error$ . The  $\Delta Error$  is plotted against the size of the input data. Accuracy in ER predictions increases with the size of input data and the number of omics.

flexible framework that (i) predicts sample sizes using different combinations of omics platforms to obtain the best predictive set, as omics technologies may have similar or complementary information content for a given classification scenario; (ii) accepts user-provided machine learning algorithms beyond the implemented partial least squares discriminant analysis (PLS-DA) and random forest (RF) methods; and (iii) offers job parallelization options to speed up calculations when high-performance computing resources are available.

We illustrate MultiML performance using TCGA Glioblastoma data<sup>47</sup>. Omic features were used to predict tumor subtypes with RF, and the target classification error rate ( $ER_{target}$ ) was set to 0.01. Sample sizes were calculated for combinations of two to five omics platforms. We found that the number of samples required to obtain  $ER_{target}$  decreased as the number of omics data platforms increased, suggesting that complementary information was captured by the multi-omics approach resulting in a more efficient predictor (Fig. 5b). Figure 5c shows the classification ER



curve fitted by MultiML for a predictor with four omics. A quadratic pattern is observed, where ERs rapidly decrease as the number of samples increases to reach a stable classification performance. This graph can be used to calculate the number of samples required at different ER levels.

To validate the estimations given by MultiML, we mimicked a prediction scenario where we took fractions of the Glioblastoma data and used MultiML to estimate the sample size for  $ER_{\text{target}}$  equal to the ER of the complete dataset (MinER) or greater. Then, for the predicted sample sizes, we obtained their actual ER from the data and compared them to  $ER_{\text{target}}$  to evaluate the accuracy of the MultiML estimate, and if the magnitude of the input dataset affected this accuracy. As expected, MultiML accuracy increased with the size of the input data and the number of included omics types (Fig. 5d). Accurate predictions (deviations < 5%) were obtained with 50 samples in a three omics combination. Results were similar for other values of  $ER_{\text{target}}$  (Supplementary Fig. 6). We concluded that the sample size could be accurately predicted when input data represents ~40–60% of the required sample size.

## Discussion

As multi-omic studies become more common, guidelines have been proposed for dealing with experimental issues (i.e., sample management) associated with specific omics data types<sup>48</sup>. However, the field has yet to address aspects that are essential to understand the complexity of multi-omic analysis, such as the definition of performance parameters across omic technologies and the formulation of an experimental design strategy in multi-omic data scenarios. In this study, we addressed both issues by proposing FoM as a language to compare omic platforms, and by providing algorithms to estimate sample sizes in multi-omic experiments aiming at differential features analysis (MultiPower) or at sample classification using ML (MultiML).

FoM have traditionally been used in analytical chemistry to describe the performance of instruments and methods. These terms can also be intuitively applied to sequencing platforms, but we noticed that the meaning and relevance of FoM slightly differ for both types of technologies. Here, we explain FoM definitions across omic platforms and discuss which of these metrics are critical to each data type. Detection limit, selectivity, coverage, and identification are FoM with critical influence on the number of features comprising the omics dataset, which in turn affects the power of the technology to identify features with true signal changes. Power diminishes as the number of features increases due to the application of multiple testing corrections to control false positives. Reproducibility and dynamic range may also be very different across platforms, and these have a direct impact on the within-condition variability and across-condition differences of the study. Moreover, while sequencing depth critically affects many of the described FoM of sequencing platforms, in MS-based methods, the choice of a targeted or untargeted method strongly influences FoM values. The number of features detected by the omic platform together with the different measurement variabilities across features and the magnitude of change to be detected, represent major components of power calculations for omics data. The highly heterogeneous nature of these factors across omics platforms calls for specific methods for power calculations in multi-omic experiments, which is addressed by MultiPower.

MultiPower solves the optimization problem of obtaining the sample size that minimizes the cost of the multi-omic experiment, while ensuring both a required power per omic and a global power. As any power computation approach, MultiPower indicates the sample size required to detect a targeted effect size given

a significance threshold. MultiPower graphically represents the relationship between sample size, dispersion, and statistical power of each omic, and facilitates exploration of alternative experimental design choices. Estimates for MultiPower parameters are optimally calculated from pilot data, although they can also be manually provided. The tool accepts normally distributed, count and binary data to facilitate the integration of omic technologies of different analytical nature. Data should have been properly preprocessed and eventual batch effects, removed. In this study, we showcase MultiPower functionalities using sequencing, microarray, and MS data, although the method could be applied to other technologies, such as NMR. Importantly, the optimal sample size can be computed under two different requirements: an equal sample size for all platforms ensuring a common minimal power, or different sample size per omic to achieve the same power. This is relevant for the choice of downstream statistical analysis. Methods that rely on co-variance analysis typically require uniform sample sizes and MultiPower will provide this while revealing the differences in power across data modalities. Methods that combine data based on effect estimates allow sample size differences and can benefit from the equally powered effect estimation. We illustrate the MultiPower method in three scenarios, where different parameterizations are assessed and include both controlled laboratory experiments and cohort data to highlight the general applicability of the method. By discussing interpretations of power plots and the factors that contribute to sample size results, we provide a means to make informed decisions on experimental design and to control the quality of their integrative analysis.

The MultiPower approach is not directly applicable to ML methods used for sample classification, as in this case the basic parameters of the power calculation—significance threshold and effect size—are not applicable. Still, sample size estimation is relevant as multi-omic approaches are frequently used to build classifiers of biological samples. This is not a trivial problem because, aside from FoM, feature relationships within the multivariate space are instrumental in ML algorithms. The MultiML strategy calculates sample size for multi-omic applications, where the classification ER can be used as a measure of performance. MultiML is itself a learning algorithm that learns the relationship between sample size and classification error, and uses this to estimate the number of observations required to achieve the desired classification performance. Hence, pilot data are a requisite for MultiML. Additionally, MultiML has been designed to be flexible for the ML algorithm and to evaluate multiple combinations of omics types in order to identify optimized multi-omic predictors.

Altogether this work establishes, for the first time, a uniform description of performance parameters across omic technologies, and offers computational tools to calculate power and sample size for the diversity of multi-omics applications. We anticipate MultiPower and MultiML will be useful resources to boost powered multi-omic studies by the genomics community.

## Methods

**Scope of the FoM analysis.** In this study, we discuss seven FoM, which we broadly classified into two groups: quantitative or analytical FoM include sensitivity, reproducibility, detection and quantification limit, and linear or dynamic range, and qualitative FoM, which are selectivity, identification, and coverage. To describe how they apply to omic technologies, we distinguish between MS and seq-based platforms.

MS platforms refer to metabolomics and proteomics, which often operate in combination with LC–MS or GC–MS. MS platforms can be used for “untargeted” (measuring a large number of features including novel compounds) and “targeted” assays. While MS is the platform used in most proteomics and metabolomics based studies, NMR (refs. 29,49) is also a relevant platform in metabolomics and is incidentally discussed.

For sequencing platforms, we also consider two subgroups: “feature-based” and “region-based”. In feature-based assays (i.e., RNA-seq or miRNA-seq), a genome annotation file defines the target features to be quantified, and hence these are known a priori. For region-based assays (i.e., CHIP-seq or ATAC-seq), the definition of the target feature to be measured (usually genomics regions) is part of the data analysis process. Applications of RNA-seq to annotate genomes could be considered a region-based assay. We consider here omic assays with a dynamic component regarding the genotype, such as RNA-seq, CHIP-seq, ATAC-seq, Methyl-seq, proteomics, and metabolomics, and also genome variation analyses. However, single-cell technologies are not included, as they require a separate discussion.

Analytical FoM quantify the quality of an analytical measurement platform and are defined at the feature level (gene, metabolite, region, etc.). We describe FoM as properties of a calibration line that displays the relationship between the measured value and the true quantity of the feature in the sample (Fig. 1). In our case, the definition of the true signal differs between omic platforms. As MS platforms and feature-based seq-based platforms measure the concentration or level of the feature of the gene, protein, or metabolite of interest, the true signal represents the average level across all cells contained in the sample (Fig. 1a). In contrast, region-based sequencing techniques aim to identify the genomic region, where a given molecular event occurs, such as the binding of a transcription factor. In this case, the true signal represents the fraction of cells in the sample (or probability), where the event actually occurs within the given coordinates (Fig. 1b).

Lastly, while analytical FoM are defined at the feature level, omic platforms by nature measure many features simultaneously and consequently, the FoM may not be uniform for all of them. For example, accuracy might be different for low vs. highly expressed genes or polar vs. apolar metabolites. In this study, we consider FoM globally and discuss how technological or experimental factors affect the FoM of different ranges of features within the same platform.

**Overview of MultiPower method.** The MultiPower R method (Fig. 3b) performs a joint power study that minimizes the cost of a multi-omics experiment, while requiring both a minimum power for each omic and an average power for all omics. MultiPower calculations are defined for a two groups contrast and implemented in the R package to support the application of the method to single and multiple pairwise comparisons. The parameters required to compute power can be estimated from multi-omic available data (pilot data or data from previous studies) or, alternatively, users can set them. The method considers multiple testing corrections by adjusting the significance level to achieve the indicated FDR. Additionally, MultiPower accepts normally distributed data, count data or binary data, and optimal sample size (number of replicates or observations per condition) can be computed either requiring the same or allowing different sample sizes for each omic. In the latter, the monetary cost is considered as an additional parameter in the power maximization problem. MultiPower can be used to both design a new multi-omic experiment and to assess if an already generated multi-omic dataset provides enough power for statistical analysis.

MultiPower minimizes the total cost of the multi-omics experiment while ensuring a minimum power per omic ( $P_i$ ) and a minimum average power for the whole experiment ( $A$ ). Equation (1) describes the optimization problem to be solved to estimate the optimal number of biological replicates for each omic ( $x_i$ ):

$$\begin{aligned} & \min \sum_{i=1}^I 2c_i x_i \\ & \text{subject to:} \\ & f(x_i, \alpha, \dots) \geq P_i \quad \forall i = 1, \dots, I \\ & \frac{\sum_{i=1}^I f(x_i, \alpha, \dots)}{I} \geq A \\ & x_i \in \mathbb{Z}^+ \end{aligned} \quad (1)$$

Where  $I$  is the number of omics,  $c_i$  is the cost of generating a replicate for omic  $i$ ,  $\alpha$  is the significance level, and  $f()$  represents the power function.

We calculate statistical power under the assumption that the means of two populations are to be compared in the case of count or normally distributed data. Consequently, power is in these cases a function of the sample size ( $x_i$ ) per condition and for a particular omic  $i$ , the significance level ( $\alpha$ ), and other parameters that depend on the nature of the omic data type. For normally distributed data, a  $t$ -test is applied and the power for omic  $i$  is expressed as  $f(x_i, \alpha, \Delta_i, \sigma_i)$ , where  $\Delta_i$  is the true difference of means in absolute value to be detected, and  $\sigma_i$  is the PSD considering two experimental groups. Count data obtained from sequencing platforms can be modeled as a negative binomial distribution (NB) and an exact test can be applied to perform differential analysis. In this case, the power of omics  $i$  is estimated as described in ref. <sup>50</sup>, where power is expressed as  $f(x_i, \alpha, \phi_i, \mu_i, \omega_i)$ , being  $\phi_i$  the dispersion parameter of the NB that relates variance and mean (see Eq. (2)). The effect size is calculated with the fold change ( $\omega_i$ ) between both groups as well as the average counts ( $\mu_i$ ).

$$\sigma^2 = \mu + \mu^2 \phi \quad (2)$$

For binary data with 0/1 or TRUE/FALSE values indicating, for instance, if a mutation is present or not, or if a transcription factor is bound or not, the goal is comparing the percentages of 1 or TRUE values between two populations. In this

case, the parameters needed to estimate power are related to the difference between proportions to be detected (the effect size) and the sample size, but variability is not considered.

As MultiPower considers multiple testing correction to control FDR, the significance level given to the power function is adapted to this correction ( $\alpha^*$ ). We followed the strategy proposed in other studies<sup>51–53</sup> given by Eq. (3).

$$\alpha^* = \frac{r_1 \alpha}{(m - m_1)(1 - \alpha)}, \quad (3)$$

where  $m$  is the number of features in a particular omic,  $m_1$  is the number of expected DE features,  $r_1$  is the expected number of true detections, and  $\alpha$  is the desired FDR.

**Sample size scenarios in MultiPower.** In multi-omic studies, an assorted range of experimental designs can be found. All omic assays may be obtained on the same biological samples or individuals, which would result in identical replicates number for all data types. However, this is not always possible due to restrictions in cost or biological material, exclusion of low-quality samples, or distributed omic data generation. In these cases, sample size differs among omic types and yet the data are to be analyzed in an integrative fashion. MultiPower contemplates these two scenarios.

Under the first scenario, adding the constraint of equal sample size for all omics ( $x_i = x$  for all  $i = 1, \dots, I$ ) to the optimization problem in Eq. (1) results in a straightforward solution. First, the minimum sample size required to meet the constraint on the minimum power per omic ( $x_i$ ) is calculated and the initial estimation of  $x$  is set to  $x = \max_i \{x_i\}$ . Next, the second constraint on the average power is evaluated for  $x$ . If true, the optimal sample size is  $x_{\text{opt}} = x$ . Otherwise,  $x$  is increased until the constraint is met. Note that, under this sample size scenario, the cost per replicate does not influence the optimal solution  $x_{\text{opt}}$ .

Under the second scenario, allowing different sample sizes for each omic, the optimization problem in Eq. (1) becomes a nonlinear integer programming problem, as the statistical power is a nonlinear function of  $x_i$ . The optimization problem can be transformed into a 0–1 linear integer programming problem by defining the auxiliary variables  $z_n^i$  for each omic  $i$  and each possible sample size  $n$  from 2 to a fixed maximum value  $n_{\text{max}}$ , where  $z_n^i = 1$  when the sample size for omic  $i$  is  $n$ . The new linear integer programming problem can be formulated as follows:

$$\begin{aligned} & \min \sum_{i=1}^I 2c_i \left( \sum_{n=2}^{n_{\text{max}}} n z_n^i \right) \\ & \text{subject to:} \\ & \sum_{n=2}^{n_{\text{max}}} f_i(n) z_n^i \geq P_i \quad \forall i = 1, \dots, I \\ & \sum_{i=1}^I \sum_{n=2}^{n_{\text{max}}} f_i(n) z_n^i \geq A \\ & \sum_{n=2}^{n_{\text{max}}} z_n^i = 1 \quad i = 1, \dots, I \\ & z_n^i \in \{0, 1\} \quad \forall i \forall n \end{aligned} \quad (4)$$

where  $f_i(n)$  is the power for sample size  $n$ , given the parameters of omic  $i$ .

**MultiPower implementation.** The MultiPower R package implements the described method together with several functionalities to support both the selection of input parameters required by the method and the subsequent interpretation of the results. The R package and user’s manual are freely available at <https://github.com/ConesaLab/MultiPower>.

The MultiPower package requires different parameters to compute the optimal sample size. As the choice of parameters can be challenging, MultiPower can estimate them from pilot or similar existing data. The multi-omic pilot data must have two groups and at least two replicates per group. The algorithm assumes data are already preprocessed, normalized, and free of technical biases. Both normally distributed and count data are accepted. We recommend raw count data to be provided for sequencing technologies as MultiPower deals with the sequencing depth bias. However, when count data contains other sources of technical noise, we recommend previous transformations to meet normality (e.g., with log or voom transformation) and indicating MultiPower that data are normally distributed. We also recommend removing low count features from sequencing data. For data containing missing values, these must be removed or imputed before running MultiPower. MultiPower also accepts binary data (e.g., SNP data, CHIP-seq transcription binding data, etc.). In this case, values must be either 0–1 or TRUE/FALSE.

Each data type requires a different sample size computation. For normally distributed data, MultiPower uses the `power.t.test()` function from `stats` R package (based on a classical  $t$ -test), while the `sample_size()` and `est_power()` functions from R Bioconductor `RnaSeqSampleSize` package<sup>54</sup> are used for count data, where a NB test is applied. For binary data, the `power.prop.test()` function from `stats` R package is used. Moreover, MultiPower transforms parameters provided by the user to fit

specific arguments required by these functions, in such a way that magnitudes are maintained roughly comparable for all data types.

The power function for normally distributed data depends on the effect size, i.e., the true difference of means to be detected (delta parameter,  $\Delta$ ). The delta parameter depends on the dynamic range of the omic and may be nonintuitive following data preprocessing. On the contrary, the fold change and mean counts are used in count data to estimate power. Therefore, MultiPower uses instead the Cohen's  $d$  ( $d_0$ ) value to homogenize input parameters. Cohen's  $d$  is defined as  $\Delta/\sigma$  and does not depend on the scale of the data as occurs for the  $\Delta$  value. Therefore, the same value can be chosen for all omic platforms. Cohen<sup>55</sup> and Sawilowsky<sup>56</sup> proposed the classification presented in Supplementary Table 7 to establish a Cohen's  $d$  value. MultiPower computes the Cohen's  $d$  value for all omic features and applies this value to estimate the set  $M_1$  of DE features (those with  $d > d_0$ ), which are called pseudo-DE features. We recommend setting the same Cohen's  $d$  initial value ( $d_0$ ) for all the omics although MultiPower allows different values for each one of them.

The equivalent to Cohen's  $d$  when comparing proportions in groups A and B is Cohen's  $h$ , which can be defined as  $h = |\varphi_A - \varphi_B|$ , where  $\varphi_i = 2 \arcsin \sqrt{p_i}$  and  $p_i$  is the proportion of 1 or TRUE values in group  $i$ . Classification in Supplementary Table 7 is also valid for Cohen's  $h$ .

The multiple testing correction is applied to each omic analysis (see Eq. (3)), taking the significance level set by the user as FDR value. The user must also provide the expected percentage of DE features ( $p_1$ ) for each omic. Given the number of features in a particular omic ( $m$ ), the expected number of DE features can be simply represented as  $m_1 = mp_1$ . The RnaSeqSampleSize package<sup>50</sup> estimates the expected number of true detections  $r_1$  for count data. For normal or binary data,  $r_1 = m_1$  is assumed (as in `ssize.fdr` R/CRAN package<sup>57</sup>).

The intrinsic characteristics of each omic feature may result in different power values for each one of them. To have a unique power estimation per omic, an average power parameter from the distribution of such parameters across pseudo-DE features must be provided. For normally distributed data, the PSD parameter ( $\sigma$ ) is computed as the percentile  $P_k$  of the PSDs for all the pseudo-DE features, where  $P_k$  is set by the user (default value  $P_{75}$ ). Thus, the  $\Delta$  estimation should equal the chosen PSD. However, to avoid dependence on a single value, MultiPower evaluates all the pseudo-DE features with PSD between percentiles  $P_{k-5}$  and  $P_{k+5}$  and the  $P_{100-k}$  of the corresponding  $\Delta$  values is taken as conservative choice. For count data, MultiPower estimates the parameter  $w$  that considers the different sequencing depth between samples as  $w = D_B/D_A$ , where  $D_i$  is the geometric mean of the sequencing depth of the samples in group  $i$  divided by median of the sequencing depth of all samples (MSD). To compute the rest of parameters, count values are normalized by dividing them by this ratio, that is, the sequencing depth of the corresponding sample divided by MSD. To be consistent with the previous choice for normally distributed data, the variance per condition ( $\sigma^2$ ) is also estimated as the percentile  $P_k$  of the variances for all pseudo-DE features and conditions. Mean counts ( $\mu$ ) are obtained as the percentile  $P_{100-k}$  of mean counts in the reference group (A) corresponding to pseudo-DE features with variance between percentiles  $P_{k-5}$  and  $P_{k+5}$ . The dispersion parameter ( $\phi$ ) is then derived from Eq. (2). Finally, the fold change of pseudo-DE features ( $\omega$ ) is estimated as the percentile  $P_{100-k}$  of the fold changes corresponding to pseudo-DE features with a PSD between percentiles  $P_{k-5}$  and  $P_{k+5}$ . For binary data, the proportions chosen to estimate power correspond to the  $P_{100-k}$  of pseudo-DE features for Cohen's  $d$ , and are stored in the delta output parameter of MultiPower. As variability is not considered for this data type, dispersion power plots are not generated in this case.

Once the power parameters are obtained for each omic and the user sets the minimum power per omic and the average power for the experiment, the optimization problem in Eq. (4) is completely defined and MultiPower makes use of `lpmodeler` and `Rsymphony` R packages to solve it. Note that the application of these packages to solve the problem is only needed when the number of replicates for each omic differs. MultiPower returns a summary table with the provided and/or estimated power parameters, the obtained optimal sample size, and corresponding power per omic.

**Power parameters in the absence of prior data.** While parameter estimation from previous data is recommended for MultiPower analysis, this might not always be feasible. In this case, MultiPower requires parameters to be provided by users, which could be challenging. Here, we provide recommendations for critical MultiPower parameters.

The average value for the standard deviation per omic feature and condition partially depends on the reproducibility of omics technology. Overestimating this parameter guarantees that the sample size fits the power needed but may lead to too large sample sizes. According to our experience, a good value for the standard deviation is 1.

Value for the expected proportion of DE features per omic should be set according to results seen in similar studies. A high percentage is expected for cell differentiation processes or diseases like cancer, while small perturbations or other types of diseases may induce fewer changes.

Typical values for the minimum fold change between conditions and the mean of counts for the DE features when using count data are 2 and 30, respectively, but again these values depend on the sequencing depth of the experiment and the magnitude of the expected molecular changes.

**Power study.** After obtaining the optimal sample size with MultiPower, some questions may still arise, especially when this sample size exceeds the available budget for the experiment:

- How much reduction of the sample size can we afford without losing too much power?
- If power cannot be decreased but the sample size has to be reduced, how will this reduction influence the effect size to be detected?
- Can we remove any omic platform with negligible changes between conditions, since this platform imposes a too large optimal sample size?

To provide answers to these and similar questions, MultiPower returns several diagnostic plots (see Fig. 4d–g for instance). The power vs. sample size plot shows variations in power as a function of the sample size. The power vs. dispersion plot also displays the power curves but for different dispersion values (PSD in normal data or  $\phi$  parameter in count data). In both plots, the power for the estimated optimal sample size or the fixed dispersion value is represented by a square dot.

If the optimal sample size estimated by MultiPower exceeds the available budget, researchers may opt for increasing the effect size (given by the Cohen's  $d$ ) to be detected and allowing a smaller sample size without modifying the required power. To perform this analysis, the `postMultiPower()` function can be applied, which computes the optimal sample size for different values of Cohen's  $d$ , from the initial value set by the user to  $d_{\max} = \min_{i \in I} \{P_{90}^i(d)\}$ , where  $P_{90}^i(d)$  is the 90th percentile of the Cohen's  $d$  values for all the features in omic  $i$ . This choice ensures sufficient pseudo-DE features to estimate the rest of the parameters needed to compute power.

**MultiPower for multiple comparisons.** Although MultiPower algorithm is essentially defined for a two groups comparison, the MultiPower R package supports experimental designs with multiple groups. Assuming that a pilot dataset is available, the `MultiGroupPower()` function automatically performs all the possible pairwise comparisons (or those comparisons indicated by the user) and returns both a summary of the power and optimal sample size for each comparison, and a numerical and graphical global summary for all the comparisons. In this global summary, the global optimal sample size is computed as the maximum of optimal sizes obtained for the individual comparisons. Therefore, the solution given by MultiPower method does not allow a different sample size for each group. Users must be aware that different optimal sample size could be obtained in this case.

**MultiML method.** The MultiML method deals with a multi-omic sample size estimation problem, in which we have prior data consisting of a list of  $O$  omic data matrices of dimension  $N$  observations  $\times P$  predictor variables, and a categorical response variable  $Y$  providing the class each observation belongs to (Fig. 5). The number of observations and variables can differ across omics. MultiML estimates the optimal sample size required to minimize the classification ER. MultiML allows users to perform analyses for different combinations of the  $O$  available omics. In each combination, the algorithm selects the common observations across omics,  $N_{\max}$ . Next, an increasing number of observations (from two observations per class to the total number of observations  $N_{\max}$ ) are selected to build a class predictor using the ML algorithm, sampling strategy (SS), and cross-validation (CV) method indicated by the user. The algorithm incorporates a LASSO variable selection step to reduce the number of predictors and computation time. LASSO regression is incorporated at the performance evaluation step to avoid overfitting. This process is repeated 15 times as default, and for each iteration the classification ER of the predictor is computed. These data are used to build a polynomial model that describes the relationship between ER and the number of observations in the multi-omics dataset, which can be then used to estimate the sample size required for a given  $ER_{\text{target}}$ . Supplementary Fig. 7 shows the pseudocode of the MultiML algorithm.

**MultiML implementation.** MultiML is implemented as a set of R functions that calculate the classification ERs at increasing number of observations, fit the sample size predictive model and graphically display results. Auxiliary functions are included to prepare multi-omic data and optimize computational requirements. The main function is `ER_calculator()`, which basically takes multi-omic pilot data and returns the estimated sample size. This function requires an  $ER_{\text{target}}$  which is the maximum classification ER that the user is willing to accept in the study. When not provided by the user, MultiML takes the ER value obtained from the pilot data. Users may select from two CV methods, tenfold and leave-one-out CV, and prediction performance results are averaged across iterations. MultiML can operate with any user-supplied ML algorithm, provided that this is a wrapped R function with input and output formats supported by MultiML. By default, MultiML includes RF and PLS-DA as ML options. For RF the R package `randomForest`<sup>58</sup> is required. In this method, the classification ER is calculated after constraining all selected omics variables into a wide matrix. For PLS-DA the `mixOmics` R package<sup>59</sup> is required. In this case, each omics data matrix is reduced to its significant variables and analysis is performed maintaining a three-way  $N \times P \times O$  structure, where  $N$  are individuals,  $P$  are the significant variables, and  $O$  are the different omics in the study. If PLS-DA is selected as ML method, the SSs may or not be balanced, returning either an overall classification ER or a balanced ER calculated on the left-

out samples. For RF method, only ER is available as both methods give similar results. MultiML also provides different prediction distances for PLS-DA and RF used to assign a category to samples. For PLS-DA, maximum distance, distance to the centroid, and Mahalanobis distance were implemented. These prediction distances can be defined as a model with  $H$  components. Given  $N_{\text{newInds}}$  new individuals and their corresponding omic data matrix  $\mathbf{X}_{\text{newInds}}$ , the predicted response variable  $\hat{\mathbf{Y}}_{\text{newInds}}$  can be computed as follows:

$$\hat{\mathbf{Y}}_{\text{newInds}} = \mathbf{X}_{\text{newInds}} \times \mathbf{W}(\mathbf{D}^T \mathbf{W})^{-1} \mathbf{B} \quad (5)$$

where  $\mathbf{W}$  is a  $p$  (variables)  $\times H$  matrix containing the loading vectors associated with  $\mathbf{X}$ ;  $\mathbf{D}$  is a  $p \times H$  matrix containing the regression coefficients of  $\mathbf{X}$  on its  $H$  latent components; and  $\mathbf{B}$  is an  $H \times n$  (individuals) matrix containing the regression coefficients of  $\mathbf{Y}$  on the  $H$  latent components associated to  $\mathbf{X}$ . The predicted scores ( $\mathbf{T}_{\text{pred}}$ ) are computed as:

$$\mathbf{T}_{\text{pred}} = \mathbf{X}_{\text{newInds}} \times \mathbf{W}(\mathbf{D}^T \mathbf{W})^{-1} \quad (6)$$

In turn, for RF, out-of-bag (OOB) estimate was included<sup>60</sup>. The RF classifier is trained using bootstrap aggregation, where each new tree is fit from a bootstrap sample of the training observations  $z_i = (x_i, y_i)$ . The OOB error is the average error for each  $z_i$  calculated using predictions from the trees that do not contain  $z_i$  in their respective bootstrap sample. The prediction distances are then applied to assign a category to each new sample. To reduce the number of omics variables and computational time, a generalized linear model via penalized maximum likelihood is applied. The regularization path is computed for the LASSO penalty using the *glmnet* R package<sup>61</sup>. This variable selection step is performed on a random selection of  $n$  observations and repeated as many times as required (15 by default) to retain the  $q$  variables that best explain the classification vector  $\mathbf{Y}$ . The algorithm then takes a new set of  $n'$  observations and, uses only the  $q$  variables to calculate the classification ER using the ML and CV method chosen by the user (Fig. 5). The sample size prediction curve is estimated with the vector of ERs  $\overline{\text{ER}}_{\text{it}}$  and the vector of number of samples  $\tau_i$ . The accuracy of MultiML depends on the number of ticks obtained to fit the sample size prediction curve. The algorithm starts with a low (5) number of ticks and iteratively increases them until the addition of new ticks does not improve the accuracy of the sample size prediction curve. The algorithm termination protocol implemented in MultiML, forces a stop when at least 12% of the range of values of three consecutive models (ticks addition steps) are equal. Finally, a first-order-smoothed penalized P-spline calculation is performed to model the relationship between the number of samples and the classification ER (refs. <sup>62,63</sup>). The degrees of freedom of this model are the number of observation subsets evaluated (ticks) minus 1. The model is used to predict the sample size required to obtain a given classification error.

MultiML is a computationally intensive algorithm. The function `RequiredTimeTest()` allows users to estimate the time required to run the full predictive model at the local installation. For users who can benefit from parallelization options, we have implemented the `slurm_creator()` R function that creates a .sh script to run MultiML calculations in a SLURM cluster. Moreover, MultiML can be run incrementally. The function `Previous_CER()` allows the utilization of a previous MultiML result with  $N$  samples in a new MultiML calculation that expands the size of the prior dataset by  $M$  samples, thereby significantly accelerating the calculation of the new model.

**MultiML output.** MultiML returns all numerical data of the ML models created to fit the sample size prediction model. This includes the evaluated subsets of observations and data types, the classification ER values obtained at each iteration, and the predicted sample size together with the margin of error of the prediction. Additionally, the function `ErrorRatePlot()` prints graphically the relationship between different sample sizes and their corresponding classification ERs. An example is shown in Fig. 5c. The plot also includes the fitted penalized smooth spline model to graphically obtain the sample size for an  $\text{ER}_{\text{target}}$  not achievable with the pilot dataset. The user can also create a comparative plot of all omic combinations to determine the best contributing data modality to an accurate classification by using `Comparative_ERPlot()` function.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The STATegra data used in this manuscript are available from: Gene Expression Omnibus with accession numbers GSE75417, GSE38200, GSE75394, GSE75393, and GSE75390; <https://identifiers.org/pride.project:PX0003263>; and <https://identifiers.org/metablights:MTBLS283>. The TCGA glioblastoma data used in this manuscript are available from <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/studied-cancers/glioblastoma>.

## Code availability

The MultiPower and MultiML methods are available at GitHub repository <https://github.com/ConesaLab/MultiPower>.

Received: 27 July 2018; Accepted: 29 May 2020;

Published online: 18 June 2020

## References

- Thingholm, L. B. et al. Strategies for integrated analysis of genetic, epigenetic, and gene expression variation in cancer: addressing the challenges. *Front. Genet.* **7**, 2 (2016).
- Blatti, C., Kazemian, M., Wolfe, S., Brodsky, M. & Sinha, S. Integrating motif, DNA accessibility and gene expression data to build regulatory maps in an organism. *Nucleic Acids Res.* **43**, 3998–4012 (2015).
- Fagan, A., Culhane, A. C. & Higgins, D. G. A multivariate analysis approach to the integration of proteomic and gene expression data. *Proteomics* **7**, 2162–2171 (2007).
- Conesa, A., Prats-Montalbán, J. M., Tarazona, S., Nueda, M. J. & Ferrer, A. A multiway approach to data integration in systems biology based on Tucker3 and N-PLS. *Chemometrics Intell. Lab. Syst.* **104**, 101–111 (2010).
- Conesa, A. et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
- Landt, S. G. et al. CHIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
- Wei, Z., Zhang, W., Fang, H., Li, Y. & Wang, X. esATAC: an easy-to-use systematic pipeline for ATAC-seq data analysis. *Bioinformatics* **34**, 2664–2665 (2018).
- Sun, Z. et al. SAAP-RRBS: streamlined analysis and annotation pipeline for reduced representation bisulfite sequencing. *Bioinformatics* **28**, 2180–2181 (2012).
- Xia, J. & Wishart, D. S. Using MetaboAnalyst 3.0 for comprehensive metabolomics data analysis. *Curr. Protoc. Bioinformatics* **55**, 14.10.1:14.10.91 (2016).
- Davidson, R. L., Weber, R. J. M., Liu, H., Sharma-Oates, A. & Viant, M. R. Galaxy-M: a Galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data. *Gigascience* **5**, 10 (2016).
- Goeminne, L. J. E., Gevaert, K. & Clement, L. Experimental design and data analysis in label-free quantitative LC/MS proteomics: a tutorial with MSqRob. *J. Proteom.* **171**, 23–36 (2018).
- Codrea, M. C. & Nahnsen, S. Platforms and pipelines for proteomics data analysis and management. *Adv. Exp. Med Biol.* **919**, 203–215 (2016).
- Park, Y., Figueroa, M., Rozek, L. & Sartor, M. MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics* **30**, 2414–2422 (2014).
- Andrews S. FASTQC. A Quality Control Tool for High Throughput Sequence Data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2014).
- García-Alcalde, F. et al. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* **28**, 2678–2679 (2012).
- Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).
- Lassmann, T., Hayashizaki, Y. & Daub, C. O. SAMStat: monitoring biases in next generation sequencing data. *Bioinformatics* **27**, 130–131 (2011).
- Ewels, P., Magnusson, M., Lundin, S. & Källér, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
- Poplawski, A. & Binder, H. Feasibility of sample size calculation for RNA-seq studies. *Brief. Bioinform.* **19**, 713–720 (2018).
- Li, C.-I., Samuels, D. C., Zhao, Y.-Y., Shyr, Y. & Guo, Y. Power and sample size calculations for high-throughput sequencing-based experiments. *Brief. Bioinform.* **19**, 1247–1255 (2018).
- Banko, M. & Brill, E. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* 26–33 (Association for Computational Linguistics, France, 2001).
- Figueroa, R. L., Zeng-Treitler, Q., Kandula, S. & Ngo, L. H. Predicting sample size required for classification performance. *BMC Med. Inf. Decis. Mak.* **12**, 8 (2012).
- Dunn, W. B. & Ellis, D. I. Metabolomics: current analytical platforms and methodologies. *TrAC Trends Anal. Chem.* **24**, 285–294 (2005).
- Chang, C.-Y. et al. Protein significance analysis in selected reaction monitoring (SRM) measurements. *Mol. Cell. Proteomics* **11**, M111.014662 <https://doi.org/10.1074/mcp.M111.014662> (2012).
- Markley, J. L. et al. The future of NMR-based metabolomics. *Curr. Opin. Biotechnol.* **43**, 34–40 (2017).
- Rocke, D. M. & Lorenzato, S. A two-component model for measurement error in analytical chemistry. *Technometrics* **37**, 176–184 (1995).

27. Van Batenburg, M. F., Coulier, L., van Eeuwijk, F., Smilde, A. K. & Westerhuis, J. A. New figures of merit for comprehensive functional genomics data: the metabolomics case. *Anal. Chem.* **83**, 3267–3274 (2011).
28. Dunn, W. B. et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat. Protoc.* **6**, 1060–1083 (2011).
29. Keun, H. C. *NMR-based Metabolomics P001–P368* (The Royal Society of Chemistry, 2018).
30. Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: a matter of depth. *Genome Res.* **21**, 2213–2223 (2011).
31. Kim, S. et al. Evaluation and optimization of metabolome sample preparation methods for *Saccharomyces cerevisiae*. *Anal. Chem.* **85**, 2169–2176 (2013).
32. Köcher, T., Swart, R. & Mechtler, K. Ultra-high-pressure RPLC hyphenated to an LTQ-Orbitrap Velos reveals a linear relation between peak capacity and number of identified peptides. *Anal. Chem.* **83**, 2699–2704 (2011).
33. Boja, E. S. & Rodriguez, H. Mass spectrometry-based targeted quantitative proteomics: achieving sensitive and reproducible detection of proteins. *Proteomics* **12**, 1093–1110 (2012).
34. Olkhov-Mitsel, E. & Bapat, B. Strategies for discovery and validation of methylated and hydroxymethylated DNA biomarkers. *Cancer Med.* **1**, 237–260 (2012).
35. Armbruster, D. A. & Pry, T. Limit of blank, limit of detection and limit of quantitation. *Clin. Biochem. Rev.* **29**, S49–S52 (2008).
36. Arsova, B., Zauber, H. & Schulze, W. X. Precision, proteome coverage, and dynamic range of Arabidopsis proteome profiling using (15)N metabolic labeling and label-free approaches. *Mol. Cell. Proteomics* **11**, 619–628 (2012).
37. Kuhn, E. et al. Interlaboratory evaluation of automated, multiplexed peptide immunoaffinity enrichment coupled to multiple reaction monitoring mass spectrometry for quantifying proteins in plasma. *Mol. Cell. Proteomics* **11**, M111.013854 <https://doi.org/10.1074/mcp.M111.013854> (2012).
38. Kondrat, R. W., McClusky, G. A. & Cooks, R. G. Multiple reaction monitoring in mass spectrometry/mass spectrometry for direct analysis of complex mixtures. *Anal. Chem.* **50**, 2017–2021 (1978).
39. Wishart, D. S. et al. HMDB: the human metabolome database. *Nucleic Acids Res.* **35**, D521–D526 (2007).
40. Kopka, J. et al. GMD@CSB.DB: the golm metabolome database. *Bioinformatics* **21**, 1635–1638 (2005).
41. Scholz, M. & Fiehn, O. SetupX—a public study design database for metabolomic projects. *Pac. Symp. Biocomput.* **12**, 169–180 (2007).
42. Bell, A. W. et al. A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat. Methods* **6**, 423–430 (2009).
43. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).
44. Roberts, A., Feng, H. & Pachter, L. Fragment assignment in the cloud with eXpress-D. *BMC Bioinformatics* **14**, 358 (2013).
45. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
46. Gomez-Cabrero, D. et al. STATegra, a comprehensive multi-omics dataset of B-cell differentiation in mouse. *Sci. Data* **6**, 256 (2019).
47. Verhaak, R. G. W. et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98–110 (2010).
48. Altmäe, S. et al. Guidelines for the design, analysis and interpretation of 'omics' data: focus on human endometrium. *Hum. Reprod. Update* **20**, 12–28 (2014).
49. Reo, N. V. NMR-based Metabolomics. *Drug Chem. Toxicol.* **25**, 375–382 (2002).
50. Li, C.-I., Su, P.-F. & Shyr, Y. Sample size calculation based on exact test for assessing differential expression analysis in RNA-seq data. *BMC Bioinformatics* **14**, 357–357 (2013).
51. Storey, J. D. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B* **64**, 479–498 (2002).
52. Jung, S.-H. Sample size for FDR-control in microarray data analysis. *Bioinformatics* **21**, 3097–3104 (2005).
53. Storey, J. D. & Tibshirani, R. Estimating the positive false discovery rate under dependence, with applications to DNA microarrays. *Stanford Stat. Rep.* **28** (2001).
54. Zhao, S., Li, C.-I., Guo, Y., Sheng, Q. & Shyr, Y. RnaSeqSampleSize: real data based sample size estimation for RNA sequencing. *BMC Bioinformatics* **19**, 191 (2018).
55. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (L. Erlbaum Associates, 1988).
56. Sawilowsky, S. New effect size rules of thumb. *J. Mod. Appl. Stat. Methods* **8**, 597–599 (2009).
57. Liu, P. & Hwang, J. T. Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics* **23**, 739–746 (2007).
58. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
59. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **13**, e1005752–e1005752 (2017).
60. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning* Vol. 112 (Springer, 2013).
61. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
62. Meyer, M. C. Inference using shape-restricted regression splines. *Ann. Appl. Stat.* **2**, 1013–1033 (2008).
63. Ramsay, J. O. Monotone regression splines in action. *Stat. Sci.* **3**, 425–441 (1988).

### Acknowledgements

This work has been funded by FP7 STATegra project agreement 306000 and Spanish MINECO grant BIO2012–40244. In addition, work in the Imhof lab has been funded by the (DFG; CIPSM and SFB1064). The work of L.B.-N. has been funded by the University of Florida Startup funds.

### Author contributions

S.T.: contributed to FOM analysis, implemented and tested MultiPower method, and drafted the manuscript; L.B.-N.: implemented and tested the MultiML method, and drafted the manuscript; D.G.-C. contributed to FOM analysis, tested MultiPower, and MultiML methods; A.S.: contributed to FOM analysis of proteomics data; A.I.: contributed to FOM analysis of proteomics data; T.H.: discussed FOM analysis of metabolomics data; J.T.: supervised parts of the work; J.A.W.: contributed to FOM definition and analysis; and A.C.: supervised, coordinated the work, and drafted the manuscript. All authors contributed to and approved the final version of the manuscript.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41467-020-16937-8>.

**Correspondence** and requests for materials should be addressed to A.C.

**Peer review information** *Nature Communications* thanks Timothy Ebbers, Daniel Rotroff and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons

Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020