Evidence-based methods in studies of biology and data analysis

by

Leslie Myint

A dissertation submitted to The Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

February 2018

Abstract

When scientists use familiar data analysis methods out of comfort or convenience, disciplines can suffer in their scientific inferences if these methods are not appropriate for their ultimate goals. Older fields experience this when long-standing methods are used simply for their longevity. Newer fields experience this when scientists transfer methods from other areas without evaluating their performance in these new domains. This work represents a collection of methods and results that contribute to evidence-based analytical practice in three different domains: mass spectrometry-based metabolomics, massively parallel reporter assays, and data science training. In the first two domains, we present new methods that improve current practice for comparative (differential) analysis in those fields. Specifically these methods are shown to be statistically calibrated and powerful compared to existing alternatives. In the third domain, we present experimental results regarding the actions and perceptions in data analysis practice. These results have implications for data analysis training and education. Broadly, in these three domains, we provide tools and discuss findings that enable higher quality work in applied research.

Thesis Committee

Primary Readers

Kasper Daniel Hansen (Primary Advisor)
Assistant Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Jeffrey Leek
Associate Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Loyal Goff
Assistant Professor
Department of Neuroscience
Johns Hopkins University
McKusick-Nathans Institute of Genetic Medicine

Thomas Hartung
Professor
Departments of Environmental Health and Engineering & Molecular Microbiology and Immunology at
Johns Hopkins Bloomberg School of Public Health

Leah Jager
Assistant Scientist
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Alternate Readers

Ingo Ruczinski
Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Floyd Bryant
Professor
Department of Biochemistry and Molecular Biology
Johns Hopkins Bloomberg School of Public Health

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

How should I analyze this data? This is the eternal question facing scientists once data have been collected. For even modestly complex situations, this question is not straightforward. There are a myriad of statistical tools, approaches, and software packages that can be used over the course of an analysis. Different tools are accompanied by different assumptions, theoretical properties, and real-data performance. In well-established fields that have close ties to computational disciplines, choices can reasonably be guided by a body of applied and theoretical literature. In newer fields and in fields separated from widespread computational ties, it can be daunting to knowledgeably consider different analysis choices because their perceived differences are influenced very strongly by speculation.

The work in this dissertation represents an attempt to create tools and increase analytic understanding for three different areas that, in some form or another, are in their nascency: metabolomics, massively parallel reporter assays (MPRAs), and data analysis/data science as a whole. A key part of this work is an emphasis on evidence-based recommendations. All methods and

conclusions that we present are based on evaluations based on real, publicly-available data.

Metabolomics is a branch of basic science that studies the small molecules that are present in biological systems. One of the main technologies that is used to collect measurements of these small molecule metabolites is mass spectrometry. Although mass spectrometry has been in widespread use for decades, existing tools that perform fundamental data processing have not been tailored to popular goals of the field, namely, comparative analysis. Mass spectrometry generates complex data that must be preprocessed to be amenable for statistical analysis. In this dissertation, we show that existing methods for preprocessing are ill-suited for the comparative analyses that practitioners are most often interested in. We develop a preprocessing method that facilitates comparisons by considering all samples simultaneously as opposed to individually. By evaluating our method on several real datasets, we show that our approach reduces unnecessary variability in preprocessing output and increases statistical power in differential analysis.

Massively parallel reporter assays (MPRAs) are newer assays that are emerging in popularity as a means of assessing the potential of a piece of DNA to regulate the transcription of a nearby gene. The main goals in these assays are to compare the regulatory activity of slightly different sequences and to explain variation in regulatory activity across sequences with genomic and biological features. Because the field is relatively new, the literature is replete with ad hoc statistical analyses. In this dissertation, we propose a unifying linear model analysis framework that draws upon established work

from RNA-sequencing literature. Using multiple publicly-available datasets, we show that our approach is well-calibrated and powerful in comparative analyses. We also formulate a mathematical model of data in this assay and use this model to provide practical advice regarding experimental design.

With reproducibility and replicability taking a more central role in scientific discourse, the research community has increasingly scrutinized the numerous stages of the scientific process, ranging from study design, to data analysis, to publication. Critiques of the data analysis stage tend to focus on specific methodology. By comparison, there has been little investigation of the cognitive aspects of data analysis. A data analysis involves numerous decisions that can be considerably subjective, and the cumulative impact of these decisions on an analyst's conclusions is likely substantial. In the final part of this dissertation, we present results from randomized experiments of human behavior and perception in data analysis situations.

Throughout this work, we place an emphasis on evidence-based decision making. In the context of methodological development in for biological studies, this consists of evaluation using real data in lieu of simulations. In the context of understanding human behavior in data analysis situations, we carry out experiments to supplement hypotheses that have been based primarily on conjecture.

Chapter 2

Metabolomics

Reproduced with permission from Myint, Leslie, Andre Kleensang, Liang Zhao, Thomas Hartung, and Kasper D. Hansen. 2017. "Joint Bounding of Peaks Across Samples Improves Differential Analysis in Mass Spectrometry-Based Metabolomics." Analytical Chemistry 89 (6):3517-23. https://doi.org/10.1021/acs.analchem.6b04719.

2.1 Introduction

As mass spectrometry-based metabolomics becomes a more mature and popular means of scientific investigation (Bouhifd et al., 2013; Bouhifd et al., 2015; Ramirez et al., 2013), it is important to revisit existing data analysis paradigms. Existing approaches to preprocessing metabolomics data focus on a two-step approach which starts by extracting features (peaks) separately from each sample, followed by a subsequent attempt to group features across samples to facilitate comparisons (Aberg, Alm, and Torgrip, 2009). In particular, there has

been considerable attention in the literature on individual stages of preprocessing, including peak detection (Hastings, Norton, and Roy, 2002; Vivó-Truyols et al., 2005; Du, Kibbe, and Lin, 2006; Noy and Fasulo, 2007; Tautenhahn, Böttcher, and Neumann, 2008; Chen et al., 2009; Nguyen et al., 2010; Shalliker et al., 2010; Vivó-Truyols, 2012; Fu et al., 2016) and alignment (Tomasi, Berg, and Andersson, 2004; Podwojski et al., 2009; Hoffmann et al., 2012; Jeong et al., 2012). Additional work has been done on specific issues with downstream differential analysis such as missing information or dependence structures (Tekwe, Carroll, and Dabney, 2012; Zhan, Patterson, and Ghosh, 2015; Taylor et al., 2017). Single sample processing methods tend to focus on reducing bias. The bias-variance tradeoff (Hastie, Tibshirani, and Friedman, 2011) shows that the overall performance of a method also depends on its noise, and experience from gene expression studies suggests that noise can be removed by processing samples jointly.

In this work, we investigate the consequences of traditional sample-specific preprocessing on the quality of differential analysis. We show that the retention time (RT) bounds that arise from preprocessing samples individually cause unnecessary variability in peak quantifications (based on integrated peak area) which leads to under-powered differential analysis. We propose a relative quantification method, called bakedpi, which addresses this shortcoming by jointly detecting and bounding peaks in the two-dimensional m/z-RT space, across all samples simultaneously. The backbone of our method is an intensity-weighted bivariate kernel density estimation that is computed on a

pooling of all samples. We show that this approach reduces unnecessary quantification variability and increases power in downstream differential analysis. Our method is open source and freely available as part of the yamss package through the Bioconductor project under Artistic License 2.0.

2.2 Results

2.2.1 Excess variability with sample-specific processing

To demonstrate issues with sample specific detection and bounding of peaks, we consider the widely used software packages XCMS (Smith et al., 2006) and MZmine2 (Pluskal et al., 2010). Output for one peak from a QTOF dataset with two sample groups is shown in Figure 2.1 (additional examples from other datasets in Supplementary Figures S1 and S2). The shape, width, and location of this peak do not appear to vary across samples. Despite this, the XCMS and MZmine2 RT bounds for this peak, indicated by blue and purple rectangles respectively, are highly heterogeneous between samples (Figure 2.1c). To a first approximation, the retention time (RT) bounds can be grouped into narrow and wide bounds; this grouping is not associated with the two sample groups (light and dark rectangles). As a consequence, the integrated peak area is completely determined by whether the RT bounds are narrow or wide (Figure 2.1d,e), and this leads to high variability in the peak quantifications (Figure 2.1f). If instead, we use the same RT bound across all samples (Figure 2.1c, orange rectangle), we substantially reduce the between-sample variability in the peak quantifications (Figure 2.1f). Excess variability results in loss of power in a differential analysis.

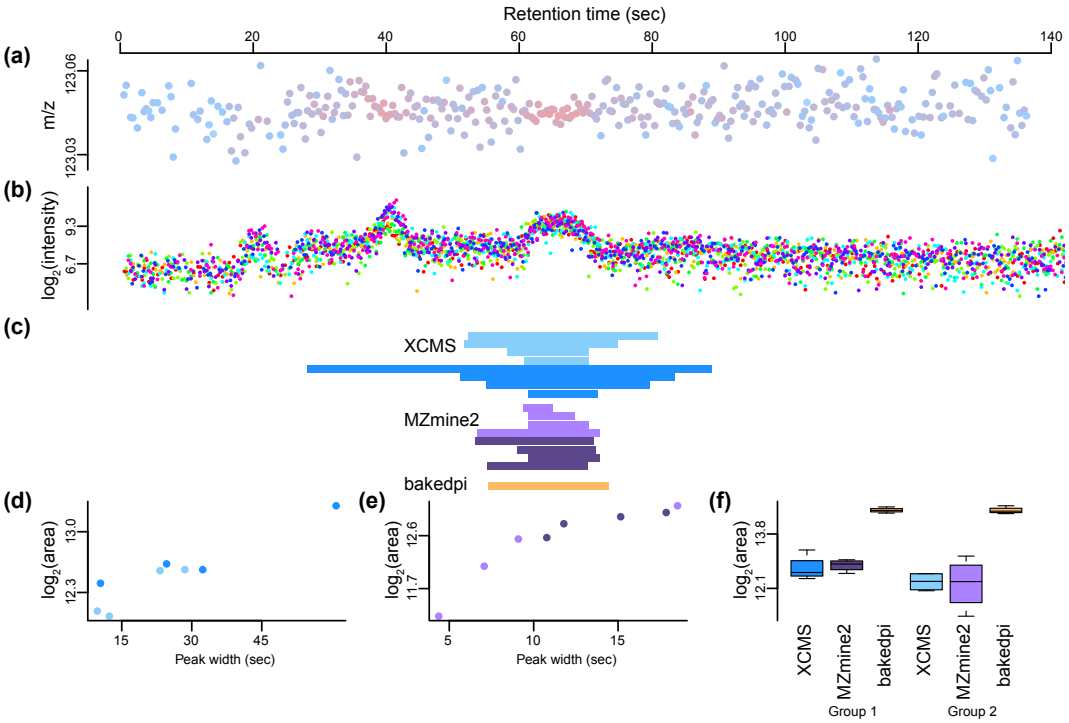


Figure 2.1: Problems with sample-specific processing in XCMS and MZmine2. Peak detection and bounding for a single peak in the MTBLS2_rep1 dataset. **(a)** The m/z-RT space surrounding this peak for a single sample, color is used to depict intensity (red is high). **(b)** Overlaid extracted ion chromatograms from all 8 samples in the experiment. Different colors denote different samples. **(c)** The peak bounds for all samples for XCMS (blue), MZmine2 (purple) and bakedpi (orange; all samples have same bounds). This experiment compares two groups of samples indicated with different color shades. **(d)** XCMS peak quantification vs. peak width. **(e)** Like (d) but for MZmine. **(f)** Distribution of peak quantifications, based on the peak bounds in (c). Substantial heterogeneity in the sample-specific bounds leads to excess variability in the quantifications; this is addressed by using the same RT bound for all samples.

2.2.2 Joint sample processing with bakedpi

To address the problem of excess variability, we propose a method which jointly detects and bounds peaks across all samples in an experiment (see Methods); an important feature of our method is the use of homogeneous RT bounds across all samples. We pool the data from all samples into a

single metasample, on which we detect and bound peaks (Figure 2.2a,b). To do this, we use intensity-weighted bivariate kernel density estimation in the two-dimensional m/z-RT space. By using the intensities as weights, we differentiate between groups of detected m/z values (data points) with high and low intensities. The output is a smooth density in the m/z-RT space, where peaks in the density correspond to clusters of high-intensity points (Figure 2.2c). To detect and bound peaks, we slice the density using a single global threshold, and form a set of contiguous regions based on the density slices. By performing this procedure on a single metasample, we ensure the same peak bounds across all samples. Like XCMS and MZmine2, we quantify the peaks by integrating the extracted ion chromatogram (EIC) for each sample across the peak's RT bounds. We can optionally perform RT alignment prior to density estimation. Our method has 3 parameters: 2 of these parameters control the bandwidth in the m/z and RT domains and are easy to set based on the resolution of the instrument. The last parameter, the only significant tuning parameter, is the global density threshold. We call our method bakedpi, for bivariate approximate kernel density estimation for peak identification.

2.2.3 Joint sample processing reduces excess variability

We applied bakedpi to 10 different datasets from 7 different experiments. Features of these datasets are summarized in Table 2.1. All datasets were subset (if necessary) to only contain two sample groups, to keep the experimental design simple and constant. For the Orbitrap dataset (MTLS216) we expect little to no differences between the sample groups, based on the design of the

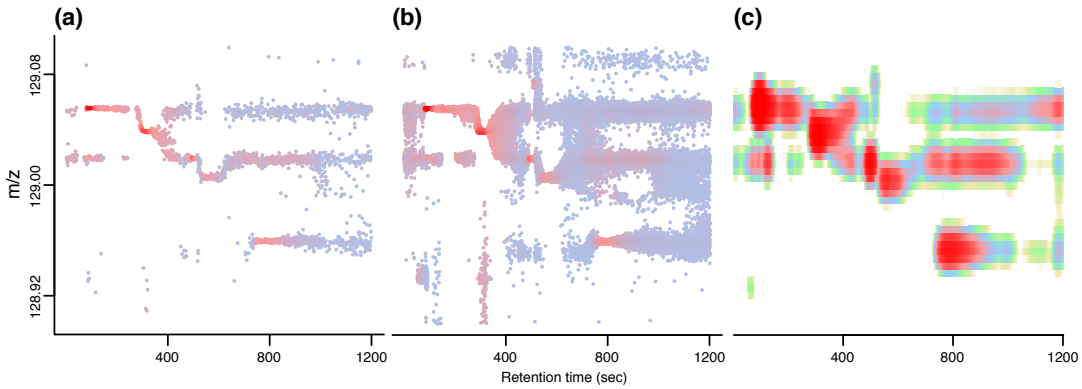


Figure 2.2: Weighted bivariate kernel density estimation. We depict a selected rectangle in m/z-RT space for **(a)** one sample and **(b)** the pooled metasample. m/z values with higher intensity are shown in red, lower with blue. **(c)** The weighted bivariate density estimate.

experiment (Murakami et al., 2014). We ran XCMS, MZmine2, and bakedpi on the 10 datasets. XCMS parameters were optimized using the IPO package available on Bioconductor (Libiseller et al., 2015) using recommended starting values for most datasets (Methods). MZmine2 parameters were set based on optimized XCMS parameters where possible (Methods). When running bakedpi, we use the higher of a fixed quantile cutoff and a data-driven cutoff to set the global tuning parameter (Methods).

To compare the quantification variability between bakedpi and XCMS and between bakedpi and MZmine2, we first identified peaks which overlapped between bakedpi and XCMS and between bakedpi and MZmine2. We will call these shared peaks. The number of peaks detected by both methods as well as the percentage of peaks that are common to both methods are shown in Supplemental Figure S3; for many datasets the overlap is around 60-80% of the peaks. On these overlapping peaks, we computed the residual standard deviation of the log-abundances to assess their variability. We used

Name (Source)	MS instrument Column	# samples (group 1, 2)
ASD_hirisk (C)	QTOF HPLC - HILIC	20, 20
timecourse_4hr (C)	QTOF HPLC - HILIC	6, 6
timecourse_24hr (C)	QTOF HPLC - HILIC	6, 6
MTBLS2_rep1 (M)	QTOF UPLC - reverse phase	4, 4
MTBLS2_rep2 (M)	QTOF UPLC - reverse phase	4, 4
CAMERA_pos (M)	QTOF UPLC - reverse phase	3, 3
CAMERA_neg (M)	QTOF UPLC - reverse phase	3, 3
MTBLS103 (M)	QTOF UPLC - HILIC	14, 12
MTBLS213 (M)	QTOF UPLC - reverse phase	6, 6
MTBLS126 (M)	Orbitrap HPLC - HILIC	3, 3

Table 2.1: Characteristics of evaluation datasets. C = CAAT, M = Metabolights

residual standard deviation to avoid being influenced by changes in the log-abundances between the two sample groups in the different experiments. Figure 2.3 shows the distribution of differences in residual standard deviation (XCMS or MZmine2 minus bakedpi) for each dataset. Values greater than zero indicate that bakedpi has smaller variability than the other method. For all datasets examined, more than half of the peaks detected by both methods had lower variability when quantified by bakedpi; for some datasets it was substantially higher.

2.2.4 Joint processing improves power in a differential analysis

We next sought to determine if the decrease in residual standard deviation of the peak quantifications leads to increased power in a differential analysis. We used the limma (Smyth, 2004) differential analysis pipeline as it has been

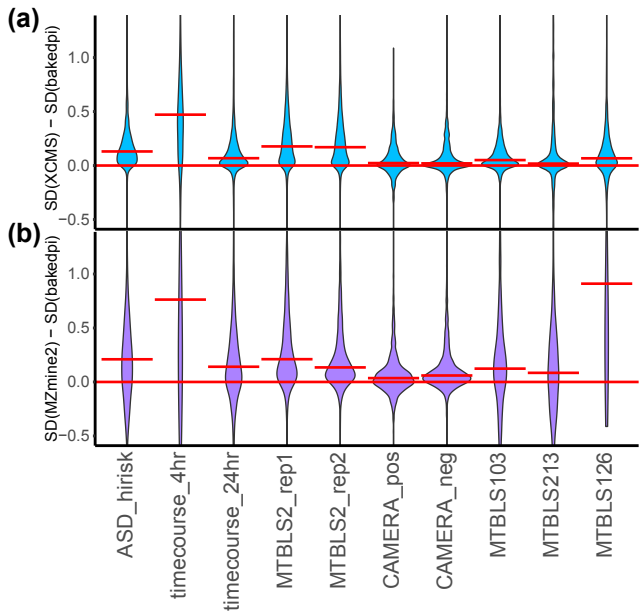


Figure 2.3: Variability comparison of peak quantifications. (a) For peaks that are detected both by bakedpi and XCMS, the distribution of the differences in residual standard deviation for all datasets are shown as violin plots. Each violin is a mirrored density plot; the median is indicated by a horizontal red line. **(b)** Like (a) but for MZmine. For all datasets, the majority of peaks detected by both methods have quantifications that are less variable when quantified with bakedpi.

shown to provide robust and powerful inference for proteomics data (Kammers et al., 2015). This method was originally developed to analyze microarray expression studies and uses empirical Bayes techniques to shrink feature (adduct)-wise variances towards a common underlying value to provide more stable inference. The resulting p-value distributions for the shared peaks in the timecourse_4hr dataset are shown in Figure 2.4a (additional datasets in Supplementary Figure S4). For the majority of the datasets, bakedpi has a p-value distribution that is more peaked around zero than XCMS and MZmine2, indicating that bakedpi detects more significant peaks among the overlapping peaks. When comparing with XCMS, the timecourse_24hr dataset is the only

one in which XCMS has a taller peak arount zero. When comparing with MZmine2, only for the CAMERA_pos dataset does MZmine2 have a taller peak around zero.

Higher detection rates alone do not necessarily indicate an increase in power. To assess power, we also evaluated the type I error control of the methods. We performed a permutation experiment in which we shuffled the sample group labels so that each of the new comparison groups were composed half of cases and half of controls. For example, in an experiment with eight cases and eight controls, the new permuted "case" group would include four true cases and four true controls, as would the new permuted "control" group. In this way, we created null datasets in which no abundance differences are expected. With datasets containing a sufficient number of samples, we performed 1000 permutations. Otherwise we enumerated all permutations satisfying the balancing characteristic just described. We again used limma to perform differential testing. Results of the permutation experiment for the timecourse_4hr dataset are shown in Figure 2.4c (additional datasets in Supplementary Figure S5). For a range of nominal type I error rates, we computed the median observed error rate over all permutations. For all ten datasets, all methods are quite conservative, showing a markedly lower error rate than the nominal value for the entire range. For most of the datasets, bakedpi is the most conservative of the three methods. The combination of more conservative type I error control and a higher detection rate indicates that bakedpi has higher power to detect differences than the sample-by-sample processing procedures of XCMS and MZmine2.

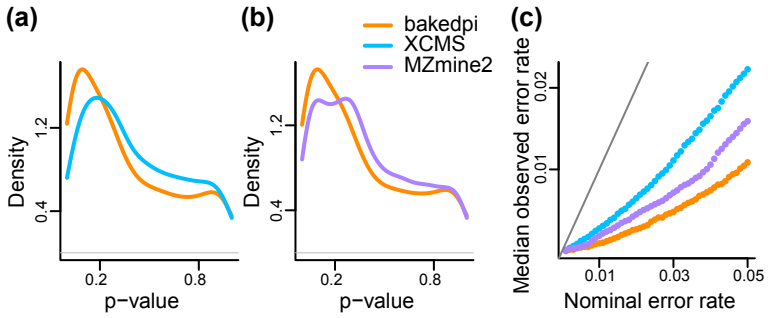


Figure 2.4: Comparison of differential analysis quality and type I error control in the timecourse_4hr dataset. (a) Distribution of p-values for peaks detected by both bakedpi and XCMS, **(b)** Like (a) but for MZmine, **(c)** Median error rate over null permutations as a function of the nominal error rate.

2.2.5 Retention time alignment

It is well established that RT deviations between experimental runs can complicate the matching of peaks across samples. We investigated the impact of correcting RT drift on the variability improvements of our method using multiple strategies. First, we used the RT warping function computed by XCMS to align the raw data before computing the density estimate. Second, we computed local sample-specific RT shifts that maximized the correlation of the chromatograms between samples and used these shifts to align the raw data. Third, we used correlation-optimal shifts to align peaks already detected from the density estimate before quantification. None of these RT alignment strategies had a large impact on the variability of detected features. The proportion of peaks detected by both bakedpi and XCMS or MZmine2 that had lower variability with bakedpi did not change appreciably with these RT corrections (Supplementary Figure S6).

2.2.6 Parameter choices

Because the detection of peaks and their bounds depend on the cutoff applied to the density estimate, it is important to investigate the sensitivity of method performance to this cutoff. We performed a sensitivity analysis by varying the density cutoff and examining the p-value distribution resulting from the detected peaks (Supplementary Figures S7 and S8). Raising the cutoff to be more stringent or lowering the cutoff to be more inclusive generally does not have a substantial impact on the global pattern of inference as assessed by p-value distributions.

2.2.7 Method-specific peaks

There are a number of peaks that are detected only by one method (Supplementary Figure S3). As comprehensive gold standard information on the true peaks corresponding to compounds was not available, we examined the characteristics of these method-specific peaks to assess their quality (Supplementary Figures S9-S12). For more than half of the datasets, XCMS-specific peaks tend to have more extreme t-statistics and lower p-values. For half the datasets, MZmine2 peaks have higher p-values than bakedpi. For nearly all datasets, bakedpi-specific peaks have greater peak heights than XCMS- and MZmine2-specific peaks with comparable peak widths. Peaks specific to bakedpi are also more likely to be supported by all samples in the experiment. The last two observations are sensible given that bakedpi relies on an intensity-weighted density estimation; a peak is more likely to be detected when a large number of high-intensity points are close together. Based on

observations about t-statistics and p-values, it is not clear that one of the two sets of methods-specific peaks is best. If peaks with greater heights or greater numbers of samples supporting them are more likely to be of scientific interest, then bakedpi-specific peaks seem to be of higher quality than XCMS- or MZmine2-specific peaks. Given the lack of gold standard data on peak identities, evaluation of method-specific peaks is less clear than evaluation of peaks common to multiple methods. On peaks common to both bakedpi and MZmine2 or XCMS, bakedpi shows a clear reduction in quantification variability and an increase in statistical power.

2.3 Conclusions

We have proposed a method for the joint processing of metabolomics data across samples, which reduces variability in peak quantification across samples, leading to increased power in a differential analysis. We take the position that the most important task in metabolomics is the identification of differentially abundant peaks, in contrast to eg. identifying all peaks in a sample. Our method compares favorably to XCMS and MZmine2 across ten datasets, and will be useful for drawing better and more substantiated inferences from untargeted metabolomics studies. We do note that the commercial software Progenesis CoMet also uses the idea of pooling samples into a metasample for processing. However, details on CoMet method are not available, making it impossible to comment further on differences between the two approaches.

A limitation of our approach is that peaks that are only truly present in a small fraction of the samples are unlikely to be detected. Such metabolites

may be of interest, but are by definition less well supported by the observed data. In developing bakedpi, we have chosen to focus on peaks with sufficient information across all samples and on obtaining for those peaks the best quality quantifications for the purposes of differential analysis.

It is important to note that the benefit of our method is dependent on using peak areas for quantification rather than peak height. As we show, the variability in quantification of a particular peak across samples is driven entirely by the variability in peak width. If peak height is used instead of peak area, our method will show the same quantification as XCMS and MZMine2, provided the sample-specific RT bounds contain the mode of the peak; this is true for two of our three examples.

In our evaluation of bakedpi, we have used both centroid-mode and profile-mode datasets with fairly stable chromatography. The RT drift we observe in these datasets is not so large that corresponding peaks from different samples do not overlap. However, stable chromatography is not required for bakedpi to work because we do implement RT alignment procedures. Our evaluation datasets also come from mass spectrometers with a range of mass accuracies from 5 ppm on Q-TOF instruments to less than 1 ppm on the Orbitrap, so bakedpi is able to handle data from a representative range of instruments. We expect lower mass accuracy to make peak merging more likely and to cause peak m/z bounds to be wider than necessary, but this is mostly a feature of low mass accuracy in general. Currently, our method is implemented as the standalone yamss package as part of the Bioconductor project.

2.4 Methods

2.4.1 Data

Also see Table 2.1.

ASD_hirisk: Prenatal serum samples from 40 mothers participating in the EARLI study whose infants had the highest (n=20) and lowest (n=20) Autism Observation Scale for Infants (AOSI) at the time of experiment (Newschaffer et al., 2012).

timecourse_4h, timecourse_24hr: Six MCF-7 cell line samples exposed to estradiol (E2) and six control samples unexposed to E2 for up to 72 hours (Kleensang et al., 2016).

MTBLS2: Four wild-type and four *cyp79b2 cyp79b3* knockout *Arabidopsis thaliana* leaves exposed to silver nitrate (Böttcher et al., 2009; Neumann, Thum, and Böttcher, 2012).

CAMERA: Spike-ins of 39 known compounds at varying concentrations on methanolic extracts of *Arabidopsis thaliana* leaves (Kuhl et al., 2012). Three samples with a spike-in concentration of 20 μM were compared to three samples with a spike-in concentration of 5 μM in both positive and negative ion mode.

MTBLS103: Serum profiling of 12 adolescent girls with hyperinsulinaemic androgen excess and 14 healthy controls matched on age, weight, and ethnicity (Samino et al., 2015).

MTBLS213: Human retinal pigment epithelium cell line (ARPE-19) batches grown labeled and unlabeled glucose media (Capellades et al., 2016).

MTBLS126: Liver concentrations of resveratrol (RESV) metabolites after application of a mixture of RESV in hydrophilic ointment to mouse skin (3 samples) compared to liver concentrations of resveratrol (RESV) metabolites after application of hydrophilic ointment without RESV to mouse skin (3 samples) (Murakami et al., 2014).

2.4.2 Processing with XCMS and MZmine2

XCMS parameters were optimized using the IPO package available on Bioconductor (Libiseller et al., 2015) using recommended starting values for most datasets. Because optimization for the MTBLS2 and MTBLS213 datasets required significant computational time (we terminated the optimization after 11 days), we either fixed parameters that could be reasonably inferred beforehand (such as ppm) or set a smaller range of values over which to optimize. MZmine2 parameters were set based on optimized XCMS parameters where possible. In particular, the "prefilter", "mzdiff", minimum and maximum peakwidth, and ppm parameters from XCMS had near equivalents in MZmine2 parameters. For XCMS, we used the "centWave" algorithm (Tautenhahn, Böttcher, and Neumann, 2008) for the nine centroid-mode datasets and the "matchedFilter" algorithm (Smith et al., 2006) for the profile-mode MTBLS126 dataset. We used the density method for peak grouping, the obiwarp method for retention time alignment, and the fillPeaks method to fill in information for peaks missing from certain samples. For MZmine2, we used the GridMass module for peak detection (Treviño et al., 2015), the join aligner for retention time alignment, and the same-range gap filler module. Details

on optimization and parameter settings for XCMS and MZmine2 are provided in the Supplemental Information.

2.4.3 Processing workflow

Our processing procedure consists of three steps. First is background correction which increases the signal to noise ratio of true peaks. Second is RT alignment which aligns the raw data to correct for drifts in compound elution times between samples; this is optional. Third is density estimation to detect peaks.

2.4.4 Background correction

Background correction is performed on each sample separately. We divide the m/z-RT space into bins and estimate background separately for each bin; this is arbitrarily done for bins of width 10 m/z units and 40 scans in the RT domain. We observe that each grid region exhibits a multi-modal intensity distribution with 2 or more modes (Supplementary Figure S13), and reason that the lowest mode is background. We estimate the location of the mode with the first peak of the kernel density estimate of the intensity distribution and subtract this value from all observations in the grid region.

2.4.5 Retention time alignment

We investigated two RT alignment procedures that could be applied to the raw data before peak detection and one procedure that could be applied after

peak detection. The first pre-peak detection approach was to use the sample-specific corrected RTs reported by XCMS to define a RT warping function that could be applied to the raw data to yield aligned RTs. In the second approach, we found tentative m/z regions containing peaks using univariate kernel density estimation and computed EICs in these regions for all samples. For each region and sample, we then found the shift that would maximize the correlation between the EICs in each sample and a reference sample (the sample with the largest area beneath the EIC). These local and sample-specific shifts were applied to the raw data to yield aligned RTs. We also investigated a correction procedure that could be applied to peaks that had already been detected. For each detected peak, we computed the sample-specific shifts that would maximize the correlation between the EICs in each sample and a reference sample (the sample with the largest area beneath the EIC). We then recomputed the peak quantifications using the original RT bounds and shifted EICs.

2.4.6 Bivariate density estimation

To detect peaks, we pool all samples into a single metasample by concatenating the spectral information from all of the samples. For example, the spectral information for the first scan of the metasample is formed by concatenating the first scan's spectral information from the individual samples. We use this metasample to estimate a two-dimensional density in the m/z-RT space. We represent the input data as a set of datapoints (M_j, T_j, I_j) where M_j is the mass over charge (m/z) of the j'th datapoint (all samples are pooled), T_j is the scan

number (RT in seconds divided by number of scans per second) and I_j is the intensity. Per sample, T typically has up to a few thousand unique values depending on the scan rate of the mass spectrometer and the duration of the experiment, and M has on the order of one hundred observations per scan in centroid-mode data and several hundred in profile-mode data. Thus the data consists of tens of thousands of datapoints such triples for each sample.

The bivariate intensity-weighted density estimator using a Gaussian kernel at a point (m, t) in m/z-RT space is given by

$$\hat{f}(m,t) = \frac{1}{h_M h_T \sum_{j=1}^{n} I_j} \sum_{j=1}^{n} I_j \phi_2 \left(\frac{m - M_j}{h_M}, \frac{t - T_j}{h_T} \right)$$

where $j = 1, \ldots, n$ indexes the n datapoints, h_M and h_T are the bandwidths in m/z and RT space respectively, and ϕ_2 is a bivariate Gaussian density. The density estimate is not highly sensitive to the RT bandwidth, and a default of bandwidth of 10 scans is recommended. The m/z bandwidth should be set based on the type of mass spectrometer used and is recommended to be 0.005 for TOF and 0.002 for Orbitrap instruments. Because the density estimate involves a sum over all n datapoints at each value of (m, t), we use various approaches to make this computationally tractable. First, we use a diagonal covariance matrix for the bivariate kernel; this implies the factorization

$$\phi_2 \left(\frac{m - M_j}{h_M}, \frac{t - T_j}{h_T} \right) = \phi_1 \left(\frac{m - M_j}{h_M} \right) \phi_1 \left(\frac{t - T_j}{h_T} \right)$$

We do this because our focus is on identifying regions of interest rather than on highly exact estimation of the density (Duong, 2007). Second, we use a simple binning strategy (Wand, 1994) where the m/z-RT space is binned and

a single representative value for each bin is chosen. In the RT domain, the default bin width is 1 scan, and in the m/z domain the default bin width is set to be equal to the bandwidth (0.005 for TOF and 0.002 for Orbitrap). Third, we use a Gaussian kernel truncated at ± 3, effectively only including points close to (m, t) in the summation (Wand, 1994). Fourth, in our implementation, we make use of sparse linear algebra as well as efficient data structures for selecting points close to (m, t) as implemented in the `data.table` package (Dowle et al., 2015).

After obtaining the density estimate, we select a cutoff using information from the strongest (most intense) features in the data. The m/z domain is divided into bins of a default width of 2 m/z. Within each bin, the most intense data point is selected. We assume that this data point belongs to a true feature and use local univariate density estimation in the m/z and RT domains to define a m/z and RT window for this feature. We compute quantiles of the density estimate values in these regions and compute the mode of this quantile distribution for various quantile values. For example, we compute the 99th percentile for each of the approximately 500 strong feature regions and compute the mode of this distribution. We repeat this for a wide range of percentiles. We then order these modes and select the first mode substantially different from zero as a cutoff. To ensure reasonable peak bounds, we enforce that this cutoff should be greater than or equal to the 99th percentile of nonzero density values. Applying the cutoff to the density estimate matrix yields a binary matrix that denotes peak and non-peak regions. In order to obtain m/z and RT bounds for these peak regions, we use a connected components

labeling algorithm (Pau et al., 2010).

2.4.7 Software availability

Our method is implemented in the `yamss` package, available from the Bioconductor project at https://www.bioconductor.org/packages/yamss.

Acknowledgements

Funding: Research reported in this publication was supported by National Institute of Environmental Health Sciences of the National Institutes of Health under award number R01ES020750 and the National Cancer Institute of the National Institutes of Health under award number U24CA180996. This research was supported by a Johns Hopkins Bloomberg School of Public Health Faculty Innovation Fund award. The EARLI study was funded by R01ES016443 and Autism Speaks 9502. Some EARLI participants were recruited with the assistance of the Interactive Autism Network (IAN) database at the Kennedy Krieger Institute, Baltimore MD.

Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest: None declared.

2.5 Supplemental Information

2.5.1 Supplementary Methods

2.5.1.1 XCMS parameters

Data available in centroid mode was processed with the `xcmsSet` command with the "centWave" method. Profile mode data was processed with the "matchedFilter" method. Parameter optimization was performed with the IPO package availabe on Bioconductor. Optimization was performed on a subset of two samples for each dataset. These two samples were chosen to be the one in each sample group that had the largest total intensity (integrated area beneath the total ion chromatogram) because these were expected to have the richest set of peak information. In most cases we used the default starting parameters for the optimizations (obtained with the `getDefaultXcmsSetStartingParams` and `getDefaultRetGroupStartingParams` functions). The signal-to-noise threshold value is not optimized by default, but we optimized it by setting the starting parameters with

```
snthresh = c(3,8)
```

We also optimized the prefilter values with the following starting parameters

```
prefilter = c(2,3)
prefilter_value = c(200,300)
```

We used the default starting parameters for retention time alignment and grouping optimization with the exception of the MTBLS2 and MTBLS213 datasets.

Due to optimization running times in excess of 11 days, we modified the starting parameters for the two MTBLS2 datasets as follows to match the parameters given in the original paper.

```
min_peakwidth = c(5,12)
max_peakwidth = c(12,35)
prefilter = 3
prefilter_value = 200
snthresh = 5
ppm = 25
minfrac = 0.75
```

For the same reason, we modified the starting parameters for the MTBLS213 dataset as follows to match the parameters given in the original paper.

```
min_peakwidth = c(5,12)
max_peakwidth = c(20,35)
ppm = 30
minfrac = 0.5
```

After obtaining optimized parameters, we ran the `xcmsSet` command followed by `group` using the "density" method, `retcor` with the "obiwarp" method, and finally `fillPeaks`.

2.5.1.2 MZmine2 parameters

We ran MZmine2 version 2.21 with the GridMass - 2D peak detection procedure, the join aligner for retention time alignment and grouping, and the

same-range gap filler module. To the best of our knowledge, there is no automated method for obtaining optimized MZmine parameters, so we translated optimized XCMS parameters to MZmine parameters as follows.

GridMass peak detection

- Minimum height: use optimized prefilter value from XCMS
- M/Z Tolerance: use optimized mzdiff from XCMS unless negative. If negative, use 100*optimized XCMS ppm/1e6.
- Min-max width time (in minutes): use optimized minimum and maximum peak width from XCMS multiplied by 60 to convert to minutes
- Smoothing M/Z: use 0.5*M/Z tolerance as this parameters is recommended to be smaller than the m/z tolerance
- Intensity similarity ratio: the default 0.5 was used
- Ignore times: the default of no times ignored was used

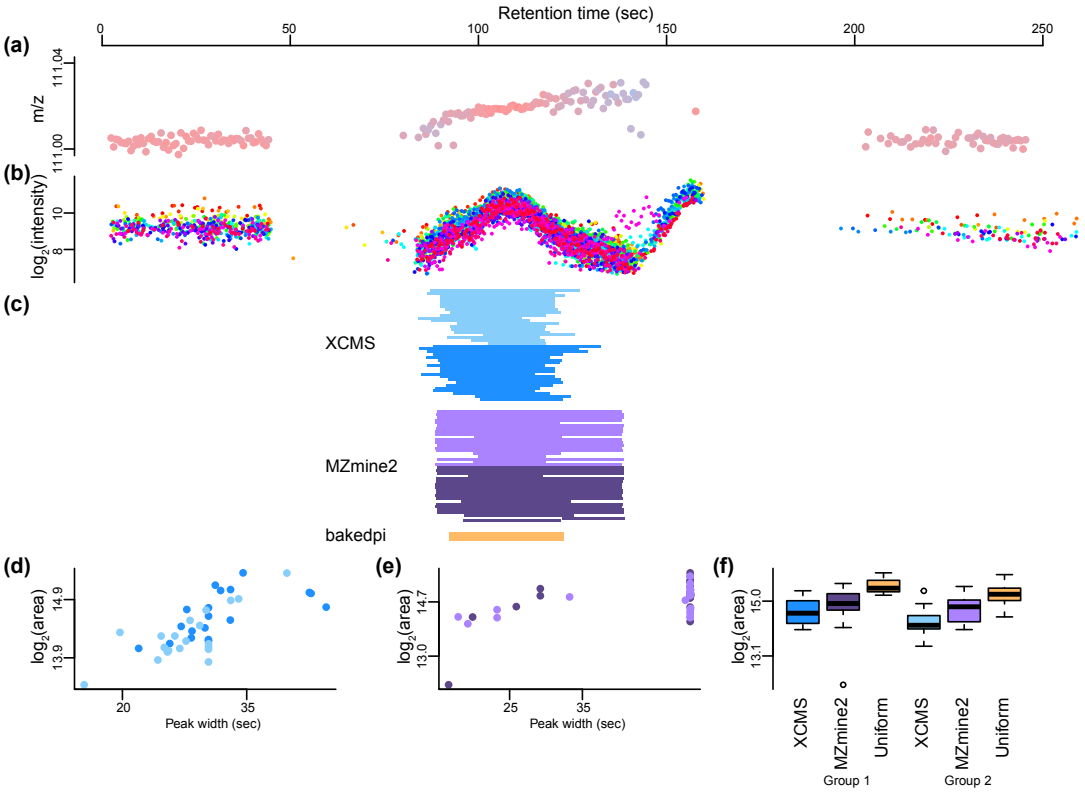
Join aligner

- m/z tolerance: We used 0.005 m/z for the absolute tolerance and the optimized XCMS ppm for the ppm tolerance.
- RT tolerance: We used the maximum peak width from XCMS
- Weight for M/Z and RT: We set these both to 1

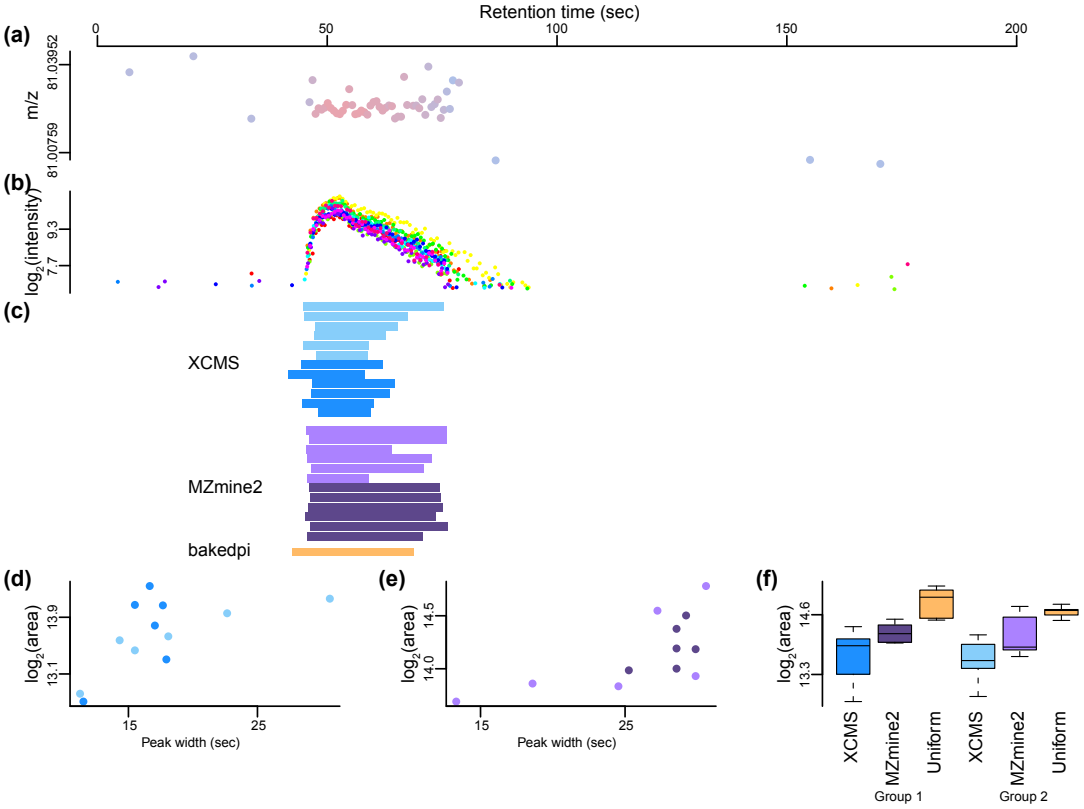
Same-range gap filler

- m/z tolerance: We used 0.005 m/z for the absolute tolerance and the optimized XCMS ppm for the ppm tolerance.

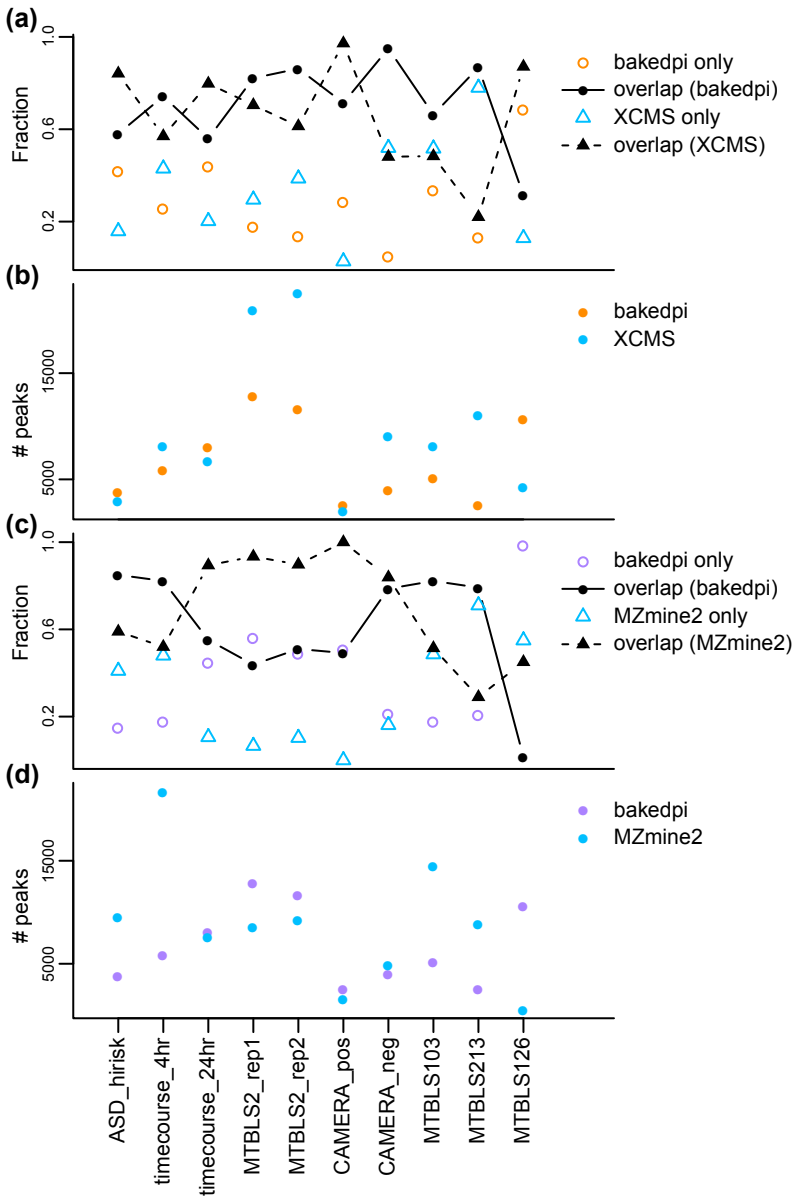
2.5.2 Supplementary Figures



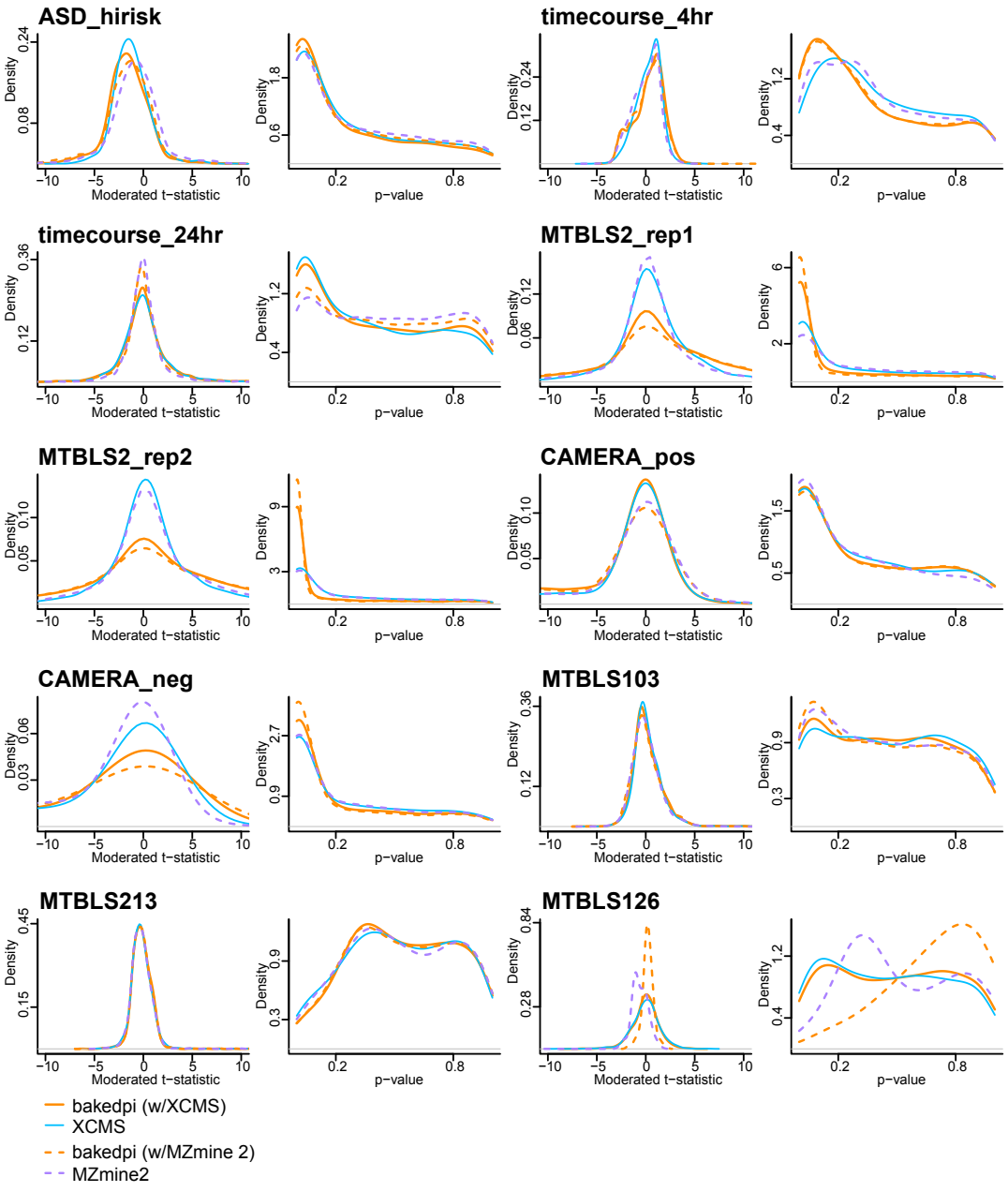
Supplementary Figure S1: Problems with XCMS and MZmine2 processing. Like Figure 1, but from the ASD_hirisk dataset. **(a)** The m/z-RT space surrounding this peak in a single sample, color is used to indicate intensity (red is high). **(b)** Overlaid extracted ion chromatograms from all 40 samples in the experiment. Different colors denote different samples. **(c)** The peak bounds for all samples for XCMS (blue), MZmine2 (purple) and bakedpi (orange; all samples have same bounds). This experiment compares two groups of samples indicated with different color shades. **(d)** XCMS peak quantification vs. peak width. **(e)** Like (d) but for MZmine. **(f)** Distribution of peak quantifications, based on the peak bounds in (c). Substantial heterogeneity in the sample-specific bounds leads to excess variability in the quantifications; this is addressed by using the same RT bound for all samples.



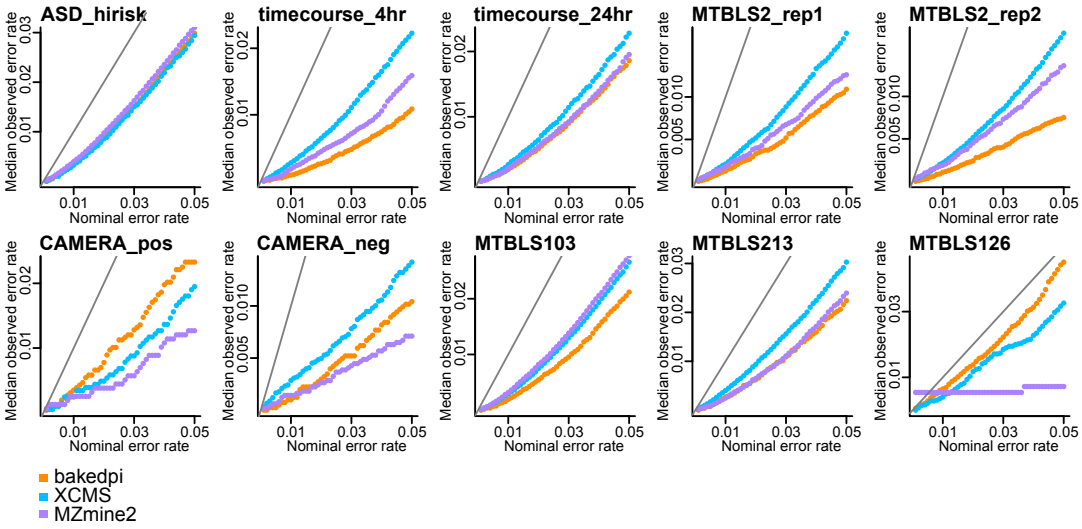
Supplementary Figure S2: Problems with XCMS and MZmine2 processing. As Supplemental Figure S1, depicting an example from the timecourse_4hr dataset.



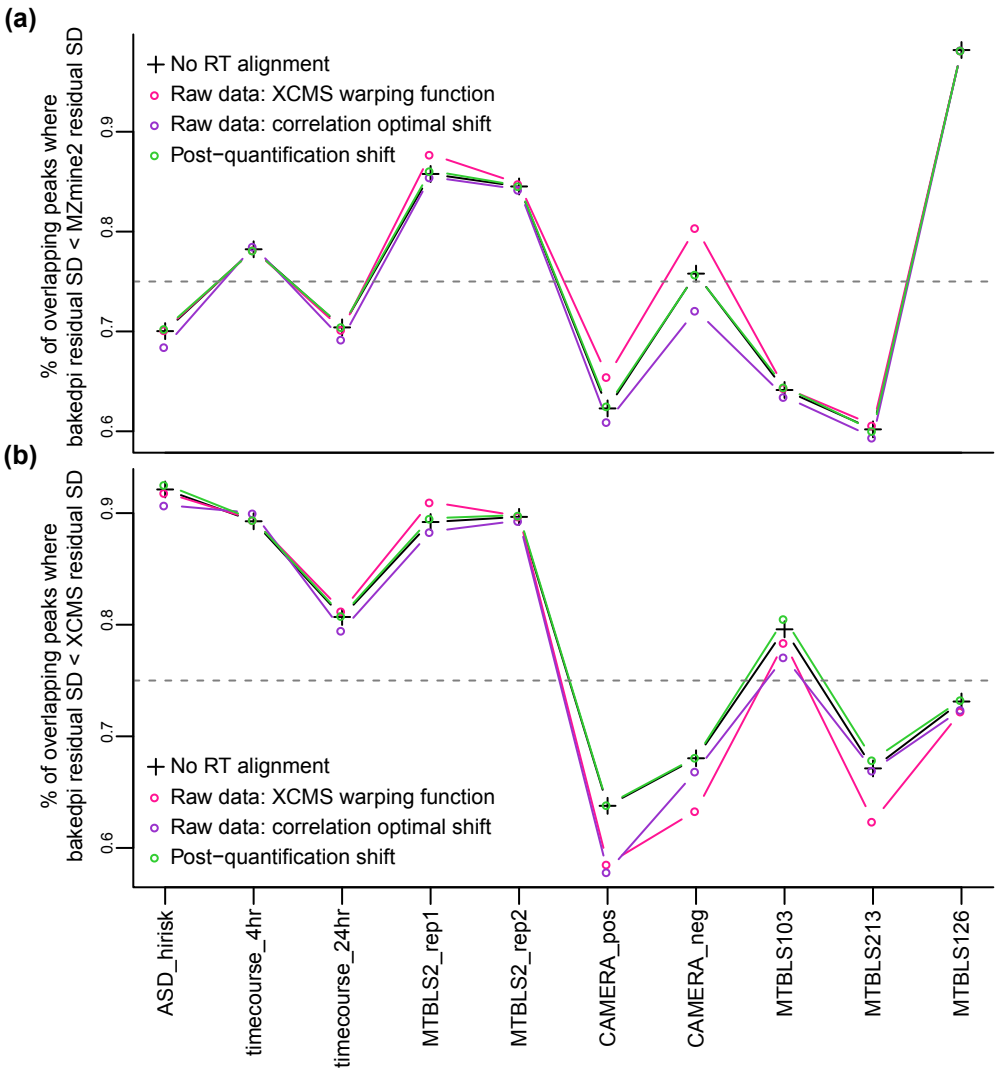
Supplementary Figure S3: Number of peaks called and overlap between methods. **(a)** The peaks detected by bakedpi are split into two groups: those that are only detected by bakedpi and those that are also detected by XCMS (orange and black circles). XCMS peaks are split similarly (blue and black triangles). **(b)** The number of peaks detected by bakedpi and XCMS. **(c), (d)** Like (a), (b) but for the bakedpi-MZmine2 comparison. In most datasets, bakedpi and the comparison method detect a similar number of peaks, a large percentage of which are found by both methods. Still for nearly all datasets, there is a sizable number of peaks which are only detected by one method.



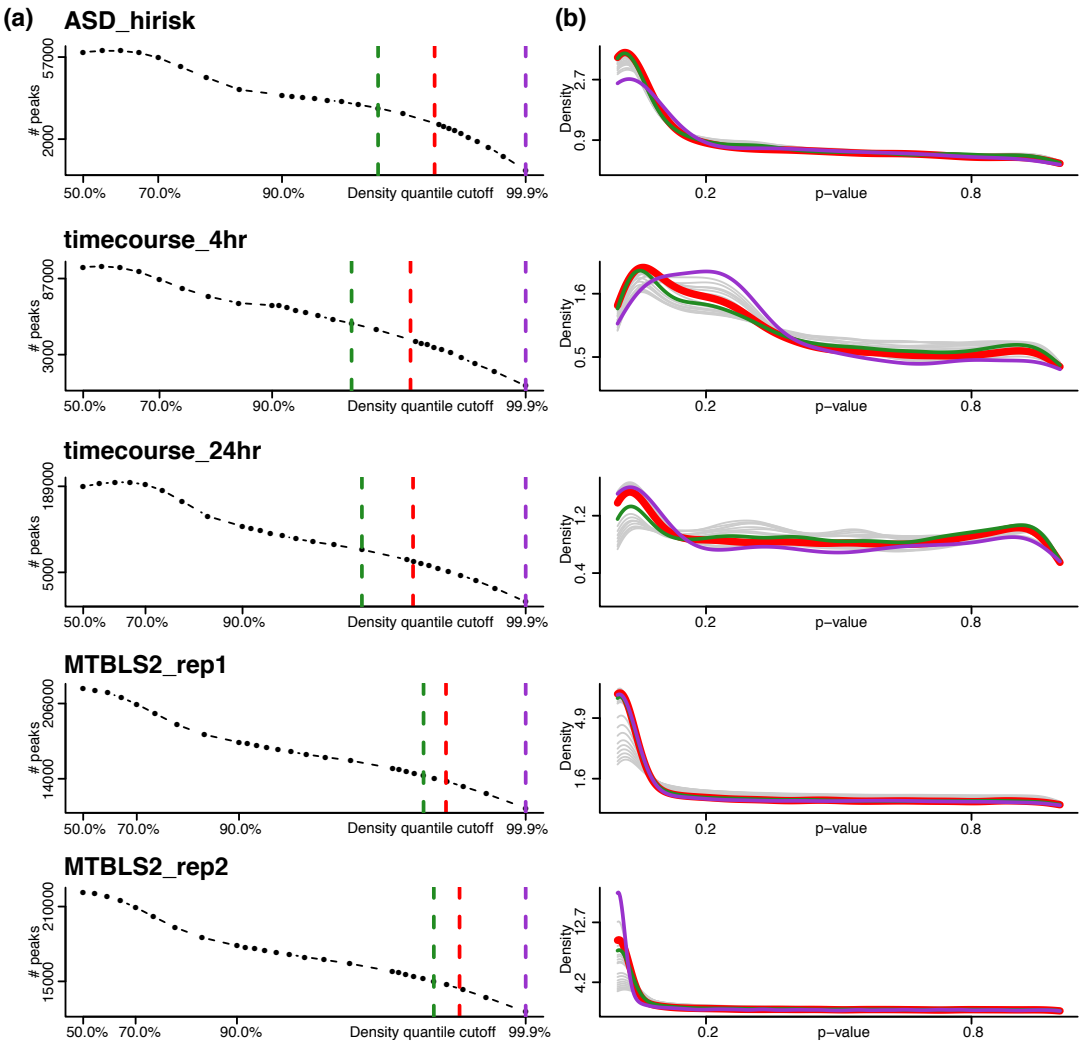
Supplementary Figure S4: Comparison of differential analysis quality in peaks detected by both bakedpi and either XCMS or MZmine2. The `limma` package was used to perform differential abundance analysis on quantifications from bakedpi and XCMS. Shown here are the distributions of the moderated t-statistics and associated p-values for the peaks detected by both bakedpi and XCMS (solid lines) and for the peaks detected by both bakedpi and MZmine2 (dotted lines).



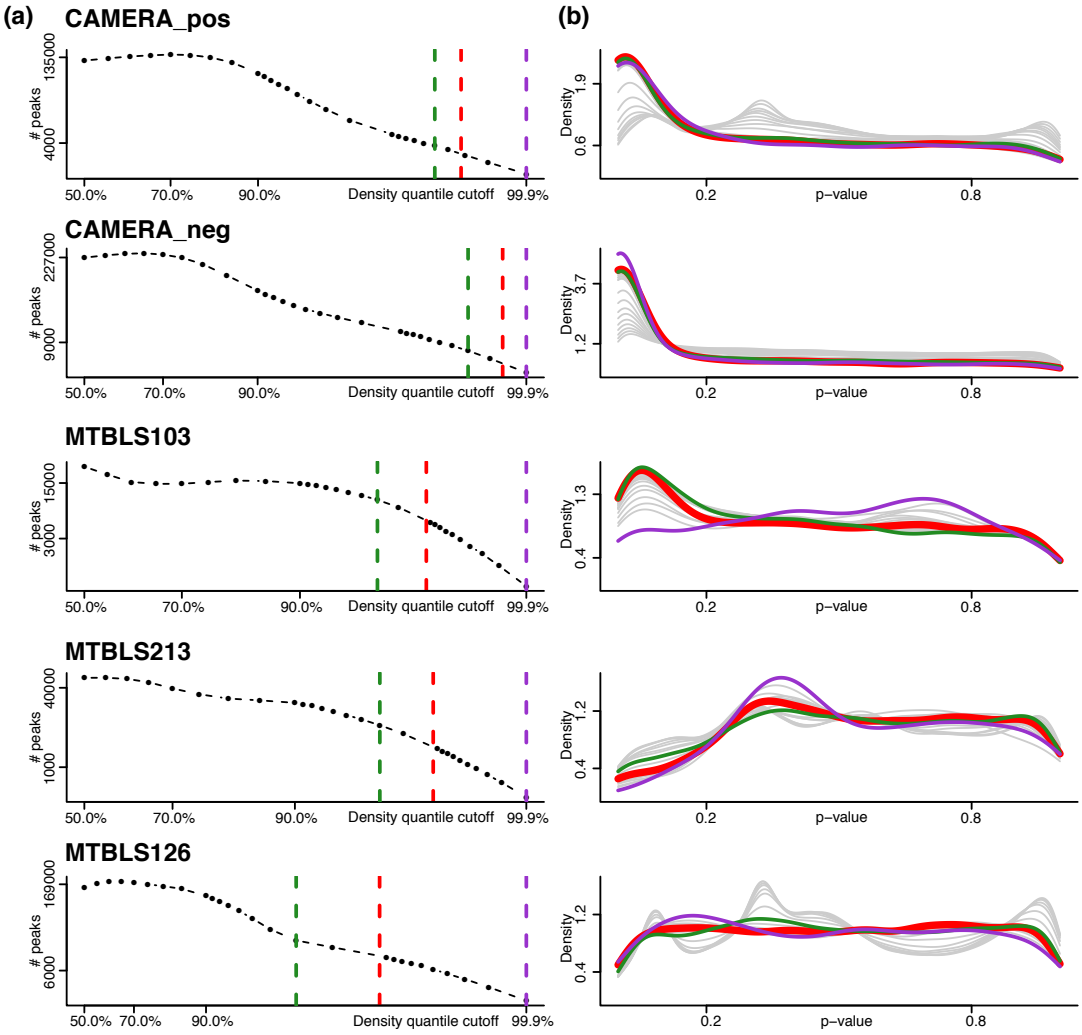
Supplementary Figure S5: bakedpi has more conservative type I error control than XCMS and MZmine2. For each dataset, sample labels were permuted to create null comparisons in which the new permuted groups both had an equal mix of original case and control samples. The median error rate over these null permutations is shown as a function of the nominal error rate. For all datasets, both bakedpi and XCMS are conservative, and for most datasets, bakedpi is as or more conservative than XCMS and MZmine2.



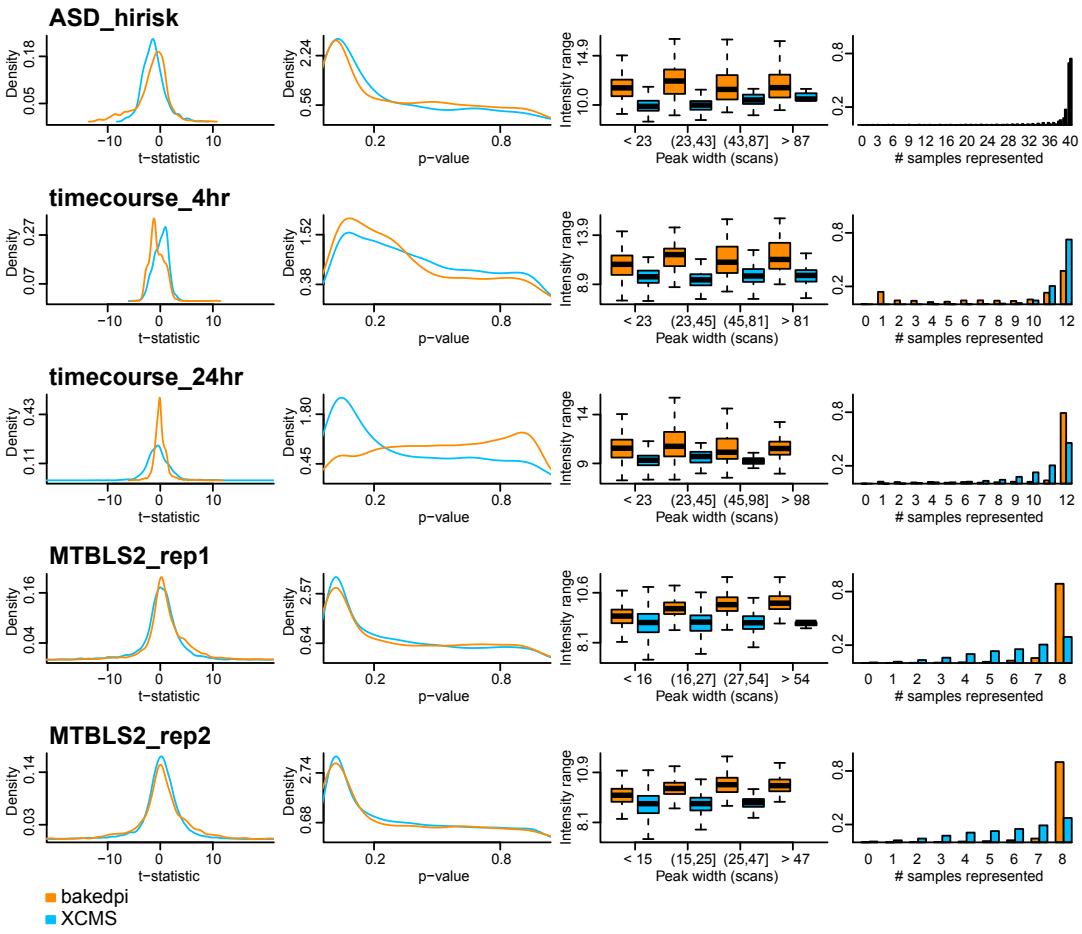
Supplementary Figure S6: Impact of RT alignment. (a) Percentage of peaks overlapping between bakedpi and MZmine2 for which quantification variability is higher in MZmine2 for various RT alignment strategies. **(b)** Like (a) but for XCMS.



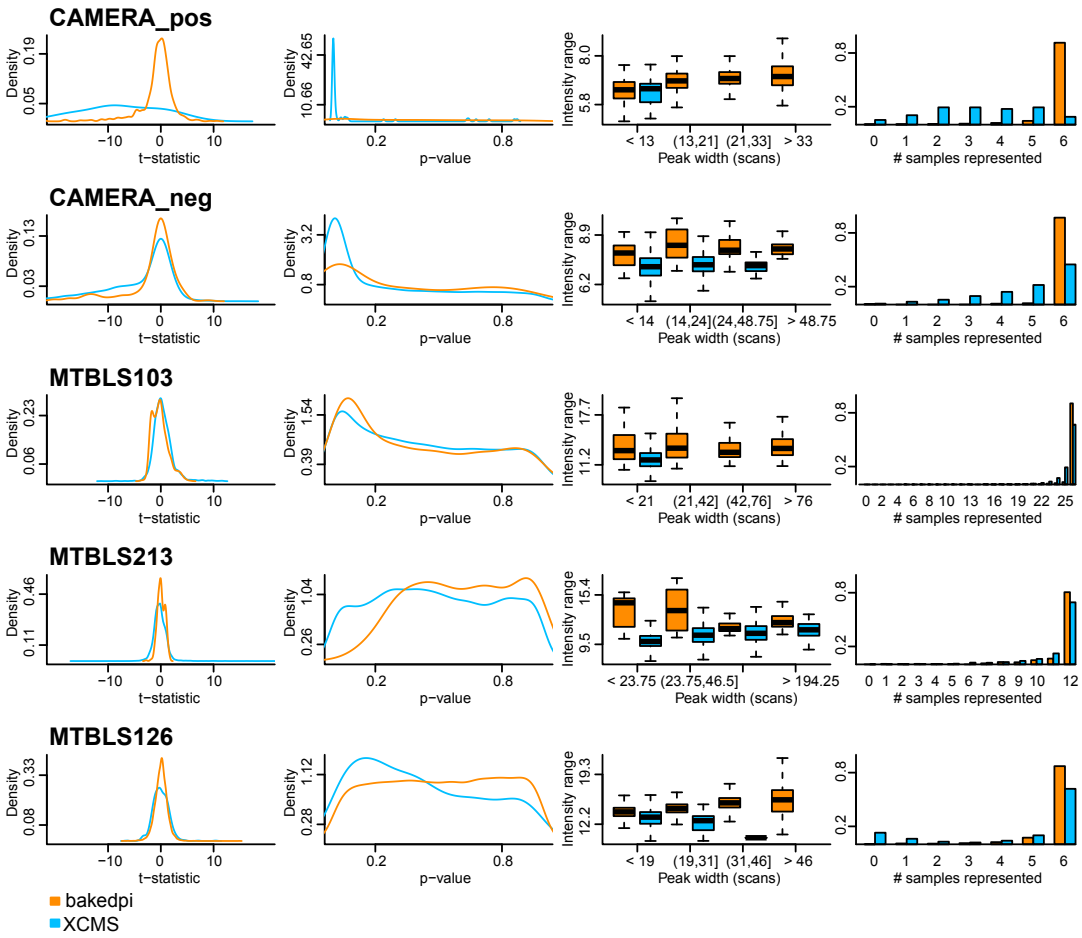
Supplementary Figure S7: Sensitivity of results to density cutoff. **(a)** Number of peaks detected by bakedpi as a function of the density cutoff. **(b)** The p-value distributions corresponding to the range of cutoffs. Shown in red is the cutoff actually picked by bakedpi. Shown in green and purple are slightly lower and slightly higher cutoffs.



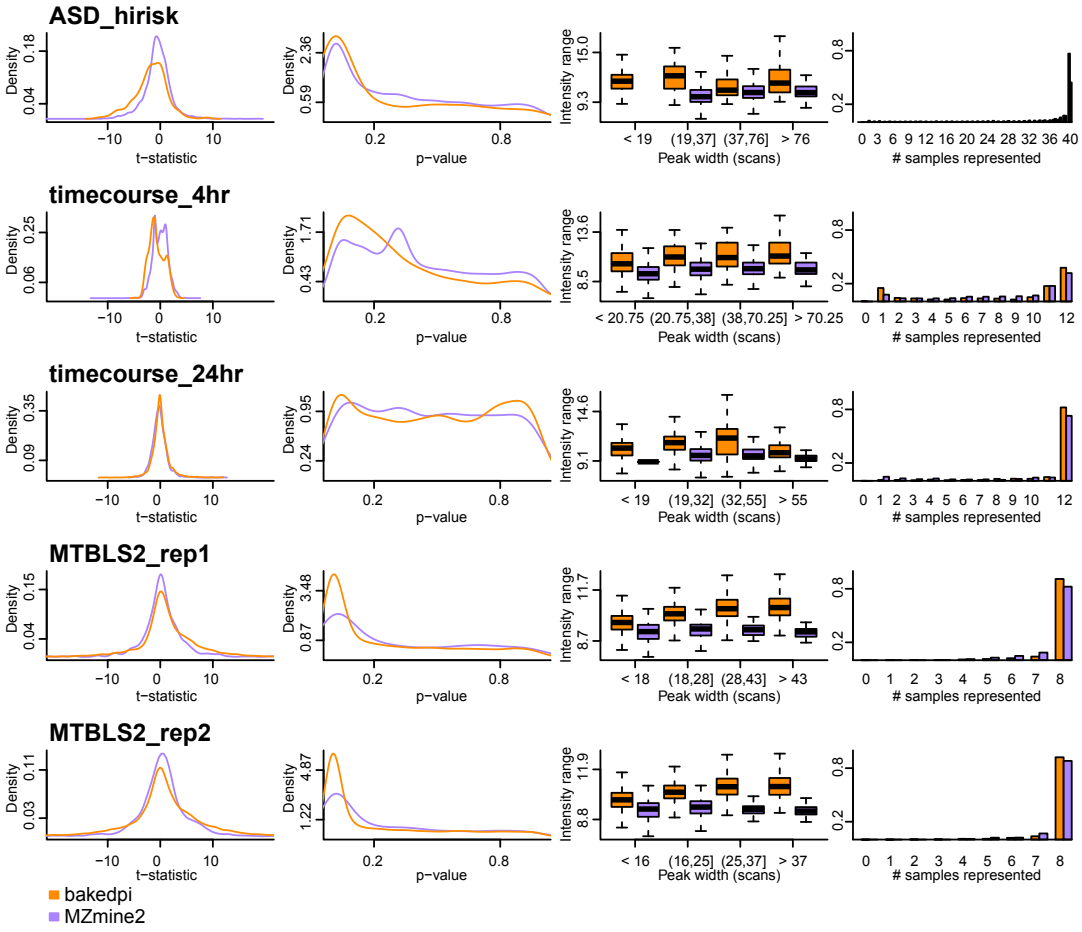
Supplementary Figure S8: Sensitivity of results to density cutoff. As Supplemental Figure S7, but for 5 additional datasets.



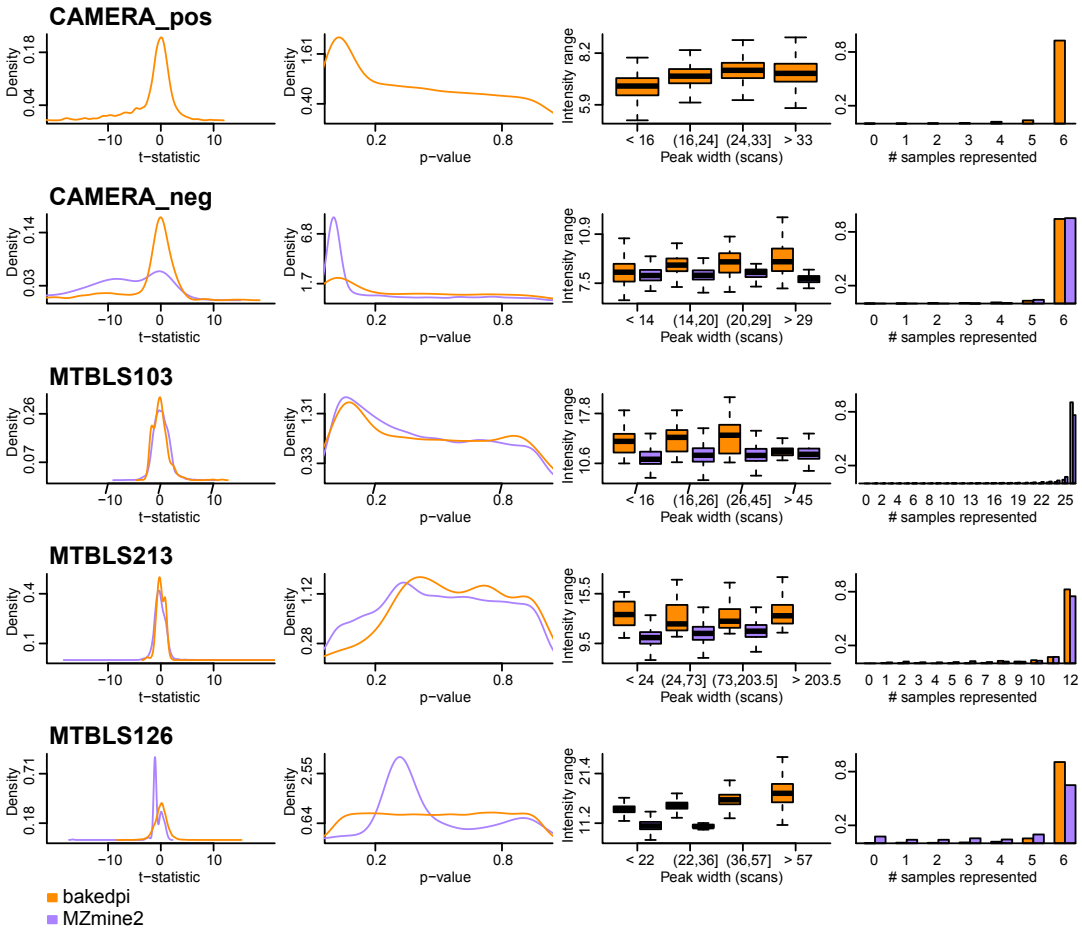
Supplementary Figure S9: Characteristics of peaks that are detected only by one method: bakedpi-XCMS comparison. Columns 1-4 show, respectively, the distribution of t-statistics, p-values, intensity ranges (log2), and number of samples represented for peaks detected only by bakedpi (orange) and only detected by XCMS (blue). The intensity range within a peak is a measure of peak height and is shown as a function of peak width.



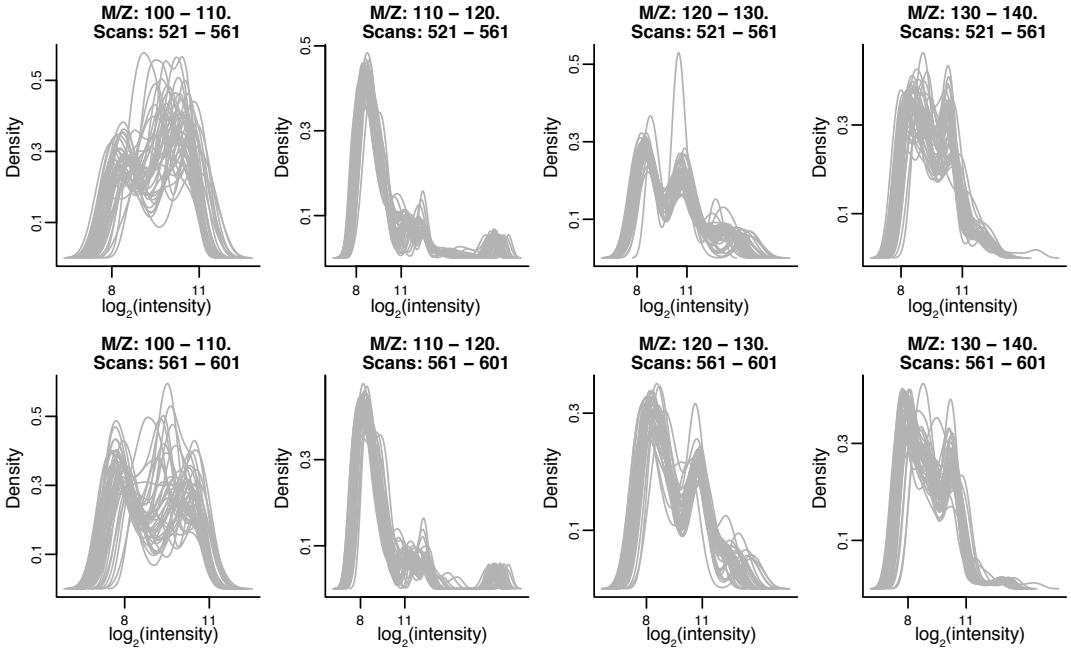
Supplementary Figure S10: Characteristics of peaks that are detected only by one method: bakedpi-XCMS comparison. As Supplementary Figure S9, but for 5 additional datasets.



Supplementary Figure S11: Characteristics of peaks that are detected only by one method: bakedpi-MZmine2 comparison. As with the bakedpi-XCMS comparisons (Supplementary Figures S9 and S10), with the first 5 datasets.



Supplementary Figure S12: Characteristics of peaks that are detected only by one method: bakedpi-MZmine2 comparison. As Supplementary Figure S11, but for 5 additional datasets.



Supplementary Figure S13: Region-specific intensity distributions. Each plot depicts the intensity distribution over a single grid region in the m/z-RT space, for the ASD_hirisk dataset. Each line corresponds to a single sample.

References

Bouhifd, Mounir, Thomas Hartung, Helena T Hogberg, Andre Kleensang, and Liang Zhao (2013). "Review: toxicometabolomics". In: *J. Appl. Toxicol.* 33.12, pp. 1365–1383. DOI: 10.1002/jat.2874.

Bouhifd, Mounir, Richard Beger, Thomas Flynn, Lining Guo, Georgina Harris, Helena Hogberg, Rima Kaddurah-Daouk, Hennicke Kamp, Andre Kleensang, Alexandra Maertens, Shelly Odwin-DaCosta, David Pamies, Donald Robertson, Lena Smirnova, Jinchun Sun, Liang Zhao, and Thomas Hartung (2015). "Quality assurance of metabolomics". In: *ALTEX* 32.4, pp. 319–326. DOI: 10.14573/altex.1509161.

Ramirez, Tzutzuy, Mardas Daneshian, Hennicke Kamp, Frederic Y Bois, Malcolm R Clench, Muireann Coen, Beth Donley, Steven M Fischer, Drew R Ekman, Eric Fabian, Claude Guillou, Joachim Heuer, Helena T Hogberg, Harald Jungnickel, Hector C Keun, Gerhard Krennrich, Eckart Krupp, Andreas Luch, Fozia Noor, Erik Peter, Bjoern Riefke, Mark Seymour, Nigel Skinner, Lena Smirnova, Elwin Verheij, Silvia Wagner, Thomas Hartung, Bennard van Ravenzwaay, and Marcel Leist (2013). "Metabolomics in toxicology and preclinical research". In: *ALTEX* 30.2, pp. 209–225.

Aberg, K Magnus, Erik Alm, and Ralf J O Torgrip (2009). "The correspondence problem for metabonomics datasets". In: *Anal. Bioanal. Chem.* 394.1, pp. 151–162. DOI: 10.1007/s00216-009-2628-9.

Hastings, Curtis A, Scott M Norton, and Sushmita Roy (2002). "New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data". In: *Rapid Commun. Mass Spectrom.* 16.5, pp. 462–467.

Vivó-Truyols, G, J R Torres-Lapasió, A M van Nederkassel, Y Vander Heyden, and D L Massart (2005). "Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals". In: *J. Chromatogr. A* 1096.1-2, pp. 133–145.

Du, Pan, Warren A Kibbe, and Simon M Lin (2006). "Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching". In: *Bioinformatics* 22.17, pp. 2059–2065.

Noy, Karin and Daniel Fasulo (2007). "Improved model-based, platform-independent feature extraction for mass spectrometry". In: *Bioinformatics* 23.19, pp. 2528–2535. DOI: 10.1093/bioinformatics/btm385.

Tautenhahn, Ralf, Christoph Böttcher, and Steffen Neumann (2008). "Highly sensitive feature detection for high resolution LC/MS". In: *BMC Bioinf.* 9, p. 504. DOI: 10.1186/1471-2105-9-504.

Chen, Shuo, Ming Li, Don Hong, Dean Billheimer, Huiming Li, Baogang J Xu, and Yu Shyr (2009). "A novel comprehensive wave-form MS data processing method". In: *Bioinformatics* 25.6, pp. 808–814. DOI: 10.1093/bioinformatics/btp060.

Nguyen, Nha, Heng Huang, Soontorn Oraintara, and An Vo (2010). "Mass spectrometry data processing using zero-crossing lines in multi-scale of Gaussian derivative wavelet". In: *Bioinformatics* 26.18, pp. i659–i665. DOI: 10.1093/bioinformatics/btq397.

Shalliker, R A, P G Stevenson, D Shock, M Mnatsakanyan, P K Dasgupta, and G Guiochon (2010). "Application of power functions to chromatographic data for the enhancement of signal to noise ratios and separation resolution". In: *J. Chromatogr. A* 1217.36, pp. 5693–5699. DOI: 10.1016/j.chroma.2010.07.007.

Vivó-Truyols, Gabriel (2012). "Bayesian approach for peak detection in two-dimensional chromatography". In: *Anal. Chem.* 84.6, pp. 2622–2630. DOI: 10.1021/ac202124t.

Fu, Hai-Yan, Jun-Wei Guo, Yong-Jie Yu, He-Dong Li, Hua-Peng Cui, Ping-Ping Liu, Bing Wang, Sheng Wang, and Peng Lu (2016). "A simple multi-scale Gaussian smoothing-based strategy for automatic chromatographic peak extraction". In: *J. Chromatogr. A* 1452, pp. 1–9. DOI: 10.1016/j.chroma.2016.05.018.

Tomasi, Giorgio, Frans van den Berg, and Claus Andersson (2004). "Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data". In: *J. Chemom.* 18.5, pp. 231–241. DOI: 10.1002/cem.859.

Podwojski, Katharina, Arno Fritsch, Daniel C Chamrad, Wolfgang Paul, Barbara Sitek, Kai Stühler, Petra Mutzel, Christian Stephan, Helmut E Meyer, Wolfgang Urfer, Katja Ickstadt, and Jörg Rahnenführer (2009). "Retention time alignment algorithms for LC/MS data must consider non-linear

shifts". In: *Bioinformatics* 25.6, pp. 758–764. DOI: 10.1093/bioinformatics/btp052.

Hoffmann, Nils, Matthias Keck, Heiko Neuweger, Mathias Wilhelm, Petra Högy, Karsten Niehaus, and Jens Stoye (2012). "Combining peak- and chromatogram-based retention time alignment algorithms for multiple chromatography-mass spectrometry datasets". In: *BMC Bioinf.* 13, p. 214. DOI: 10.1186/1471-2105-13-214.

Jeong, Jaesik, Xue Shi, Xiang Zhang, Seongho Kim, and Changyu Shen (2012). "Model-based peak alignment of metabolomic profiling from comprehensive two-dimensional gas chromatography mass spectrometry". In: *BMC Bioinf.* 13, p. 27. DOI: 10.1186/1471-2105-13-27.

Tekwe, Carmen D, Raymond J Carroll, and Alan R Dabney (2012). "Application of survival analysis methodology to the quantitative analysis of LC-MS proteomics data". In: *Bioinformatics* 28.15, pp. 1998–2003. DOI: 10.1093/bioinformatics/bts306.

Zhan, Xiang, Andrew D Patterson, and Debashis Ghosh (2015). "Kernel approaches for differential expression analysis of mass spectrometry-based metabolomics data". In: *BMC Bioinf.* 16, p. 77. DOI: 10.1186/s12859-015-0506-3.

Taylor, Sandra L, L Renee Ruhaak, Robert H Weiss, Karen Kelly, and Kyoungmi Kim (2017). "Multivariate two-part statistics for analysis of correlated mass spectrometry data from multiple biological specimens". In: *Bioinformatics* 33, pp. 17–25. DOI: 10.1093/bioinformatics/btw578.

Hastie, Trevor J, Robert John Tibshirani, and Jerome H Friedman (2011). *The elements of statistical learning: data mining, inference, and prediction*. New York, NY: Springer.

Smith, Colin A, Elizabeth J Want, Grace O'Maille, Ruben Abagyan, and Gary Siuzdak (2006). "XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification". In: *Anal. Chem.* 78.3, pp. 779–787. DOI: 10.1021/ac051437y.

Pluskal, Tomás, Sandra Castillo, Alejandro Villar-Briones, and Matej Oresic (2010). "MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data". In: *BMC Bioinf.* 11, p. 395. DOI: 10.1186/1471-2105-11-395.

Murakami, Itsuo, Romanas Chaleckis, Tomáš Pluskal, Ken Ito, Kousuke Hori, Masahiro Ebe, Mitsuhiro Yanagida, and Hiroshi Kondoh (2014). "Metabolism of skin-absorbed resveratrol into its glucuronized form in

mouse skin". In: *PLoS One* 9.12, e115359. DOI: 10.1371/journal.pone.0115359.

Libiseller, Gunnar, Michaela Dvorzak, Ulrike Kleb, Edgar Gander, Tobias Eisenberg, Frank Madeo, Steffen Neumann, Gert Trausinger, Frank Sinner, Thomas Pieber, and Christoph Magnes (2015). "IPO: a tool for automated optimization of XCMS parameters". In: *BMC Bioinf.* 16, p. 118. DOI: 10.1186/s12859-015-0562-8.

Smyth, Gordon K (2004). "Linear models and empirical bayes methods for assessing differential expression in microarray experiments". In: *Stat. Appl. Genet. Mol. Biol.* 3.1, pp. 1–26. DOI: 10.2202/1544-6115.1027.

Kammers, Kai, Robert N Cole, Calvin Tiengwe, and Ingo Ruczinski (2015). "Detecting Significant Changes in Protein Abundance". In: *EuPA Open Proteom* 7, pp. 11–19. DOI: 10.1016/j.euprot.2015.02.002.

Newschaffer, Craig J, Lisa A Croen, M Daniele Fallin, Irva Hertz-Picciotto, Danh V Nguyen, Nora L Lee, Carmen A Berry, Homayoon Farzadegan, H Nicole Hess, Rebecca J Landa, Susan E Levy, Maria L Massolo, Stacey C Meyerer, Sandra M Mohammed, Mckenzie C Oliver, Sally Ozonoff, Juhi Pandey, Adam Schroeder, and Kristine M Shedd-Wise (2012). "Infant siblings and the investigation of autism risk factors". In: *J. Neurodev. Disord.* 4.1, p. 7. DOI: 10.1186/1866-1955-4-7.

Kleensang, Andre, Marguerite M Vantangoli, Shelly Odwin-DaCosta, Melvin E Andersen, Kim Boekelheide, Mounir Bouhifd, Albert J Fornace Jr, Heng-Hong Li, Carolina B Livi, Samantha Madnick, Alexandra Maertens, Michael Rosenberg, James D Yager, Liang Zhaog, and Thomas Hartung (2016). "Genetic variability in a frozen batch of MCF-7 cells invisible in routine authentication affecting cell function". In: *Sci. Rep.* 6, p. 28994. DOI: 10.1038/srep28994.

Böttcher, Christoph, Lore Westphal, Constanze Schmotz, Elke Prade, Dierk Scheel, and Erich Glawischnig (2009). "The multifunctional enzyme CYP71B15 (PHYTOALEXIN DEFICIENT3) converts cysteine-indole-3-acetonitrile to camalexin in the indole-3-acetonitrile metabolic network of Arabidopsis thaliana". In: *Plant Cell* 21.6, pp. 1830–1845. DOI: 10.1105/tpc.109.066670.

Neumann, Steffen, Andrea Thum, and Christoph Böttcher (2012). "Nearline acquisition and processing of liquid chromatography-tandem mass spectrometry data". In: *Metabolomics* 9.1, pp. 84–91. DOI: 10.1007/s11306-012-0401-0.

Kuhl, Carsten, Ralf Tautenhahn, Christoph Böttcher, Tony R Larson, and Steffen Neumann (2012). "CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets". In: *Anal. Chem.* 84.1, pp. 283–289. DOI: 10.1021/ac202450g.

Samino, Sara, Maria Vinaixa, Marta Díaz, Antoni Beltran, Miguel A Rodríguez, Roger Mallol, Mercedes Heras, Anna Cabre, Lorena Garcia, Nuria Canela, Francis de Zegher, Xavier Correig, Lourdes Ibáñez, and Oscar Yanes (2015). "Metabolomics reveals impaired maturation of HDL particles in adolescents with hyperinsulinaemic androgen excess". In: *Sci. Rep.* 5, p. 11496. DOI: 10.1038/srep11496.

Capellades, Jordi, Miriam Navarro, Sara Samino, Marta Garcia-Ramirez, Cristina Hernandez, Rafael Simo, Maria Vinaixa, and Oscar Yanes (2016). "geoRge: A Computational Tool To Detect the Presence of Stable Isotope Labeling in LC/MS-Based Untargeted Metabolomics". In: *Anal. Chem.* 88.1, pp. 621–628. DOI: 10.1021/acs.analchem.5b03628.

Treviño, Victor, Irma-Luz Yañez-Garza, Carlos E Rodriguez-López, Rafael Urrea-López, Maria-Lourdes Garza-Rodriguez, Hugo-Alberto Barrera-Saldaña, José G Tamez-Peña, Robert Winkler, and Rocío-Isabel Díaz de-la Garza (2015). "GridMass: a fast two-dimensional feature detection method for LC/MS". en. In: *J. Mass Spectrom.* 50, pp. 165–174. DOI: 10.1002/jms.3512.

Duong, Tarn (2007). "ks: Kernel Density Estimation and Kernel Discriminant Analysis for Multivariate Data in R". In: *J. Stat. Softw.* 21.1, pp. 1–16. DOI: 10.18637/jss.v021.i07.

Wand, M P (1994). "Fast Computation of Multivariate Kernel Estimators". In: *J. Comput. Graph. Stat.* 3.4, pp. 433–445. DOI: 10.1080/10618600.1994.10474656.

Dowle, M, A Srinivasan, T Short, S Lianoglou with contributions from R Saporta, and E Antonyan (2015). *data.table: Extension of Data.frame*. URL: https://CRAN.R-project.org/package=data.table.

Pau, Grégoire, Florian Fuchs, Oleg Sklyar, Michael Boutros, and Wolfgang Huber (2010). "EBImage–an R package for image processing with applications to cellular phenotypes". In: *Bioinformatics* 26.7, pp. 979–981. DOI: 10.1093/bioinformatics/btq046.

Chapter 3

Massively parallel reporter assays

3.1 Introduction

Noncoding regions in the human genome represent the overwhelming majority of genomic sequence, but their function remains largely uncharacterized. Better understanding of the functional consequences of these regions has the potential to greatly enrich our understanding of biology. It is well understood that some noncoding regions are regulatory in nature. It has been straightforward to experimentally test the regulatory ability of a given DNA sequence with standard reporter assays, but these assays are low throughout and do not scale to the testing of large numbers of sequences. Massively parallel reporter assays (MPRA) have emerged as a high-throughput means of measuring the ability of sequences to drive expression (White, 2015; Melnikov et al., 2014). These assays build on the traditional reporter assay framework by coupling each putative regulatory sequence with several short DNA tags, or barcodes, that are incorporated into the RNA output. These tags are counted in the RNA reads and the input DNA, and the resulting counts are used to quantify

the activity of a given putative regulatory sequence, typically involving the ratio of RNA counts to DNA counts (Figure 3.1). The applications of MPRA have been diverse, and there have been correspondingly diverse and ad hoc methods used in statistical analysis. There are three broad categories of MPRA applications: characterization studies, saturation mutagenesis, and differential analysis.

Characterization studies examine thousands of different putative regulatory elements that have a wide variety of sequence features and try to correlate these sequence features with measured activity levels (Grossman et al., 2017; Guo et al., 2017; Safra et al., 2017; Levo et al., 2017; Maricque, Dougherty, and Cohen, 2017; Groff et al., 2016; Ernst et al., 2016; White et al., 2016; Ferreira et al., 2016; Fiore and Cohen, 2016; Farley et al., 2015; Kamps-Hughes et al., 2015; Dickel et al., 2014; Kwasnieski et al., 2014; Mogno, Kwasnieski, and Cohen, 2013; Gisselbrecht et al., 2013; White et al., 2013; Smith et al., 2013). Typical statistical analyses use regression to study the impact of multiple features simultaneously. They also compare continuous activity measures or categorized (high/low) activity measures across groups using paired and unpaired t-, rank, Fisher's exact, and chi-squared tests.

Saturation mutagenesis studies look at only a few established enhancers and examine the impact on activity of every possible mutation at each base as well as interactions between these mutations (Patwardhan et al., 2009; Melnikov et al., 2012; Patwardhan et al., 2012; Kwasnieski et al., 2012; Kheradpour et al., 2013; Birnbaum et al., 2014; Zhao et al., 2014). Analyses have uniformly used linear regression where each position in the enhancer sequence is a

predictor.

Differential analysis studies look at thousands of different elements, each of which has two or more versions. Versions can correspond to allelic versions of a sequence (Ulirsch et al., 2016; Tewhey et al., 2016; Vockley et al., 2015) or different environmental contexts (Inoue et al., 2017), such as different cell or tissue types (Shen et al., 2016). These studies have compared different sequence versions using paired t-tests, rank sum tests, and Fisher's exact test (by pooling counts over biological replicates).

Despite the increasing popularity of this assay, guiding principles for statistical analysis have not been put forth. Researchers use a large variety of ad hoc methods for analysis. For example, there has been considerable diversity in the earlier stages of summarization of information over barcodes. Barcodes are viewed as technical replicates of the regulatory element sequences, and groups have considered numerous methods for summarizing barcode-level information into one activity measure per enhancer. On top of this, a large variety of statistical tests are used to make comparisons.

Recently, a method called QuASAR-MPRA was developed to identify regulatory sequences that have allele-specific activity (Kalita et al., 2017). This method uses a beta-binomial model to model RNA counts as a function of DNA counts, and it provides a means for identifying sequences that show a significant difference in regulatory activity between two alleles. While it provides a framework for two group differential analysis within MPRAs, QuASAR-MPRA is limited in this regard because experiments might have several conditions and involve arbitrary comparisons.

To our knowledge, no method has been developed that provides tools for general purpose differential analysis of activity measures from MPRA. General purpose methods are ones that can flexibly analyze data from a range of study designs. We present mpralm, a method for testing for differential activity in MPRA experiments. Our method uses linear models as opposed to count-based models to identify differential activity. This approach provides desired analytic flexibility for more complicated experimental designs that necessitate more complex models. It also builds on an established method that has a solid theoretical and computational framework (Law et al., 2014). We show that mpralm can be applied to a wide variety of MPRA datasets and has good statistical properties related to type I error control and power. Furthermore, we examine proper techniques for combining information over barcodes and provide guidelines for choosing sample sizes and sequencing depth when considering power. Our method is open source and freely available in the *mpra* package for R on the Bioconductor repository: https://bioconductor.org/packages/mpra.

3.2 Results

3.2.1 The structure of MPRA data and experiments

MPRA data consists of measuring the activity of some putative regulatory sequences, henceforth referred to as "elements". First a plasmid library of oligos is constructed, where each element is coupled with a number of short DNA tags, or barcodes. This plasmid library is then transfected into one or more cellular contexts, either as free-floating plasmids or integrated into

the genome (Inoue et al., 2017). Next, RNA output is measured using RNA sequencing, and DNA output as a proxy for element copy number is measured using DNA sequencing (occasionally, element copy number is unmeasured), giving the data structure shown in Figure 3.1. The log-ratio of RNA to DNA counts is commonly used as an activity outcome measure.

Since each element is measured across a number of barcodes, it is useful to summarize this data into a single activity measure a for a single element in a single sample. Multiple approaches have been proposed for this summarization step. We consider two approaches. First is averaging, where a log-ratio is computed for each barcode, then averaged across barcodes. This treats the different barcodes as technical replicates. The second approach is aggregation, where RNA and DNA counts are each summed across barcodes, followed by formation of a log-ratio. This approach effectively uses the barcodes to simply increase the sequencing counts for that element.

In our investigation of the characteristics of MPRA data we use a number of datasets listed in Table 3.1. We have divided them into 3 categories. Two categories are focused on differential analysis: one on comparing different alleles and one on comparing the same element in different conditions (retina vs. cortex and episomal vs. chromosomal integration). The two allelic studies naturally involve paired comparisons in that the two elements being compared are always measured together in a single sample (which is replicated). We also use two saturation mutagenesis experiments.

3.2.2 The variability of MPRA data depends on element copy number

It is well established that count data from RNA sequencing studies exhibit a mean-variance relationship (McCarthy, Chen, and Smyth, 2012). On the log scale, low counts are more variable across replicates than high counts, at least partly due to inherent Poisson variation in the sequencing process (Marioni et al., 2008; Bullard et al., 2010). This relationship has been leveraged in both count-based analysis methods (Robinson, McCarthy, and Smyth, 2010; Love, Huber, and Anders, 2014) and, more recently, linear model-based methods (Law et al., 2014). For count-based methods, this mean-variance relationship helps improve dispersion estimates, and for linear model-based methods, the relationship allows for estimation of weights reflecting inherent differences in variability for count observations in different samples and genes.

Because MPRAs are fundamentally sequencing assays, it is useful to know whether similar variance relationships hold in these experiments. Due to the construction of MPRA libraries, each element is present in a different (random) copy number, and this copy number ought to impact both background and signal measurements from the element. We are therefore specifically interested in the functional relationship between element copy number and the variability of the activity outcome measure. As outcome measure we use the log-ratio of RNA counts to DNA counts (aggregate estimator), and we use aggregated DNA counts, averaged across samples, as an estimate of DNA copy number. We compute empirical standard deviations of the library size-corrected outcome measure across samples. In Figure 3.2 we depict this relationship

across the previously discussed publicly available datasets (Table 3.1). For all datasets, with one exception, there is higher variation associated with lower copy number. The functional form is reminiscent of the mean-variance relationship in RNA sequencing data (Law et al., 2014), despite that we here show variance of a log-ratio of sequencing counts.

3.2.3 Statistical modeling of MPRA data

To model MPRA data we propose to use a simple variant of the voom methodology (Law et al., 2014), proposed for analysis of RNA sequencing data. This methodology is based on standard linear models, which are coupled with inverse variance weights representing the mean-variance relationship inherent in RNA sequencing data. The weights are derived from smoothing an empirical mean-variance plot. Similar to voom, we propose to use linear models to model log-ratio activity data from MPRAs, but we estimate weights by smoothing the relationship between empirical variance of the log-ratios and log-DNA copy number, as depicted in Figure 3.2. This approach has a number of advantages. (1) It is flexible to different functional forms of the variance-copy number relationship. (2) It allows for a unified approach to modeling many different types of MPRA design using the power of design matrices. (3) It allows for borrowing of information across elements using empirical Bayes techniques. (4) It allows for different levels of correlation between elements using random effects. We call this approach mpralm.

The current literature on analysis of MPRA experiments contains many variant methods (see Introduction). To evaluate mpralm, we compare the

method to the following variants used in the literature: QuASAR-MPRA, t-tests, and Fisher's exact test. QuASAR-MPRA is a recently developed method that is targeted for the differential analysis of MPRA data (Kalita et al., 2017). It specifically addresses a two group differential analysis where the two groups are elements with two alleles and uses base-calling error rate in the model formulation. It collapses count information across samples to create three pieces of information for each element: one count for RNA reads for the reference allele, one count for RNA reads for the alternate allele, and one proportion that gives the fraction of DNA reads corresponding to the reference allele. Fisher's exact test similarly collapses count information across samples. To test for differential activity, a 2-by-G table is formed with RNA and DNA designation forming one dimension and condition designation (with G groups) in the second dimension. The t-test operates on the log ratio outcomes directly; we use the aggregate estimator to summarize over barcodes. Either a paired or unpaired t-test is used based on experimental design.

Both edgeR and DESeq2 are popular methods for analysis of RNA-sequencing data represented as counts. The two methods are both built on negative binomial models, and both attempt to borrow information across genes. These methods allow for the inclusion of an offset. Because both methods use a logarithmic link function, including log-DNA as an offset allows for the modeling of log-ratios of RNA to DNA. This makes these methods readily applicable to the analysis of MPRA data, and they carry many of the same advantages as mpralm. In addition to QuASAR, t-tests, and Fisher's exact test, we examine the performance of edgeR and DESeq2 for differential activity

analysis in our evaluations.

3.2.4 Simulations shed light on permutation strategies for assessing error rates

Because comparison of type I error rates forms an important part of our methods evaluation (next section), we first present simulation study results regarding the accuracy of permutation procedures for estimating type I error rates. These procedures consist of creating curated null permutations in which the comparison groups are composed of half of the samples from the two original groups. The error rate at different nominal levels is estimated with the median error rate over permutations.

Figure 3.3 shows how permutation-estimated error rates compare to true type I error rates in a simulation setting with increasing prevalence of differential activity. For all methods we show error estimates resulting from permuting the raw data. For mpralm and the t-test, which operate on the continuous log-ratios, we explore the permutation of residuals proposed in Jiang (2017). We uniformly see that permuting residuals results in substantial overestimation of the error for both methods. Permuting the raw data results in accurate estimation of the error rates in most situations. For this reason, we choose to estimate error rates in real datasets (Table 3.1) with raw data permutations. We note, however, that permutation of the raw data consistently results in overestimation of QuASAR's error rates and underestimation of error rates for mpralm for 30% and 50% differential activity. The degree of over- and underestimation increases with the proportion (p) of differential elements, with the effect being more dramatic for QuASAR than for mpralm. We draw

on these results when comparing method performance on real datasets in the next section.

3.2.5 mpralm is a powerful method for differential analysis

First, we focus on evaluating the performance of mpralm for differential analysis. We compare to QuASAR-MPRA, t-tests, Fisher's exact test, edgeR, and DESeq2. We use four of the previously discussed studies, specifically the Tewhey, Inoue, Ulirsch, and Shen studies. Two of these studies (Tewhey, Ulirsch) focus on comparing the activity of elements with two alleles, whereas the other two (Inoue, Shen) compare the activity of each element in two different conditions. For the allelic studies, we use a random effects model for mpralm and paired t-tests. Both Tewhey et al. (2016) and Ulirsch et al. (2016) compare alleles in different cellular contexts; we observe similar behavior of all evaluations in all contexts (data not shown) and have therefore chosen to depict results from one cellular context for both of these studies. For Tewhey et al. (2016) we depict results both from a large pool of elements used for initial screening and a smaller, targeted pool.

Figure 3.4 shows p-value distributions that result from running all methods. Across these datasets, all methods except for QuASAR show a well-behaved p-value distribution; high p-values appear uniformly distributed, and there is a peak at low p-values. QuASAR-MPRA consistently shows conservative p-value distributions. We were unable to run QuASAR-MPRA for the Shen dataset. Fisher's exact test has a very high peak around zero, likely due to the extreme sensitivity of the test with high counts. We examine mpralm using

both an average estimator and an aggregation estimator for summarizing across barcodes; this cannot be done for the Tewhey dataset where we do not have access to barcode-level data. To fully interpret these p-value distributions, we need to assess error rates.

To estimate empirical type I error rates, we performed null permutations as described in the previous section. Figure 3.5 shows estimated error rates (median error rate over the permutations). We observe that Fisher's exact test has wildly inflated type I error, presumably because the data is overdispersed. QuASAR-MPRA appears well calibrated across datasets, but these error rates might be overestimated. mpralm, t-tests, edgeR, and DESeq2 control the type I error rate but tend to be conservative.

To investigate the trade-off between observed power (number of rejected tests) and type I error rates, we combine these quantities in two ways: (1) we look at the number of rejections as a function of observed type I error rates and (2) we look at estimated FDR as a function of the number of rejections.

In Figure 3.6 we display the number of rejections as a function of observed type I error rates. In this display, we have essentially used the observed type I error rate displayed in Figure 3.5 to calibrate the nominal alpha-level. For a fixed error rate, we interpret a high number of rejections to suggest high power. Both Fisher's exact test and QuASAR-MPRA show poor performance. Because our simulations suggest that the type I error rate of QuASAR can be overestimated with permutations, we expect that it should have better performance than depicted. However, given that its largest number of detections (Figure 3.6 bottom row) is nearly always as low as the smallest number of

detections from other methods, we expect that QuASAR still has poor performance in this regard. Across these datasets, mpralm tends to have the best performance, but edgeR and DESeq2 are competitive. Because our simulations suggest an underestimation of the type I error rate for mpralm, we expect these methods to be closely comparable for this metric.

If we know the proportion of true null hypotheses, π_0, we can estimate false discovery rates (FDR). This proportion is an unknown quantity, but we estimate it using a method developed by Phipson (2013) and thereby compute an estimated FDR. In Figure 3.7 the estimated FDR (for a given π_0) is displayed as a function of the number of rejections. QuASAR-MPRA, t-tests, and Fisher's exact test tend to have the highest false discovery rates. mpralm tends to have the lowest FDRs. For the Inoue dataset, all methods except for QuASAR have very low FDR, presumably because a very high fraction of elements are expected to be differential given the extreme expected differences between the comparison groups. For this metric, we again expect that QuASAR has better performance than depicted due to error rate overestimation but not enough to be comparable to the other methods. We also expect mpralm to be more comparable to edgeR and DESeq2 given its error rate underestimation.

In conclusion, we observe that Fisher's exact test has too high of an error rate and that QuASAR-MPRA is underpowered; based on these results we cannot recommend either method. T-tests perform better than these two methods but are still outperformed by mpralm, edgeR, and DESeq2, which all have similar performance.

3.2.6 Comparison of element rankings between methods

While power and error calibration are important evaluation metrics for a differential analysis method, they do not have a direct relation with element rankings, which is often of practical importance. We observe fairly different rankings between mpralm and the t-test and examine drivers of these differences in Figure 3.8. For each dataset, we find the MPRA elements that appear in the top 200 elements with one method but not the other. We will call these uniquely top ranking elements, and they make up 24% to 64% of the top 200 depending on dataset. For most datasets, DNA, RNA, and log-ratio activity measures are higher in uniquely top ranking mpralm elements (top three rows of Figure 3.8). It is desirable for top ranking elements to have higher values for all three quantities because higher DNA levels increase confidence in the activity measure estimation, and higher RNA and log-ratio values give a stronger indication that a particular MPRA element has regulatory activity. In the last two rows of Figure 3.8, we compare effect sizes and variability measures (residual standard deviations). The t-test uniformly shows lower variability but also lower effect sizes for its uniquely top ranking elements. This follows experience from gene-expression studies where standard t-tests tend to underestimate the variance and thereby exhibit t-statistics which are too large, leading to false positives. In MPRA studies, as with most other high-throughput studies, it is typically more useful to have elements with high effect sizes at the top of the list. Such elements are able to picked out in mpralm due to its information sharing and weighting framework.

We similarly compare mpralm rankings with edgeR and DESeq2 rankings

in Figures 3.9 amd 3.10. The ranking concordance between mpralm and these two methods is much higher than with the t-test. Generally, uniquely top ranking mpralm elements have higher DNA and RNA levels, but lower log-ratio activity measures. Uniquely top ranking mpralm elements also tend to have larger effect sizes. The variability of activity measures (residual SD) is similar among the methods.

3.2.7 mpralm enables modeling for complex comparisons

While many comparisons of interest in MPRA studies can be posed as a two group comparison (e.g. major allele vs. minor allele), more complicated experimental designs are also of interest. For example, in the allelic study conducted by Ulirsch et al. (2016), putative biallelic enhancer sequences are compared in two cellular contexts. The first is a standard culture of K562 cells, and the second is a K562 culture that induces over-expression of GATA1 for a more terminally-differentiated phenotype. A straightforward question is whether an allele's effect on enhancer activity differs between cellular contexts. Let y_{eia} be the enhancer activity measure (log ratio of RNA over DNA counts) for element e, in sample i for allele a. Let x_{1eia} be a binary indicator of the mutant allele. Let x_{2eia} be a binary indicator of the GATA1 over-expression condition. Then the following model

$$Y_{eia} = \beta_{0e} + \beta_{1e}x_{1eia} + \beta_{2e}x_{2eia} + \beta_{3e}x_{1eia}x_{2eia} + b_i + \epsilon_{eia}$$

is a linear mixed effects model for activity measures, where b_i is a random effect that induces correlation between the two alleles measured within the

same sample. We can perform inference on the β_{3e} parameters to determine differential allelic effects. Such a model is easy to fit within the mpralm framework, since our framework supports model specifications by general design matrices. In contrast, this question cannot be formulated in the QuASAR, t-test, and Fisher's exact test frameworks. Neither edgeR nor DESeq2 support the fitting of mixed effects models.

3.2.8 Accuracy of activity measures and power of differential analysis depends on summarization technique over barcodes

MPRA data initially contain count information at the barcode level, but we typically desire information summarized at the element level for the analysis stage. We examine the theoretical properties of two summarization methods: averaging and aggregation. Under the assumption that DNA and RNA counts follow a count distribution with a mean-variance relationship, we first show that averaging results in activity estimates with more bias. Second, we examine real data performance of these summarization techniques.

Let R_b and D_b denote the RNA and DNA count, respectively, for barcode $b = 1, \ldots, B$ for a putative regulatory element in a given sample. We suppress the dependency of these counts on sample and element. Typically, B is approximately 10 to 15 (for examples, see Table 3.1). We assume that R_b has mean μ_r and variance $k_r\mu_r$ and that D_b has mean μ_d and variance $k_d\mu_d$. Typically the constants k_d and k_r are greater than 1, modeling overdispersion. Negative binomial models are a particular case with $k = 1 + \phi\mu$, where ϕ is an overdispersion parameter. Also let N_d and N_r indicate the library size for

DNA and RNA, respectively, in a given sample. Let p_d and p_r indicate the fraction of reads mapping to element e for DNA and RNA, respectively, in a given sample so that $\mu_r = N_r p_r$ and $\mu_d = N_d p_d$. Let a be the true activity measure for element e defined as $a := \log(p_r / p_d)$. When performing total count normalization, the RNA and DNA counts are typically scaled to a common library size L.

The average estimator of a is an average of barcode-specific log activity measures:

$$\hat{a}^{AV} = \frac{1}{B} \sum_{b=1}^{B} \log \left(\frac{R_b L / N_r + 1}{D_b L / N_d + 1} \right)$$

Using a second order Taylor expansion (Methods), it can be shown that this estimator has bias approximately equal to

$$\text{bias}^{AV} \approx \frac{1}{2} \left(\frac{k_d}{\mu_d} - \frac{k_r}{\mu_r} \right) = \frac{1}{2} \left(\frac{k_d}{N_d p_d} - \frac{k_r}{N_r p_r} \right)$$

The aggregate estimator of a first aggregates counts over barcodes:

$$\hat{a}^{AGG} = \log \left(\frac{1 + (L/N_r) \sum_{b=1}^{B} R_b}{1 + (L/N_d) \sum_{b=1}^{B} D_b} \right)$$

Using an analogous Taylor series argument, we can show that this estimator has bias approximately equal to

$$\text{bias}^{AGG} \approx \frac{1}{B} \text{bias}^{AV}$$

The aggregate estimator has considerably less bias than the average estimator for most MPRA experiments because most experiments use at least 10 barcodes per element. Bias magnitude depends on count levels and the true activity measure a. Further, the direction of bias depends on the relative variability of RNA and DNA counts. Similar Taylor series arguments show that the variance of the two estimators is approximately the same.

The choice of estimator can impact the estimated log fold-changes (changes in activity) in a differential analysis. In Figure 3.11 we compare the log fold-changes inferred using the two different estimators. For the Inoue dataset, these effect sizes are very similar, but there are larger differences for the Ulirsch and Shen datasets.

Aggregation technique affects power in a differential analysis. In the last three columns of Figures 3.4, 3.5, 3.6, and 3.7, we compare aggregation to averaging using mpralm. The two estimators have similar type I error rates but very different detection rates between datasets. The average estimator is more favorable for the Ulirsch and Shen datasets, and the aggregate estimator is more favorable in the Inoue dataset.

3.2.9 Recommendations for sequencing depth and sample size

To aid in the design of future MPRA experiments, we used the above mathematical model to inform power calculations. Power curves are displayed in Figure 3.12. We observe that the variance of the aggregate estimator depends minimally on the true unknown activity measure but is greatly impacted by

sequencing depth. We fix one of the two true activity measures to be 0.8 as this is common in many datasets. We use a nominal type I error rate of 0.05 that has been Bonferroni adjusted for 5000 tests to obtain conservative power estimates. We also use ten barcodes per element as this is typical of many studies.

Our model suggests different impacts of sample size, and a marked impact of increasing the number of replicates, especially between 2 and 6 samples. From Figure 3.13, we can see that large effect sizes (effect sizes of 1 or greater) are typical for top ranking elements in many MPRA studies. In this situation it is advisable to do 4 or more replicates per group.

3.3 Discussion

The field of MPRA data analysis has been fragmented and consists of a large collection of study-specific ad hoc methods. Our objective in this work has been to provide a unified framework for the analysis of MPRA data. Our contributions can be divided into three areas. First, we have investigated techniques for summarizing information over barcodes. In the literature, these choices have always been made without justification and have varied considerably between studies. Second, we have developed a linear model framework, mpralm, for powerful and flexible differential analysis. To our knowledge, this is the second manuscript evaluating for statistical analysis in MPRA studies. The first proposed the QuASAR-MPRA method (Kalita et al., 2017), which we show to have worse performance than mpralm. In our comparisons, we provide the largest and most comprehensive comparison of

analysis methods so far; earlier work used only two datasets for comparisons. Third, we have analyzed the impact of sequencing depth and number of replicates on power. To our knowledge, this is the first mathematically-based power investigation, and we expect this information to be useful in the design of MPRA studies.

The activity of a regulatory element can be quantified with the log ratio of RNA counts to DNA counts. In the literature, groups have generally taken two approaches to summarizing barcode information to obtain one such activity measure per element per sample. One approach is to add RNA and DNA counts from all barcodes to effectively increase sequencing depth for an element. This is termed the aggregate estimator. Another approach is to compute the log ratio measure for each barcode and use an average of these measures as the activity score for an element. This is termed the average estimator, and we have shown that it is more biased than the aggregate estimator. Because of this bias, we caution against the use of the average estimator when comparing activity scores in enhancer groups (often defined by sequence features). However, it is unclear which of the two estimators is more appropriate for differential analysis.

In addition to barcode summarization recommendations, we have proposed a linear model framework, mpralm, for the differential analysis of MPRA data. Our evaluations show that it produces calibrated p-values and is as or more powerful than existing methods being used in the literature. Its type I error rates appear conservative, so in practice, we recommend performing permutations to estimate error rates.

While the count-based tools, edgeR and DESeq2, would seem like natural methods to use for the analysis of MPRA data, they have not been used for differential analysis of MPRA activity measures. There has been some use of DESeq2 to identify (filter) elements with regulatory activity (differential expression of RNA relative to DNA) (Tewhey et al., 2016; Gisselbrecht et al., 2013). However, these tools have not been used for comparisons of activity measures between groups. In this work we propose the use of log-DNA offsets as potential sensible uses of these software for differential analysis. In our evaluations, we see that this approach is most competitive with mpralm. For the allelic studies (Tewhey et al., 2016; Ulirsch et al., 2016), we observe that the degree of within-sample correlation affects the power of mpralm relative to comparison methods. In particular, there is little difference in the performance of the different methods for the large pool experiment of Tewhey et al. (2016), and this experiment had overall low within-sample correlation. Both the targeted pool experiment of Tewhey et al. (2016) and the Ulirsch experiment had larger within-sample correlations, and we observe that mpralm has increased power over the comparison methods for these datasets. We expect that mpralm will generally be more powerful for paired designs with high within-pair correlations.

In terms of element rankings, mpralm, edgeR, and DESeq2 are similar. However, we observe a substantial difference in ranking between t-tests and mpralm and believe top ranked mpralm elements exhibit better properties compared to those from t-tests.

Linear models come with analytic flexibility that is necessary to handle diverse MPRA designs. Paired designs involving alleles, for example, are easily handled with linear mixed effects models due to computational tractability. The studies we have analyzed here only consider two alleles per locus. It is possible to have more than two alleles at a locus, and such a situation cannot be addressed with paired t-tests, but is easily analyzed using mpralm. This is important because we believe such studies will eventually become routine for understanding results from genome-wide association studies.

While we have focused on characterizing the mpralm linear model framework for differential analysis, it is possible to include variance weights in the multivariate models used in saturation mutagenesis and characterization studies. We expect that modeling the copy number-variance relationship will improve the performance of these models.

For power, we find a substantial impact of even small increases in sample size. This is an important observation because many MPRA studies use 2 or 3 replicates per group, and our results suggest that power can be substantially increased with even a modest increase in sample size. We caution that using less than 4 replicates can be quite underpowered.

In short, the tools and ideas set forth here will aid in making rigorous conclusions from a large variety of future MPRA studies.

3.4 Methods

3.4.1 Data

See Table 1. Dataset labels used in figures are accompanied by short descriptions below.

Melnikov: Study of the base-level impact of mutations in two inducible enhancers in humans (Melnikov et al., 2012): a synthetic cAMP-regulated enhancer (CRE) and a virus-inducible interferon-beta enhancer (IFNB). We do not look at the IFNB data because it contains only one sample. We consider 3 datasets:

Melnikov: CRE, single-hit, induced state: Synthetic cAMP-regulated enhancer, single-hit scanning, induced state.

Melnikov: CRE, multi-hit, uninduced state: Synthetic cAMP-regulated enhancer, multi-hit sampling, uninduced state.

Melnikov: CRE, multi-hit, induced state: Synthetic cAMP-regulated enhancer, multi-hit sampling, induced state.

Kheradpour: Study of the base-level impact of mutations in various motifs (Kheradpour et al., 2013). Transfection into HepG2 and K562 cells.

Tewhey: Study of allelic effects in eQTLs (Tewhey et al., 2016). Transfection into two lymphoblastoid cell lines (NA12878 and NA19239) as well as HepG2. In addition two pools of plasmids are considered: a large screening pool and a smaller, targeted pool, designed based on the results of the large pool. We use data from both the large and the targeted pool in NA12878.

Inoue: chromosomal vs. episomal: Comparison of episomal and

chromosomally-integrated constructs (Inoue et al., 2017). This study uses a wild-type and mutant integrase to study the activity of a fixed set of putative regulatory elements in an episomal and a chromosomally-integrated setting, respectively.

Ulirsch: Study of allelic effects in GWAS to understand red blood cell traits (Ulirsch et al., 2016). Transfection into K562 cells as well as K562 with GATA1 overexpressed. We use the data from K562.

Shen: mouse retina vs. cortex: Comparison of cis-regulatory elements in-vivo in mouse retina and cerebral cortex (Shen et al., 2016). Candidate CREs that tile targeted regions are assayed in-vivo in these two mouse tissues with adeno-associated virus delivery.

3.4.2 Count preprocessing

We use total count normalization to account for differences in library size for both DNA and RNA. Specifically, each count in a sample is divided by that sample's library size and scaled so that the library size in all samples is the same. We perform minimal filtering on the counts to remove elements from the analysis that have low counts across all samples. Specifically, we require that DNA counts must be at least 10 in all samples to avoid instability of the log-ratio activity measures. We also remove elements in which these log-ratios are identical across all samples. This is necessary for sensible differential analysis. In practice, log-ratios are only identical across all samples if RNA counts are zero across all samples. Both steps also improve the estimation of the copy number-variance relationship used in subsequent modeling by

removing clear outliers.

3.4.3 Estimating the copy number-variance relationship

After preprocessing the first step is to estimate the copy number-variance relationship that will allow for the estimation of element-specific reliability weights. These weights are ultimately used in element-specific weighted regressions. The square root of the standard deviation of the log-ratios over samples are taken as a function of average log DNA levels over samples, and this relationship is fit with a lowess curve. Predicted variances are inverted to form observation-level precision weights.

3.4.4 Modeling

Once the observation-specific weights are calculated, the log-ratios and weights are used in the voom analysis pipeline. If, as in allele-specific activity studies, the different versions of the elements being compared are correlated due to being measured in the same sample, a mixed model is fit for each element using the `duplicateCorrelation` module in the limma Bioconductor package (Smyth, Michaud, and Scott, 2005).

3.4.5 Running mpralm, QuASAR, t-test, Fisher's exact test

For all methods, DNA and RNA counts were first corrected for library size with total count normalization. For edgeR and DESeq2, DNA counts were included as offset terms on the log scale before standard analysis. For the t-test we computed the aggregate estimator of the log-ratio as the outcome measure.

For Fisher's exact test, we summed DNA and RNA counts in the two conditions to form a 2-by-2 table as input to the procedure. For QuASAR-MPRA, we summed RNA counts in each condition to get one reference condition count and one alternative condition count per element. We also summed DNA counts in all samples and in the reference condition to get one DNA proportion for each element. These were direct inputs to the method.

3.4.6 Permutation tests

We performed null permutation experiments to estimate empirical type I error rates (denoted by α) at different nominal levels. Specifically, we created permuted sample groups that each were composed half of group 1 samples and half of group 2 samples. For example, in a six versus six comparison, we would select three samples from group 1 and three samples from group 2 to be in the first comparison group. The remaining samples would be in the second comparison group. In this way, we expect no differences in activity measures between the comparison groups. In paired experiments, we maintained the linking between samples but swapped group labels to create null comparisons.

3.4.7 Estimation of π_0

The proportion of truly null hypotheses for each dataset was estimated using the "lfdr" method in the `propTrueNull` function within limma (Phipson, 2013). This proportion was estimated for mpralm, t-test, QuASAR, edgeR, and DESeq2, and the median of these estimates was used as the estimate for π_0 for that dataset. Fisher's exact test was excluded from this estimate because it

gave an estimate of π_0 that was considerably smaller than the other methods, and which was dubious in light of its uncontrolled type I error rate. These π_0 estimates are used in the FDR calculations of Figure 3.7.

3.4.8 Simulation studies to assess accuracy of permutations for error rate estimation

To model MPRA data, we simulated negative binomial data for both DNA and RNA with a range of means and dispersion parameters, and we fix a proportion p to have differential activity between conditions. We simulated both paired and unpaired data to respectively model allelic and environmental studies.

3.4.9 Bias and variance of estimators

We use Taylor series arguments to approximate the bias and variance of the aggregate and average estimators. The following summarizes our parametric assumptions:

$$\mathrm{E}[R_b] = \mu_r = N_r p_r \qquad \mathrm{Var}(R_b) = k_r \mu_r$$

$$\mathrm{E}[D_b] = \mu_d = N_d p_d \qquad \mathrm{Var}(D_b) = k_d \mu_d$$

We suppress the dependency of these parameters on sample and element. Library sizes are given by N. The fraction of reads coming from a given element is given by p. Dispersion parameters are given by k. The common library size resulting from total count normalization is given by L. The true

activity measure of a given element is given by $a := \log(p_r/p_d)$.

3.4.10 Average estimator

The "average estimator" of a is an average of barcode-specific log activity measures and is written as:

$$\hat{a}^{AV} = \frac{1}{B}\sum_{b=1}^{B} \log\left(\frac{R_b L/N_r + 1}{D_b L/N_d + 1}\right)$$

The second-order Taylor expansion of the function

$$f(R_b, D_b) = \log(R_b L/N_r + 1) - \log(D_b L/N_d + 1)$$

about the point $(\mathrm{E}[R_b], \mathrm{E}[D_b]) = (\mu_r, \mu_d)$ is:

$$\begin{aligned}\log\left(\frac{R_b L/N_r + 1}{D_b L/N_d + 1}\right) \approx{}& \log\left(\mu_r L/N_r + 1\right) - \log\left(\mu_d L/N_d + 1\right) \\ &+ (R_b - \mu_r)\frac{L/N_r}{\mu_r L/N_r + 1} - (D_b - \mu_d)\frac{L/N_d}{\mu_d L/N_d + 1} \\ &- \frac{(L/N_r)^2}{2(\mu_r L/N_r + 1)^2}(R_b - \mu_r)^2 + \frac{(L/N_d)^2}{2(\mu_d L/N_d + 1)^2}(D_b - \mu_d)^2\end{aligned}$$

We use the expansion above to approximate the expectation of the average estimator:

$$\begin{aligned}
\mathrm{E}\left[\hat{a}^{AV}\right] &\approx \log\left(\frac{\mu_r L/N_r + 1}{\mu_d L/N_d + 1}\right) + \frac{(L/N_d)^2 k_d \mu_d}{2(\mu_d L/N_d + 1)^2} - \frac{(L/N_r)^2 k_r \mu_r}{2(\mu_r L/N_r + 1)^2} \\
&\approx \log\left(\frac{p_r}{p_d}\right) + \frac{k_d}{2\mu_d} - \frac{k_r}{2\mu_r} \\
&= a + \frac{k_d}{2\mu_d} - \frac{k_r}{2\mu_r}
\end{aligned}$$

We can also approximate the variance under the assumption that the barcode-specific log-ratios are uncorrelated:

$$\begin{aligned}
\mathrm{Var}(\hat{a}^{AV}) &= \frac{1}{B}\mathrm{Var}\left(\log\left(\frac{R_b L/N_r + 1}{D_b L/N_d + 1}\right)\right) \\
&\approx \frac{(L/N_r)^2 k_r \mu_r}{B(\mu_r L/N_r + 1)^2} + \frac{(L/N_d)^2 k_d \mu_d}{B(\mu_d L/N_d + 1)^2} - \frac{2(L/N_r)(L/N_d)\mathrm{Cov}(R_b, D_b)}{B(\mu_r L/N_r + 1)(\mu_d L/N_d + 1)}
\end{aligned}$$

3.4.11 Aggregate estimator

The "aggregate estimator" of a first aggregates counts over barcodes and is written as:

$$\hat{a}^{AGG} = \log\left(\frac{1 + (L/N_r)\sum_{b=1}^{B} R_b}{1 + (L/N_d)\sum_{b=1}^{B} D_b}\right) = \log\left(\frac{1 + (L/N_r)R^{AGG}}{1 + (L/N_d)D^{AGG}}\right)$$

The second-order Taylor expansion of the function

$$f(R^{AGG}, D^{AGG}) = \log((L/N_r)R^{AGG} + 1) - \log((L/N_d)D^{AGG} + 1)$$

about the point $(\mathrm{E}[R^{AGG}], \mathrm{E}[D^{AGG}]) = (B\mu_r, B\mu_d)$ is:

$$\log\left(\frac{1+(L/N_r)R^{AGG}}{1+(L/N_d)D^{AGG}}\right) \approx \log\left(B\mu_r L/N_r + 1\right) - \log\left(B\mu_d L/N_d + 1\right)$$
$$+ (R^{AGG} - B\mu_r)\frac{L/N_r}{B\mu_r L/N_r + 1} - (D^{AGG} - B\mu_d)\frac{L/N_d}{B\mu_d L/N_d + 1}$$
$$- \frac{(L/N_r)^2}{2(B\mu_r L/N_r + 1)^2}(R^{AGG} - B\mu_r)^2 + \frac{(L/N_d)^2}{2(B\mu_d L/N_d + 1)^2}(D^{AGG} - B\mu_d)^2$$

We use the expansion above to approximate the expectation:

$$\begin{aligned}
\mathrm{E}\left[\hat{a}^{AGG}\right] &\approx \log\left(\frac{B\mu_r L/N_r + 1}{B\mu_d L/N_d + 1}\right) + \frac{Bk_d\mu_d(L/N_d)^2}{2(B\mu_d L/N_d + 1)^2} - \frac{Bk_r\mu_r(L/N_r)^2}{2(B\mu_r L/N_r + 1)^2} \\
&\approx \log\left(\frac{p_r}{p_d}\right) + \frac{k_d}{2B\mu_d} - \frac{k_r}{2B\mu_r} \\
&= a + \frac{k_d}{2B\mu_d} - \frac{k_r}{2B\mu_r}
\end{aligned}$$

We can also approximate the variance:

$$\mathrm{Var}(\hat{a}^{AGG}) \approx$$
$$\frac{(L/N_r)^2 Bk_r\mu_r}{(B\mu_r L/N_r + 1)^2} + \frac{(L/N_d)^2 Bk_d\mu_d}{(B\mu_d L/N_d + 1)^2} - \frac{2(L/N_r)(L/N_d)\mathrm{Cov}(R^{AGG}, D^{AGG})}{(B\mu_r L/N_r + 1)(B\mu_d L/N_d + 1)}$$

3.4.12 Acknowledgements

Funding: Research reported in this publication was supported by the National Cancer Institute and the National Institute of General Medical Sciences of the National Institutes of Health under award numbers U24CA180996 and R01GM121459.

Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest: None declared.

3.5 Tables and Figures

Dataset	Description	Cell culture	Replicates	Barcodes
Differential analysis: alleles				
Tewhey	Study of 39,479 oligos coming from 29,173 variants to follow up on prior eQTL results. Large initial oligo pool: 79k. Second pool: 7.5k.	NA12878 (LCL) NA19239 (LCL) HepG2	NA12878: 5 NA19239: 3 HepG2: 5	79k pool: ∼73 7.5k pool: ∼350
Ulirsch	Study of 2756 variants in strong LD with 75 main variants to identify loci that affect RBC traits.	K562, K562 with GATA1 over-expr.	K562: 6 K562+GATA1: 4	14
Differential analysis: conditions				
Inoue	Comparison of episomal and lentiviral MPRA.	HepG2	3	Max: 99.
Shen	Study of tissue specificity of cis-regulatory elements in-vivo in mouse.	Mouse retina and cerebral cortex	3	∼8
Saturation mutagenesis				
Melnikov	Two inducible enhancers: (1) a synthetic cAMP-regulated enhancer and (2) the virus-inducible interferon-beta enhancer. Single-hit scanning alters one base at a time. Multi-hit sampling alters several bases at a time.	HEK293T	Single: 2 Multi: 2	Single: 13 Multi: 1
Kheradpour	Study of 2104 wild-type sequences and 3314 variant sequences containing targeted motif disruptions to understand base-level effects in motifs.	K562, HepG2	2	10

Table 3.1: Datasets used for investigations in this paper. All datasets are publicly available.

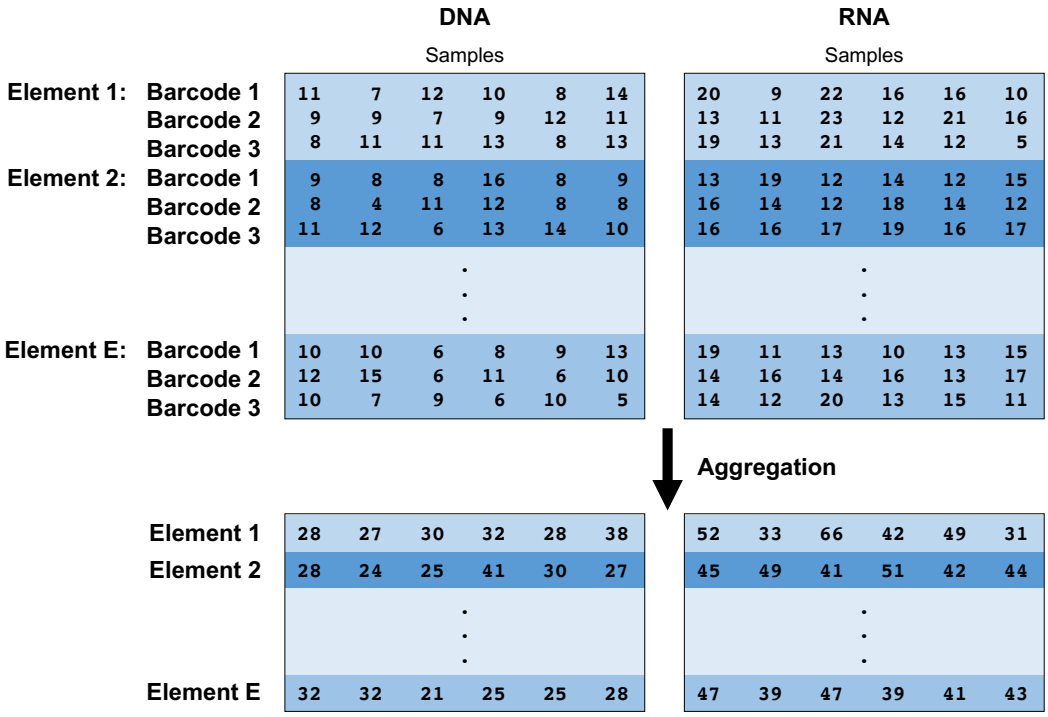


Figure 3.1: Structure of MPRA data. Thousands of putative regulatory elements can be assayed at a time in an MPRA experiment. Each element is linked to multiple barcodes. A plasmid library containing these barcoded elements is transfected into several cell populations (samples). Cellular DNA and RNA can be isolated and sequenced. The barcodes associated with each putative regulatory element can be counted to obtain relative abundances of each element in DNA and RNA. The process of aggregation sums counts over barcodes for element in each sample. Aggregation is one method for summarizing barcode level data into element level data.

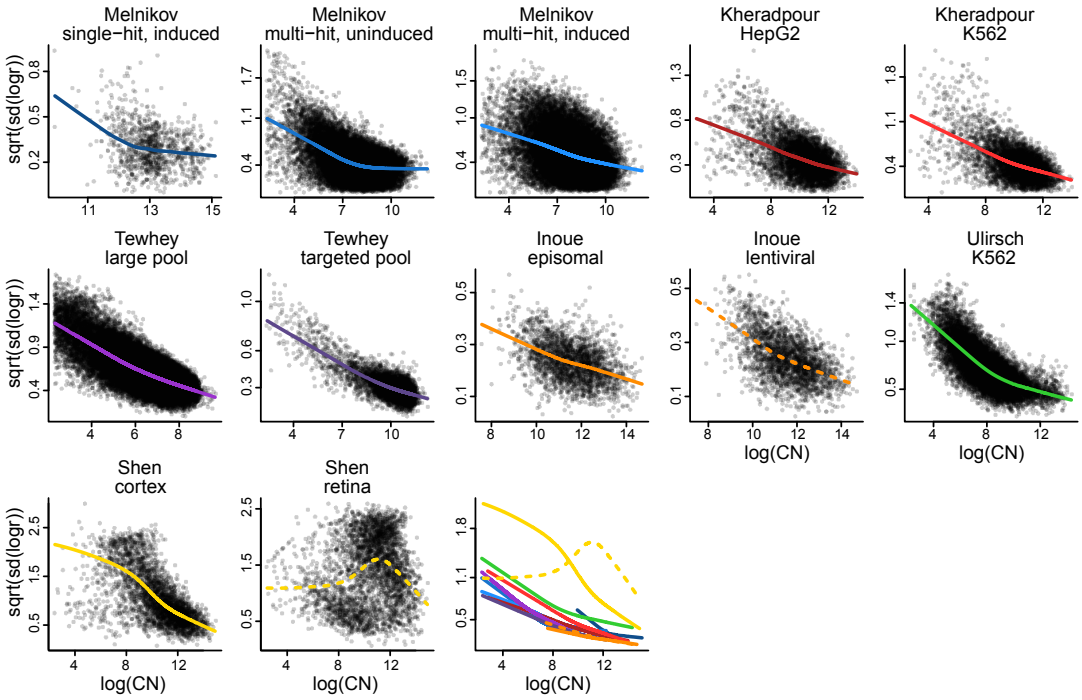


Figure 3.2: Variability of MPRA activity measures depends on element copy number. For multiple publicly available datasets we compute activity measures of putative regulatory element as the log2 ratio of aggregated RNA counts over aggregated DNA counts. Each panel shows the relationship between variability (across samples) of these activity measures and the average log2 DNA levels (across samples). Smoothed relationships are lowess curves representing the local average variability. The last plot shows all lowess curves on the same figure.

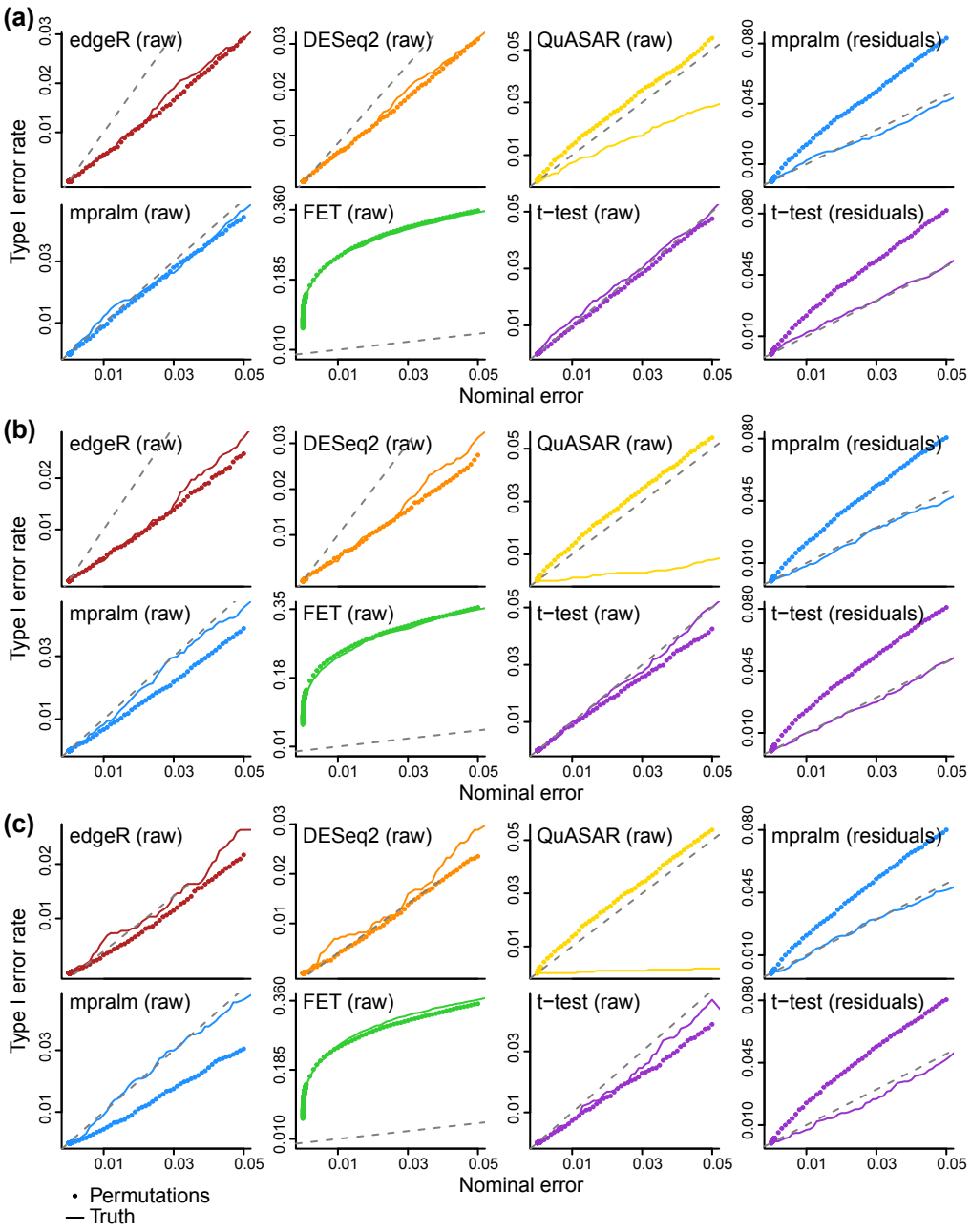


Figure 3.3: Estimation accuracy of type I error rates using permutations on simulated data. The three sets of panels vary the true proportion (p) of elements with differential activity. (a) $p = 0.1$, (b) $p = 0.3$, (c) $p = 0.5$. Each panel shows one method used for differential analysis and compares the true type I error rate to that estimated from null permutations. For mpralm and the t-test, we show error rates from permuting both the raw data and residuals.

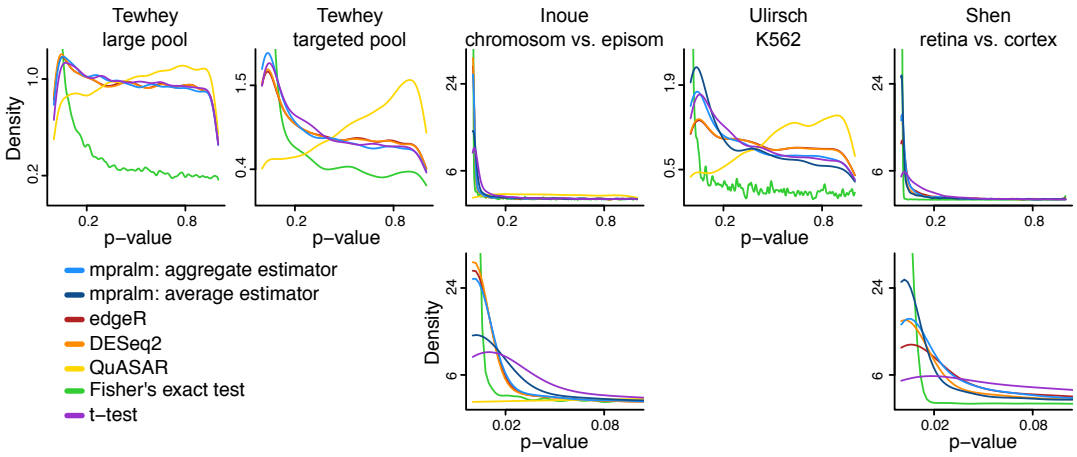


Figure 3.4: Comparison of detection rates and p-value calibration over datasets. The distribution of p-values for all datasets, including a zoom of the $[0, 0.1]$ interval for some datasets. Over all datasets, most methods show p-values that closely follow the classic mixture of uniformly distributed p-values with an enrichment of low p-values for differential elements. For the datasets which had barcode level counts (Inoue, Ulirsch, and Shen), we used two types of estimators of the activity measure (log ratio of RNA/DNA) with mpralm, shown in light and dark blue. We were not able to run QuASAR on the Shen mouse dataset.

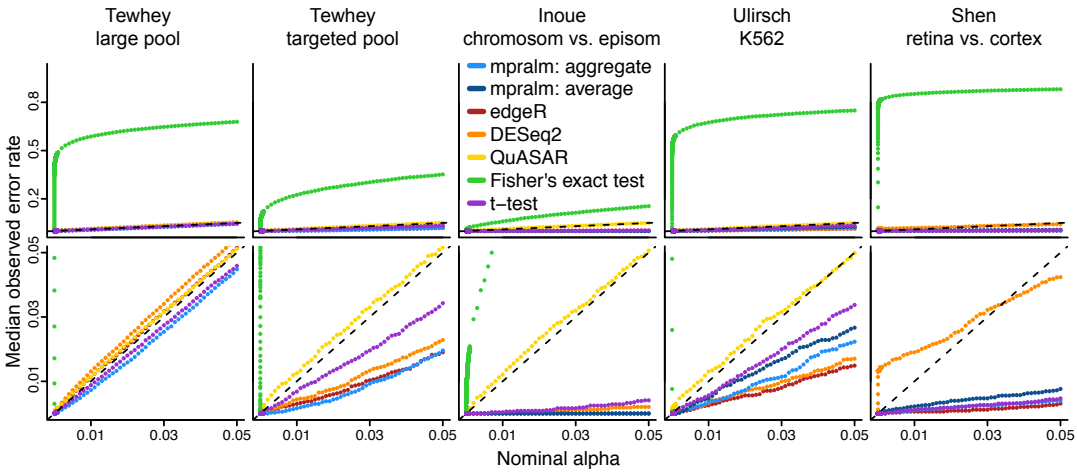


Figure 3.5: Empirical type I error rates. Type I error rates were estimated for all methods at different nominal levels with null permutation experiments (Methods). For the datasets which had barcode level counts (Inoue, Ulirsch, and Shen), we used two types of estimators of the activity measure (aggregate and average estimator) with mpralm, shown in dark and light blue.

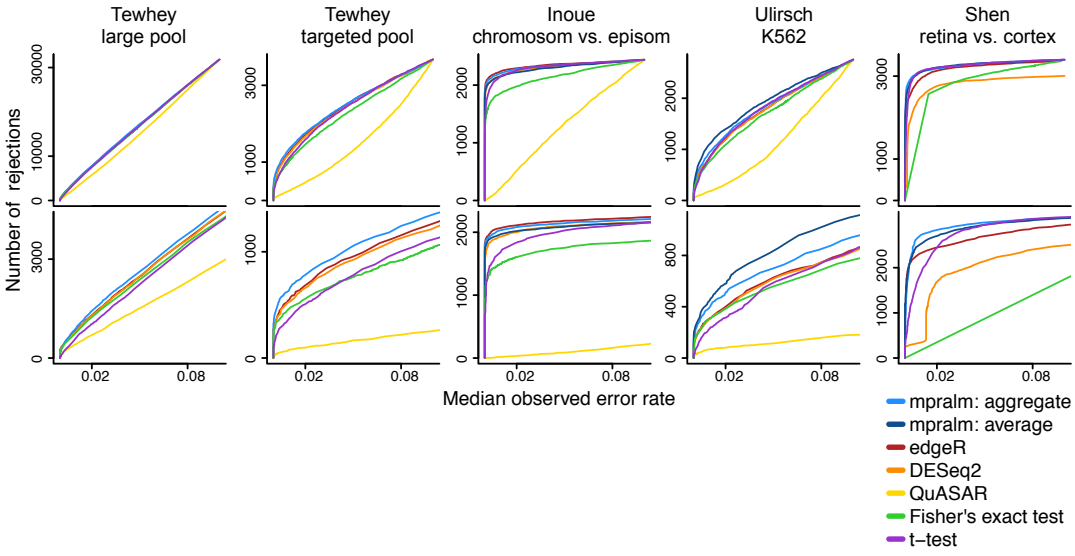


Figure 3.6: Number of rejections as a function of observed error rate. To compare the detection (rejection) rates of the methods fairly, we compare them at the same observed type I error rates, estimated in Figure 3.5. The bottom row is a zoomed-in version of the top row. We see that mpralm, edgeR, and DESeq2 consistently have the highest detection rates.

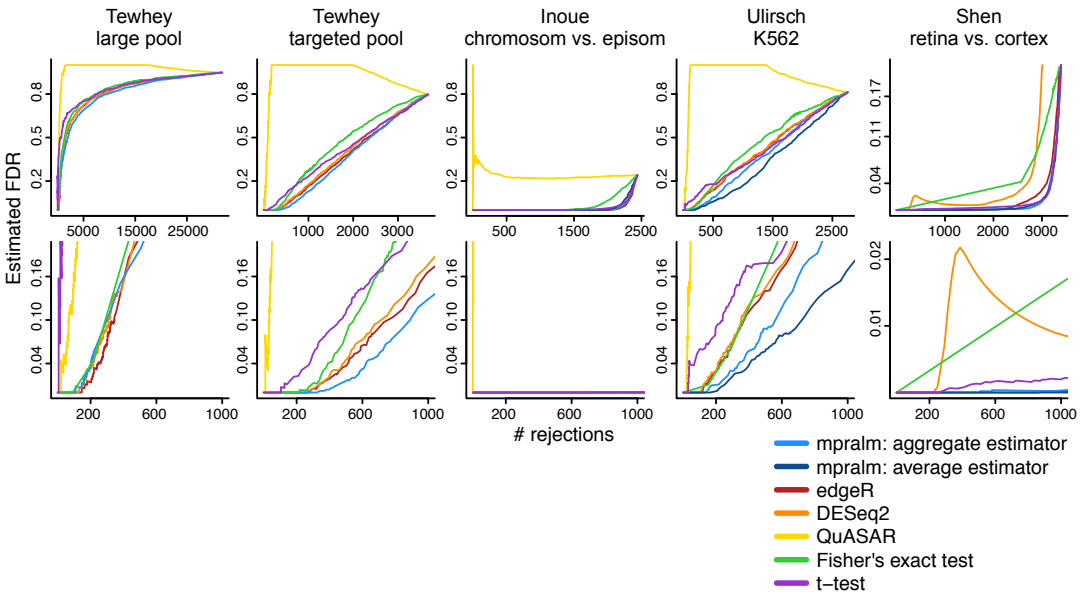


Figure 3.7: Estimated FDR. For each dataset and method, the false discovery rate is estimated as a function of the number of rejections. This requires estimation of the proportion of true null hypotheses (Supplemental Methods). The bottom row is a zoomed-in version of the top row.

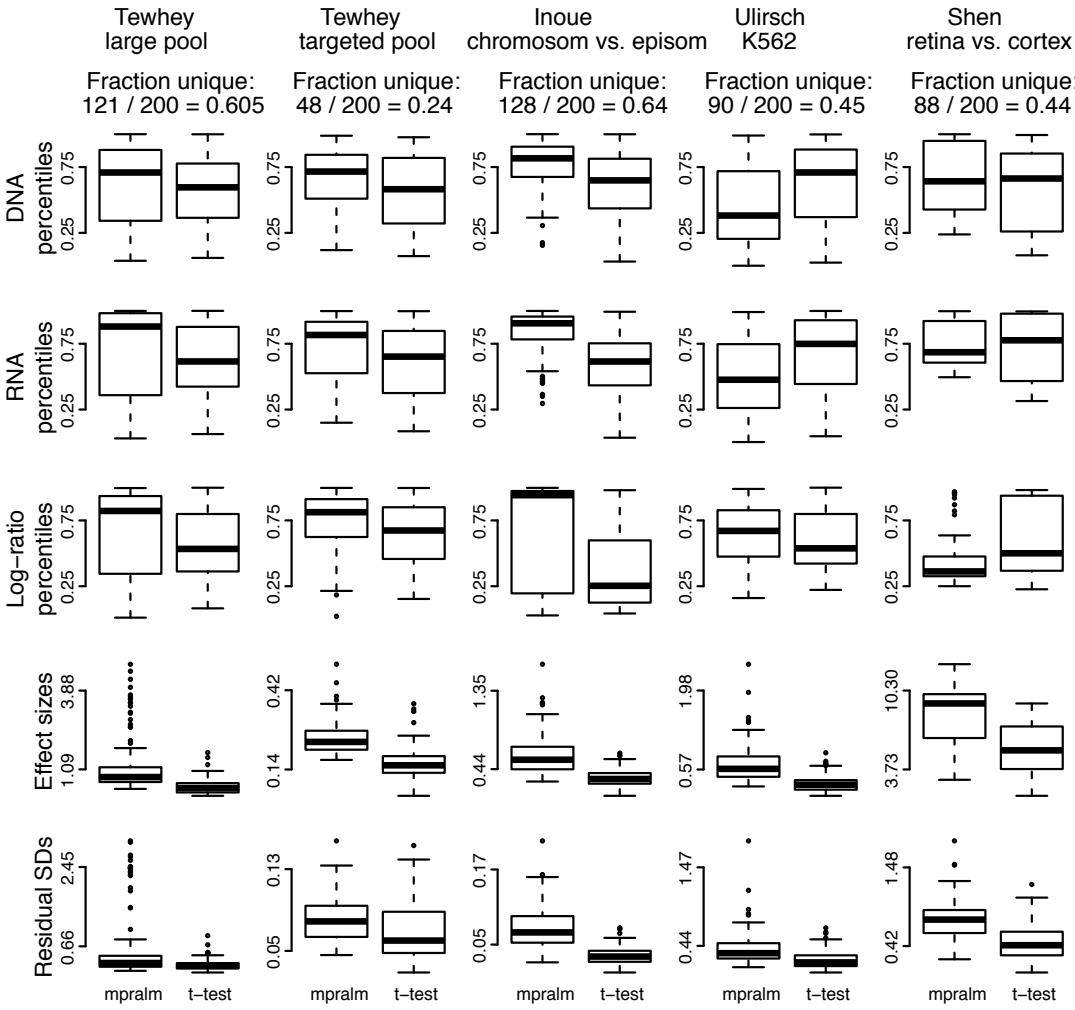


Figure 3.8: Distribution of quantities related to statistical inference in top ranked elements with mpralm and t-test. MPRA elements that appear in the top 200 elements with one method but not the other are examined here. For these uniquely top ranking elements, the DNA, RNA, and log-ratio percentiles are shown in the first three rows. The effect sizes (difference in mean log-ratios) and residual standard deviations are shown in the last two rows. Overall, uniquely top ranking elements for the t-test tend to have lower log-ratio activity measures, effect sizes, and residual standard deviations.

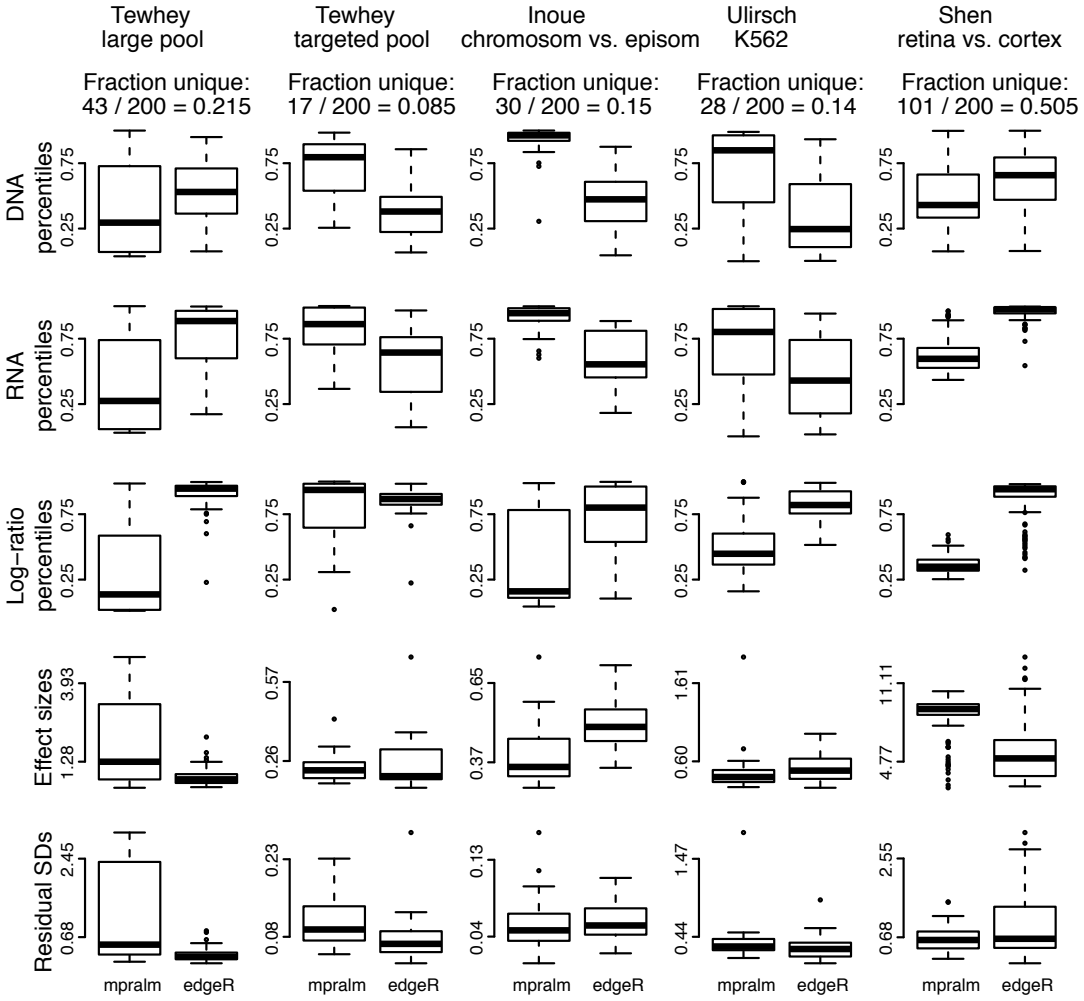


Figure 3.9: Distribution of quantities related to statistical inference in top ranked elements with mpralm and edgeR. Similar to Figure 3.8.

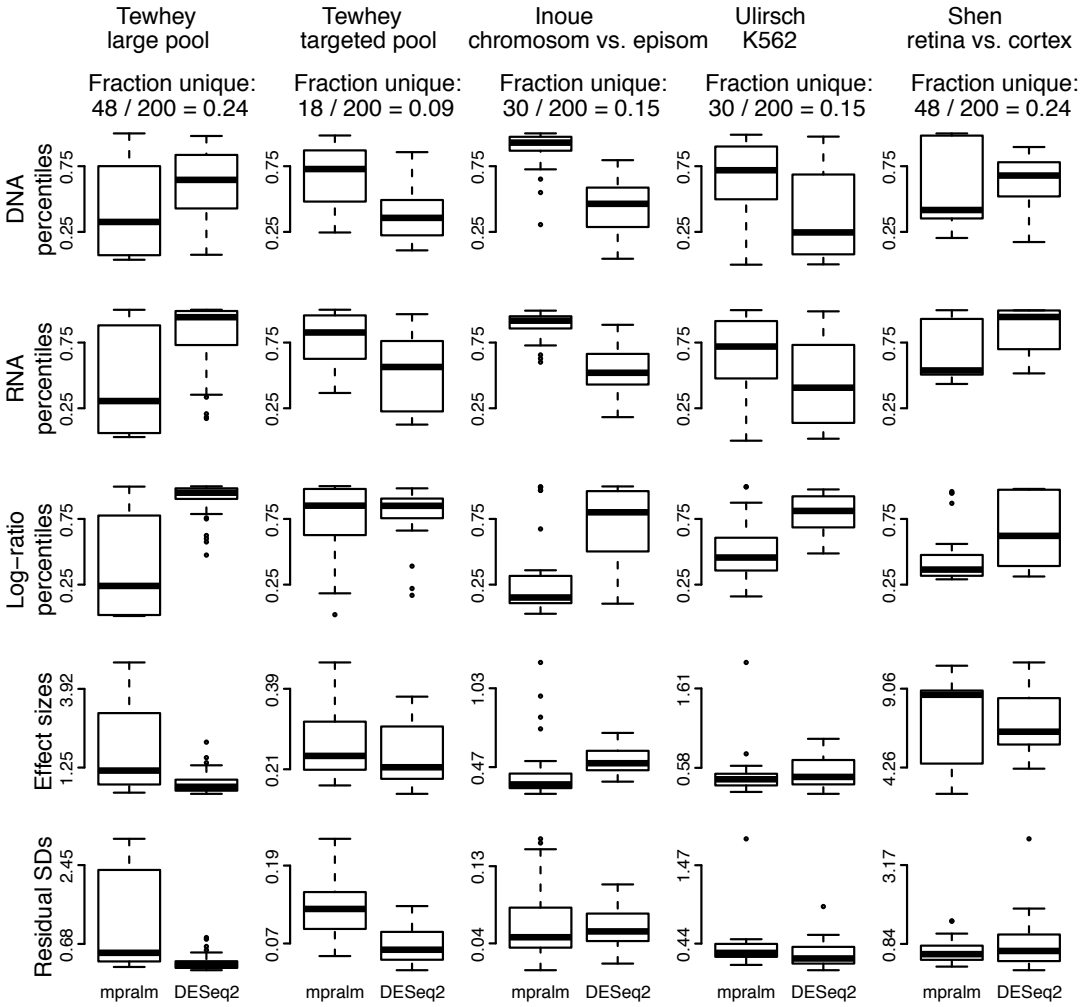


Figure 3.10: Distribution of quantities related to statistical inference in top ranked elements with mpralm and DESeq2. Similar to Figure 3.8.

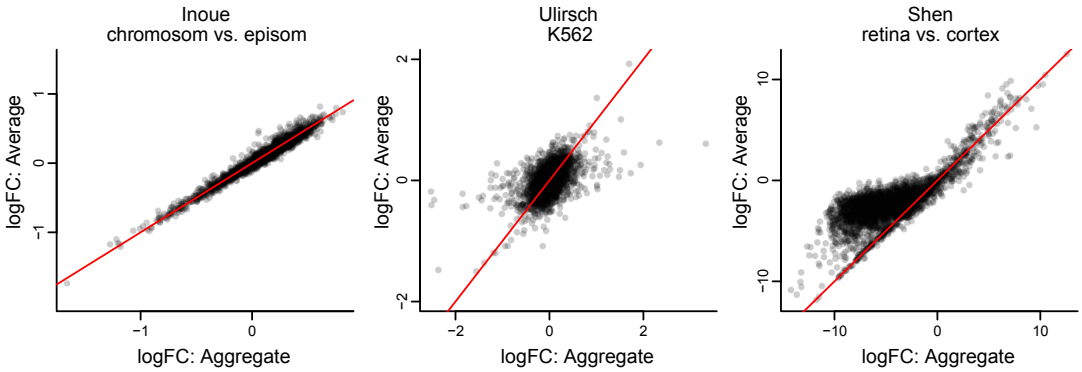


Figure 3.11: Comparison of the average and aggregate estimators For the three datasets containing barcode-level information, we compare the effect sizes (log fold changes in activity levels) resulting from use of the aggregate and average estimators. The $y = x$ line is shown in red.

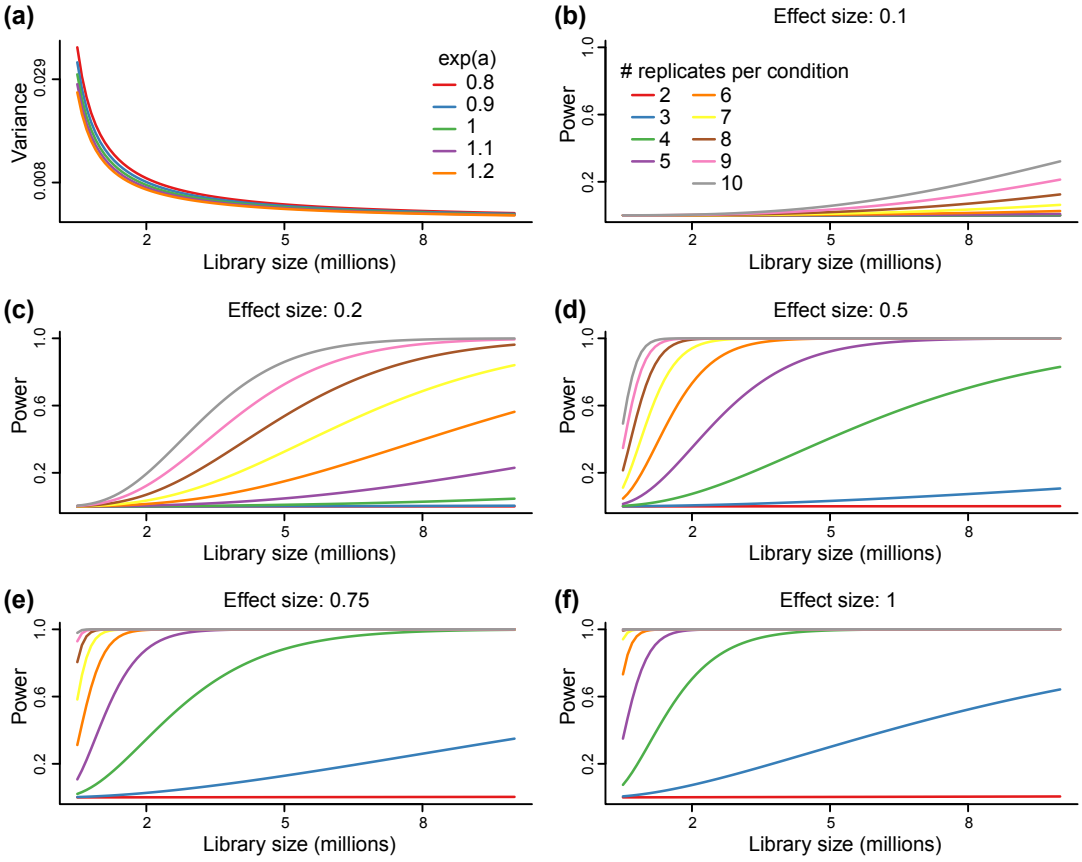


Figure 3.12: Power analysis. Variance and power calculated based on our theoretical model. **(a)** Variance of the aggregate estimator depends on library size and the true unknown activity level but not considerably on the latter. **(b)-(f)** Power curves as a function of library size for different effect sizes and sample sizes. Effect sizes are $\log_2$ fold-changes.

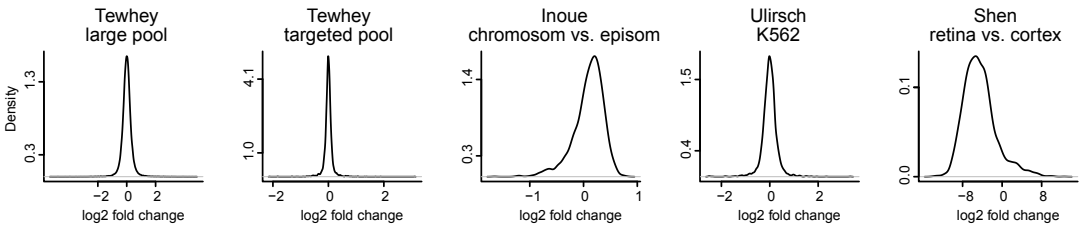


Figure 3.13: Effect size distributions across datasets. Effect sizes in MPRA differential analysis are the (precision-weighted) differences in activity scores between groups, also called $\log_2$ fold-changes. The distribution of $\log_2$ fold changes resulting from using mpralm with the aggregate estimator are shown here.

References

White, Michael A (2015). "Understanding how cis-regulatory function is encoded in DNA sequence using massively parallel reporter assays and designed sequences". In: *Genomics* 106, pp. 165–170. DOI: 10.1016/j.ygeno.2015.06.003.

Melnikov, Alexandre, Xiaolan Zhang, Peter Rogov, Li Wang, and Tarjei S Mikkelsen (2014). "Massively parallel reporter assays in cultured mammalian cells". In: *J. Vis. Exp.* DOI: 10.3791/51719.

Grossman, Sharon R, Xiaolan Zhang, Li Wang, Jesse Engreitz, Alexandre Melnikov, Peter Rogov, Ryan Tewhey, Alina Isakova, Bart Deplancke, Bradley E Bernstein, Tarjei S Mikkelsen, and Eric S Lander (2017). "Systematic dissection of genomic features determining transcription factor binding and enhancer function". In: *PNAS* 114, E1291–E1300. DOI: 10.1073/pnas.1621150114.

Guo, Cong, Ian C McDowell, Michael Nodzenski, Denise M Scholtens, Andrew S Allen, William L Lowe, and Timothy E Reddy (2017). "Transversions have larger regulatory effects than transitions". In: *BMC Genomics* 18, p. 394. DOI: 10.1186/s12864-017-3785-4.

Safra, Modi, Ronit Nir, Daneyal Farouq, Ilya Vainberg Slutskin, and Schraga Schwartz (2017). "TRUB1 is the predominant pseudouridine synthase acting on mammalian mRNA via a predictable and conserved code". In: *Genome Research* 27, pp. 393–406. DOI: 10.1101/gr.207613.116.

Levo, Michal, Tali Avnit-Sagi, Maya Lotan-Pompan, Yael Kalma, Adina Weinberger, Zohar Yakhini, and Eran Segal (2017). "Systematic Investigation of Transcription Factor Activity in the Context of Chromatin Using Massively Parallel Binding and Expression Assays". In: *Mol. Cell* 65, 604–617.e6. DOI: 10.1016/j.molcel.2017.01.007.

Maricque, Brett B, Jose Dougherty, and Barak A Cohen (2017). "A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of cis-regulatory activity in neural cells". In: *Nucleic Acids Research* 45, e16–e16. DOI: 10.1093/nar/gkw942.

Groff, Abigail F, Diana B Sanchez-Gomez, Marcela M L Soruco, Chiara Gerhardinger, A Rasim Barutcu, Eric Li, Lara Elcavage, Olivia Plana, Lluvia V Sanchez, James C Lee, Martin Sauvageau, and John L Rinn (2016). "In Vivo Characterization of Linc-p21 Reveals Functional cis-Regulatory DNA Elements". In: *Cell Reports* 16, pp. 2178–2186. DOI: 10.1016/j.celrep.2016.07.050.

Ernst, Jason, Alexandre Melnikov, Xiaolan Zhang, Li Wang, Peter Rogov, Tarjei S Mikkelsen, and Manolis Kellis (2016). "Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions". In: *Nature Biotechnology.* 34, pp. 1180–1190. DOI: 10.1038/nbt.3678.

White, Michael A, Jamie C Kwasnieski, Connie A Myers, Susan Q Shen, Joseph C Corbo, and Barak A Cohen (2016). "A Simple Grammar Defines Activating and Repressing cis-Regulatory Elements in Photoreceptors". In: *Cell Reports* 5, pp. 1247–1254. DOI: 10.1016/j.celrep.2016.09.066.

Ferreira, Leonardo M R, Torsten B Meissner, Tarjei S Mikkelsen, William Mallard, Charles W O'Donnell, Tamara Tilburgs, Hannah A B Gomes, Raymond Camahort, Richard I Sherwood, David K Gifford, John L Rinn, Chad A Cowan, and Jack L Strominger (2016). "A distant trophoblast-specific enhancer controls HLA-G expression at the maternal-fetal interface". In: *PNAS* 113, pp. 5364–5369. DOI: 10.1073/pnas.1602886113.

Fiore, Chris and Barak A Cohen (2016). "Interactions between pluripotency factors specify cis-regulation in embryonic stem cells". In: *Genome Research* 26, pp. 778–786. DOI: 10.1101/gr.200733.115.

Farley, Emma K, Katrina M Olson, Wei Zhang, Alexander J Brandt, Daniel S Rokhsar, and Michael S Levine (2015). "Suboptimization of developmental enhancers". In: *Science* 350, pp. 325–328. DOI: 10.1126/science.aac6948.

Kamps-Hughes, Nick, Jessica L Preston, Melissa A Randel, and Eric A Johnson (2015). "Genome-wide identification of hypoxia-induced enhancer regions". In: *PeerJ* 3, e1527. DOI: 10.7717/peerj.1527.

Dickel, Diane E, Yiwen Zhu, Alex S Nord, John N Wylie, Jennifer A Akiyama, Veena Afzal, Ingrid Plajzer-Frick, Aileen Kirkpatrick, Berthold Göttgens, Benoit G Bruneau, Axel Visel, and Len A Pennacchio (2014). "Function-based identification of mammalian enhancers using site-specific integration". In: *Nature Methods* 11, pp. 566–571. DOI: 10.1038/nmeth.2886.

Kwasnieski, Jamie C, Christopher Fiore, Hemangi G Chaudhari, and Barak A Cohen (2014). "High-throughput functional testing of ENCODE segmentation predictions". In: *Genome Research* 24, pp. 1595–1602. DOI: 10.1101/gr.173518.114.

Mogno, Ilaria, Jamie C Kwasnieski, and Barak A Cohen (2013). "Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants". In: *Genome Research* 23, pp. 1908–1915. DOI: 10.1101/gr.157891.113.

Gisselbrecht, Stephen S, Luis A Barrera, Martin Porsch, Anton Aboukhalil, Preston W Estep 3rd, Anastasia Vedenko, Alexandre Palagi, Yongsok Kim, Xianmin Zhu, Brian W Busser, Caitlin E Gamble, Antonina Iagovitina, Aditi Singhania, Alan M Michelson, and Martha L Bulyk (2013). "Highly parallel assays of tissue-specific enhancers in whole Drosophila embryos". In: *Nature Methods* 10, pp. 774–780. DOI: 10.1038/nmeth.2558.

White, Michael A, Connie A Myers, Joseph C Corbo, and Barak A Cohen (2013). "Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks". In: *PNAS* 110.29, pp. 11952–11957. DOI: 10.1073/pnas.1307449110.

Smith, Robin P, Leila Taher, Rupali P Patwardhan, Mee J Kim, Fumitaka Inoue, Jay Shendure, Ivan Ovcharenko, and Nadav Ahituv (2013). "Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model". In: *Nature Genetics* 45, pp. 1021–1028. DOI: 10.1038/ng.2713.

Patwardhan, Rupali P, Choli Lee, Oren Litvin, David L Young, Dana Pe'er, and Jay Shendure (2009). "High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis". In: *Nature Biotechnology* 27, pp. 1173–1175. DOI: 10.1038/nbt.1589.

Melnikov, Alexandre, Anand Murugan, Xiaolan Zhang, Tiberiu Tesileanu, Li Wang, Peter Rogov, Soheil Feizi, Andreas Gnirke, Curtis G Callan Jr, Justin B Kinney, Manolis Kellis, Eric S Lander, and Tarjei S Mikkelsen (2012). "Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay". In: *Nature Biotechnology* 30, pp. 271–277. DOI: 10.1038/nbt.2137.

Patwardhan, Rupali P, Joseph B Hiatt, Daniela M Witten, Mee J Kim, Robin P Smith, Dalit May, Choli Lee, Jennifer M Andrie, Su-In Lee, Gregory M Cooper, Nadav Ahituv, Len A Pennacchio, and Jay Shendure (2012). "Massively parallel functional dissection of mammalian enhancers in vivo". In: *Nature Biotechnology* 30, pp. 265–270. DOI: 10.1038/nbt.2136.

Kwasnieski, Jamie C, Ilaria Mogno, Connie A Myers, Joseph C Corbo, and Barak A Cohen (2012). "Complex effects of nucleotide variants in a mammalian cis-regulatory element". In: *PNAS* 109, pp. 19498–19503. DOI: 10.1073/pnas.1210678109.

Kheradpour, Pouya, Jason Ernst, Alexandre Melnikov, Peter Rogov, Li Wang, Xiaolan Zhang, Jessica Alston, Tarjei S Mikkelsen, and Manolis Kellis (2013). "Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay". In: *Genome Research* 23, pp. 800–811. DOI: 10.1101/gr.144899.112.

Birnbaum, Ramon Y, Rupali P Patwardhan, Mee J Kim, Gregory M Findlay, Beth Martin, Jingjing Zhao, Robert J A Bell, Robin P Smith, Angel A Ku, Jay Shendure, and Nadav Ahituv (2014). "Systematic dissection of coding exons at single nucleotide resolution supports an additional role in cell-specific transcriptional regulation". In: *PLOS Genetics* 10, e1004592. DOI: 10.1371/journal.pgen.1004592.

Zhao, Wenxue, Joshua L Pollack, Denitza P Blagev, Noah Zaitlen, Michael T McManus, and David J Erle (2014). "Massively parallel functional annotation of 3' untranslated regions". In: *Nature Biotechnology* 32, pp. 387–391. DOI: 10.1038/nbt.2851.

Ulirsch, Jacob C, Satish K Nandakumar, Li Wang, Felix C Giani, Xiaolan Zhang, Peter Rogov, Alexandre Melnikov, Patrick McDonel, Ron Do, Tarjei S Mikkelsen, and Vijay G Sankaran (2016). "Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits". In: *Cell* 165, pp. 1530–1545. DOI: 10.1016/j.cell.2016.04.048.

Tewhey, Ryan, Dylan Kotliar, Daniel S Park, Brandon Liu, Sarah Winnicki, Steven K Reilly, Kristian G Andersen, Tarjei S Mikkelsen, Eric S Lander, Stephen F Schaffner, and Pardis C Sabeti (2016). "Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay". In: *Cell* 165.6, pp. 1519–1529. DOI: 10.1016/j.cell.2016.04.027.

Vockley, Christopher M, Cong Guo, William H Majoros, Michael Nodzenski, Denise M Scholtens, M Geoffrey Hayes, William L Lowe Jr, and Timothy E Reddy (2015). "Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort". In: *Genome Research* 25, pp. 1206–1214. DOI: 10.1101/gr.190090.115.

Inoue, Fumitaka, Martin Kircher, Beth Martin, Gregory M Cooper, Daniela M Witten, Michael T McManus, Nadav Ahituv, and Jay Shendure (2017). "A systematic comparison reveals substantial differences in chromosomal

versus episomal encoding of enhancer activity". In: *Genome Research* 27, pp. 38–52. DOI: 10.1101/gr.212092.116.

Shen, Susan Q, Connie A Myers, Andrew E O Hughes, Leah C Byrne, John G Flannery, and Joseph C Corbo (2016). "Massively parallel cis-regulatory analysis in the mammalian central nervous system". In: *Genome Research* 26, pp. 238–255. DOI: 10.1101/gr.193789.115.

Kalita, Cynthia A, Gregory A Moyerbrailean, Christopher Brown, Xiaoquan Wen, Francesca Luca, and Roger Pique-Regi (2017). "QuASAR-MPRA: Accurate allele-specific analysis for massively parallel reporter assays". In: *Bioinformatics*. DOI: 10.1093/bioinformatics/btx598.

Law, Charity W, Yunshun Chen, Wei Shi, and Gordon K Smyth (2014). "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts". In: *Genome Biology* 15, R29. DOI: 10.1186/gb-2014-15-2-r29.

McCarthy, Davis J, Yunshun Chen, and Gordon K Smyth (2012). "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation". In: *Nucleic Acids Research* 40, pp. 4288–4297. DOI: 10.1093/nar/gks042.

Marioni, John C, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays". In: *Genome Research* 18, pp. 1509–1517. DOI: 10.1101/gr.079558.108.

Bullard, James H, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit (2010). "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments". In: *BMC Bioinformatics* 11, p. 94. DOI: 10.1186/1471-2105-11-94.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26, pp. 139–140. DOI: 10.1093/bioinformatics/btp616.

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15, p. 550. DOI: 10.1186/s13059-014-0550-8.

Jiang, Dan (2017). "Adjustment Procedure to Permutation Tests in Epigenomic Differential Analysis". PhD thesis. Johns Hopkins Bloomberg School of Public Health.

Phipson, Belinda (2013). "Empirical bayes modelling of expression profiles and their associations". PhD thesis.

Smyth, Gordon K, Joëlle Michaud, and Hamish S Scott (2005). "Use of within-array replicate spots for assessing differential expression in microarray experiments". In: *Bioinformatics* 21, pp. 2067–2075. DOI: 10.1093/bioinformatics/bti270.

Chapter 4

Evidence-based data analysis

4.1 Introduction

Data analysis is a multistage process that begins with the statement of a question; spans iterative phases of cleaning, exploration, and modeling; and ends with the communication of results. At each of these stages, analysts must interact with data and their beliefs to make judgments about what to do next. Understanding the factors that influence these judgments is important for improving the general practice of data analysis and the training of data analysts. In this chapter, we describe in detail one randomized experiment that examines the impact of explanation on perception of causality. We also briefly review results from another randomized experiment that studies the qualities of plots made in two different R graphics systems.

4.2 Explanation and causal interpretation

4.2.1 Introduction

Facebook could raise your risk of cancer (*How using Facebook could raise your risk of cancer*), drinking too much tea causes prostate cancer (PTI, 2016), eating chocolate helps people stay thin (*Eating lots of chocolate helps people stay thin, study finds*). We all know that correlation does not imply causation, but we have also all seen exaggerated headlines in the media that fall short in capturing the true results of a scientific study. A recent report in the British Medical Journal found the fault may not lie entirely with the media (Sumner et al., 2014), but may be aided by exaggerated press releases from universities themselves. In fact, in their study of 462 press releases, the study authors found that 33% (26% to 40%) contained exaggerated causal claims. Regardless, of where the exaggeration happens, a result seems more realistic if you can explain why you think it is happening.

Most researchers do not deliberately claim causal results in an observational study. But do we lead our readers to draw a causal conclusion unintentionally by explaining why significant correlations and relationships may exist? Once we discover that an association exists, it is natural to want to explain why it does. We may describe potential mechanisms, make connections to previous literature, or put an observation in context. Despite these explanations, causal relationships are not proven in a single observational study and are only increasingly substantiated over the course of many such studies. There is observational evidence suggesting a noticeable prevalence

of inappropriate causal language in both nutritional (Cofield, Corona, and Allison, 2010) and educational (Robinson et al., 2007) research studies.

The distinction between correlational and causal evidence is not merely a pedantic formality. Because causal statements carry moral underpinnings, they can have dangerous consequences for societal perceptions of certain groups, products, or practices when consumed and interpreted by the general public (Lombrozo, 2017). For example, researchers of developmental origins of health and disease published a cautionary commentary in response to a collection of headlines (Mother's diet during pregnancy alters baby's DNA, Pregnant 9/11 survivors transmitted trauma to their children) that seemed to vilify mothers for developmental outcomes in babies (Richardson et al., 2014). In research areas dealing with human subjects, mistakes in perceptions about evidence can be harmful, and reporters must use great care in the language they use to describe scientific findings. The danger in these headlines and in related causal language (e.g. explanatory statements, jargon) lies not in the words themselves but in their interpretation by the public.

In this work, we investigate how interpretation of scientific evidence is affected by a specific area of causal language: explanation. We report the results of a randomized experiment performed on an online educational platform that suggest a strong effect of explanatory language on students' perception of whether a study is correlational or causal.

Type of analysis	Goal of analysis
Descriptive	Summarizing the data without interpretation
Exploratory	Summarizing the data with interpretation, but without generalization beyond the original sample
Inferential	Generalizing beyond the original sample, with the goal of describing an association in a larger population
Predictive	Generalizing beyond the original sample, with the goal of predicting a measurement for a new individual
Causal	Generalizing beyond the original sample, with the goal of learning how changing the average of one measurement affects, on average, another measurement
Mechanistic	Generalizing beyond the original sample, with the goal of learning how changing one measurement deterministically affects another variable's measurement

Table 4.1: Goals for different analysis types. These analysis types form the set of possible answer choices in our randomized experiment and were taught to students before the experiment was performed.

4.2.2 Study Design

Different types of studies have different analysis goals (Table 4.1) (Leek and Peng, 2015). We were interested in whether people can distinguish between a study whose goal was inferential and one whose goal was actually causal, as this is a common error often termed "correlation does not equal causation". We wanted to know whether including language explaining an observed association leads people to believe that an inferential study is causal. To test this hypothesis, we ran an experiment in a large online open-access data analysis course. This introductory-level course covered basic data analytic concepts. Our experiment involved a single randomized quiz question administered during the course. We originally ran the experiment in January 2013, but later independently replicated our experiment in a separate offering of the course in October 2013. Between these two replications, over 22,000 students completed versions of our experimental question.

Early in the course, students were presented with the definitions of six possible types of data analysis (descriptive, exploratory, inferential, predictive, causal, and mechanistic) consistent with those shown in Table 4.1. In the subsequent course quiz, we provided students with an description of an inferential study - from which we can only infer correlation:

> *We take a random sample of individuals in a population and identify whether they smoke and if they have cancer. We observe that there is a strong relationship between whether a person in the sample smoked or not and whether they have lung cancer. We claim that the smoking is related to lung cancer in the larger population.*

We randomized students to see or not see an explanatory interpretation accompanying this description. Students in this explanatory interpretation group saw an additional sentence:

> *We explain we think that the reason for this relationship is because cigarette smoke contains known carcinogens such as arsenic and benzene, which make cells in the lungs become cancerous.*

All students were then asked to identify the type of analysis for these results. In addition to the correct answer (inferential), students were presented at random with three of four possible incorrect answer choices (descriptive, causal, predictive, mechanistic). That is, approximately 25% of students made their choice from inferential, descriptive, causal, and predictive, approximately 25% from inferential, descriptive, causal, and mechanistic, and so on.

Although the described analysis is inferential in nature, we hypothesized that students who saw the explanatory language would be more likely to identify the analysis as causal if given that choice. Because students were able to retake this quiz multiple times in order to achieve a passing grade, we collected answers from each student's first attempt (Table 4.3).

4.2.3 Results

In our original experiment (January 2013), 20,257 students completed our experimental quiz question. These students were randomly assigned to one of four arms, where each arm contained the correct answer choice (inferential) and three incorrect answer choices (from among causal, descriptive, predictive, and mechanistic). Sample sizes are given in Table 4.2. We present detailed results for two arms: (1) students who chose between inferential, causal, predictive, and mechanistic analyses and (2) students who were not given causal as a choice, but instead chose between inferential, descriptive, predictive, and mechanistic analyses. Table 4.2 shows summary results for the four groups of students corresponding to the four sets of answer choices seen.

Among students selecting from inferential, causal, predictive, and mechanistic answer choices, the majority (68.5%) correctly answered that the description referred to an inferential data analysis (Table 4.3). However, a significantly higher percentage of students who were shown the explanatory language claimed it was a causal analysis compared to students who did not see the additional language: 31.8% compared to 16.6%. These results indicate that

	Difference in percentage choosing "causal" when seeing explanatory language vs. not seeing explanatory language (95% CI)	
Answer choices seen	January 2013 course	October 2013 course
inferential, causal, descriptive, predictive	14.5% (12.2%, 16.8%) N = 5061	14.3% (6.4%, 22.2%) N = 447
inferential, causal, descriptive, mechanistic	15.8% (13.4%, 18.1%) N = 5092	14.8% (6.6%, 23.0%) N = 463
inferential, causal, predictive, mechanistic	15.2% (12.8%, 17.5%) N = 5088	19.9% (11.5%, 28.3%) N = 437
	Difference in percentage choosing "inferential" when seeing explanatory language vs. not seeing explanatory language (95% CI)	
inferential, descriptive, predictive, mechanistic	-7.3% (-9.3%, -5.2%) N = 5016	-4.9% (-12.6%, 2.9%) N = 416

Table 4.2: Effect of explanatory language on student responses. Students were randomized to one of four arms containing different sets of answer choices. Differences in the percentage choosing the "causal" and "inferential" answer choices are given, as well as 95% confidence intervals for the differences and sample sizes.

explanatory language increases the chance a student will mistake an inferential result as causal. In this case, students who saw the additional explanation were almost twice as likely to claim the results as causal.

This increase in the choice of a causal analysis when faced with explanatory language corresponded to a decrease in choice of an inferential analysis. The percentages of students who chose either a predictive or descriptive analysis were similar between the two treatment groups. However, there was an increase in the percentage of students who claimed the result was mechanistic in the explanatory language group: 3.5% compared to 1.2%. This is not surprising since a mechanistic result is similar to a causal result in that it describes a deterministic process by which one variable affects another.

Among students who were not given the option to select "causal" as an answer (selecting instead from inferential, predictive, descriptive, and mechanistic analyses), a higher percentage (84.6%) correctly answered that the description referred to an inferential data analysis (Table 4). In this case, a

This is an example of a/an _____ data analysis.	January 2013 course N = 5088		October 2013 course N = 437	
	Saw explanatory language N = 2516	No explanatory language N = 2572	Saw explanatory language N = 199	No explanatory language N = 238
inferential	1508 (59.9%)	1977 (76.9%)	116 (58.3%)	190 (79.8%)
causal	799 (31.8%)	427 (16.6%)	68 (34.2%)	34 (14.3%)
predictive	120 (4.8%)	138 (5.4%)	8 (4.0%)	11 (4.6%)
mechanistic	89 (3.5%)	30 (1.2%)	7 (3.5%)	3 (1.3%)

Table 4.3: Detailed results for the arm with answer choices: inferential, causal, predictive, and mechanistic. Results for randomized controlled experiment asking students to identify the type of data analysis in a scenario. The quiz question described an inferential analysis. Students were randomized to see or not see explanatory language that hypothesized why the association occurred. In the presence of explanatory language, nearly twice as many students selected "causal" as the answer. The presence of explanatory language also corresponds to a decrease the in the percentage of students correctly selecting "inferential" as the answer.

This is an example of a/an _____ data analysis.	January 2013 course N = 5016		October 2013 course N = 416	
	Saw explanatory language N = 2485	No explanatory language N = 2531	Saw explanatory language N = 199	No explanatory language N = 217
inferential	2011 (80.9%)	2232 (88.2%)	160 (80.4%)	185 (85.3%)
predictive	196 (7.9%)	181 (7.2%)	10 (5.0%)	12 (5.5%)
descriptive	138 (5.6%)	82 (3.2%)	14 (7.0%)	14 (6.5%)
mechanistic	140 (5.6%)	36 (1.4%)	15 (7.5%)	6 (2.8%)

Table 4.4: Detailed results for the arm with answer choices: inferential, descriptive, predictive, and mechanistic (no causal). In the presence of explanatory language, a lower percentage of students correctly selected "inferential" as the answer and a higher percentage of students incorrectly selected "mechanistic" as the answer.

significantly higher percentage of students correctly claimed the analysis was inferential when not shown the explanatory language: 88.2% compared to 80.9% These results indicate that, even without the ability to identify the analysis as causal, students had a harder time correctly identifying an inferential study when given hypothesized information about the reason for a correlation. The size of the effect is much smaller than with the causal answer option, however. The decrease in correct answers again corresponded to an increase in choice of a mechanistic analysis.

To confirm our results, we performed an independent replication of our experiment in a later offering of the same data analysis course. In the replication (October 2013), 1762 students completed our experimental quiz question. The results of this replication were consistent with those in the original experiment (Tables 4.2, 4.3, 4.4). Differences in percentages for the causal and inferential answer choices were always of the same sign between the two courses, and the magnitudes of the differences were also similar (Table 4.2). While the sample size in this course is much smaller, the concordance of results and the maintenance of experimental procedures between courses align with a statistical definition of replicability that has been put forth (Patil, Peng, and Leek, 2016).

4.2.4 Discussion

We know that the way data is visualized can affect how well people derive information from graphs (Cleveland and McGill, 1985). The results of this experiment suggest that the way we write about a data analysis is also critical. By performing a randomized controlled trial, we have shown a clear effect of explanatory statements on perceptions of research results, and we have replicated this effect in a second experiment. The nature of our study design justifies the use of causal language to describe the precise effect of explanatory language on categorical perceptions of research findings, but it is important to keep in mind that these effects are specific to a certain population of learners and to our specific quiz question. In the remainder of this section, we discuss these limitations and avenues for further research.

One limitation of our study is the population used. We performed this randomized trial in a population of learners in a massive open online course (MOOC) as opposed to a representative sample of the general population. While we do not have demographic information on the learners in our trial, surveys of various MOOCs indicate that these learners are slightly more likely to be male, often have bachelor's degrees, and typically have some level of employment (Bayeck, 2016). Learners in these online courses report a variety of motivations for taking the courses, suggesting at least some lifestyle diversity.

A second limitation of our study was the choice to use smoking as the study example. A well-studied phenomenon in cognitive science, the availability heuristic, describes how people often unduly use readily available examples to guide their thinking. The causal link between smoking and lung cancer has been firmly established over time with the accumulation of studies, so although the wording of our quiz question does not describe a causal study, the availability heuristic likely nudges learners to think otherwise. Had we used a different example, the effect of the explanatory text would likely have been smaller.

The scope of our findings is also limited in that we have not investigated any strategies for combating causal misinterpretations arising from explanation. We recognize that is quite difficult to avoid any explanation when communicating scientific results because explanation is a key means of interpreting research findings. Interpretation is essential for combining different sources of information and advancing our understanding. In both academic

and mainstream scientific writing, there is a desire to put results into context, including hypothesized mechanistic explanations to enhance the narrative around a set of empirical results. Nearly every study includes this type of explanation in the discussion section. However, our results suggest that such efforts may actually cause a certain population of readers to be misled about the strength of the scientific evidence. The misinterpretation may be exacerbated by the phenomenon that readers are swayed to believe a statement when they are told scientists understand it (Sloman and Rabb, 2016). Because interpretation, and thus explanation, is an essential aspect of science communication, we should not aim to avoid explanation but to understand how certain characteristics of explanation help or hinder perception.

We hypothesize that it may be beneficial for readers' perceptions to follow up any explanations with warnings against interpreting results causally. Further research is needed to determine if this could counteract the effect of explanations on causal perceptions. It will also be important in further work to try to generalize the findings we present here in a population that is more representative of the general public and to dissect the nature of misinterpretation. In this study, we focused on categorical perception of knowledge, but it is also worthwhile to allow more flexibility in responses to understand how subjects' actions are affected.

The code and data used to perform this analysis are available at: https://github.com/leekgroup/explanatory_language.

4.3 Learner perceptions of plotting systems

One of the more commonly debated aspects of data science education within the R community is the plotting system used to introduce learners to statistical graphics. There has been some online and informal debate about the general strengths and weaknesses of the base and ggplot2 (Wickham, 2009) plotting systems within R for both research and teaching (Leek, 2016; Robinson, 2016). More recently there has been discussion of the relative merits of the two plotting systems in teaching the specific student population of beginner analysts (Robinson, 2016) and some investigation of learning outcomes when using base R and ggplot2 in the classroom (Stander and Dalla Valle, 2017). In the latter investigation, Stander et al provide instruction in both plotting systems in the classroom but do not compare the systems in terms of student learning outcomes.

We conducted a randomized experiment within the Coursera platform to better understand student perceptions of statistical graphics created in the two plotting systems. Students were randomly assigned to a peer-graded assignment in which they had to make two plots using only the base R graphics system or only the ggplot2 graphics system. The first of these plots was a simple scatterplot between two continuous variables. The second of these plots was more complex, asking for a grid of scatterplots resulting from stratification along two factor variables.

Students were asked to grade their classmates' submissions using a rubric. Results for the simple plot are shown in Figure 4.1, and results for the complex plot are shown in Figure 4.2. Generally, positive characteristics were more

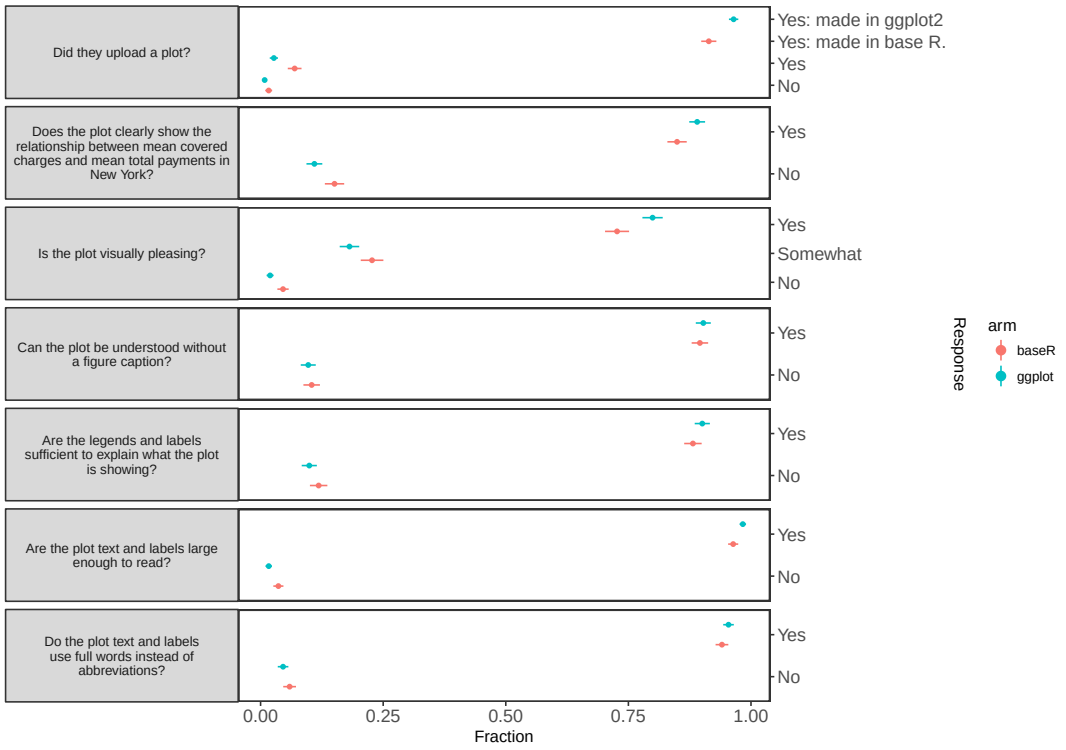


Figure 4.1: Peer review responses for the simple plot. Generally, reviews indicated that ggplot2 graphics were slightly more likely to contain desirable aesthetic qualities than base R graphics but that the two plotting systems were overall similar in these attributes. Plots made with ggplot2 were more likely to clearly show the intended relationship.

likely to be seen in figures made with ggplot2. In particular, students found ggplot2 figures to more clearly show the intended relationship, and this clarity seemed to be more pronounced for the complicated plot.

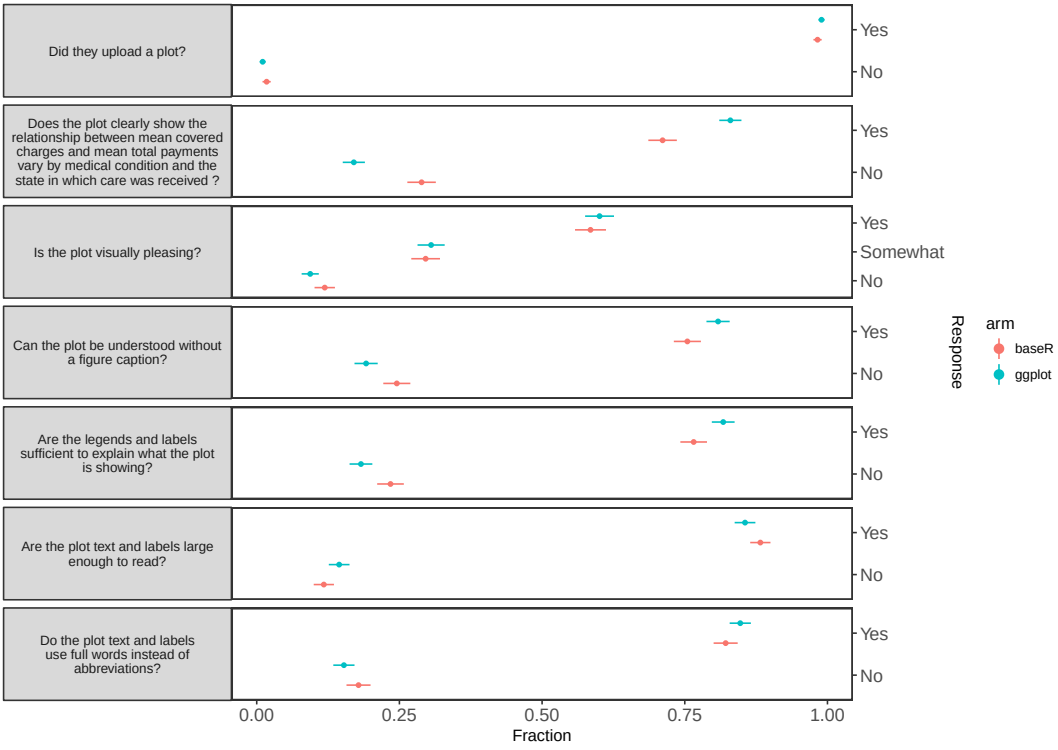


Figure 4.2: Peer review responses for the complex plot. As with the simple plot, reviews indicated that ggplot2 graphics were generally more likely to contain desirable aesthetic qualities than base R graphics but that the two plotting systems were overall similar in these attributes. For this more complex plot, the increased clarity in ggplot2 graphics was more pronounced.

References

Reporter, By Daily Mail. *How using Facebook could raise your risk of cancer*. http://www.dailymail.co.uk/health/article-1149207/How-using-Facebook-raise-risk-cancer.html.

PTI (2016). *Drinking too much tea can cause prostate cancer: study*. http://www.thehindu.com/sci-tech/health/medicine-and-research/drinking-too-much-tea-can-cause-prostate-cancer-study/article3547285.ece.

Jaslow, Ryan. *Eating lots of chocolate helps people stay thin, study finds*. http://www.cbsnews.com/news/eating-lots-of-chocolate-helps-people-stay-thin-study-finds/.

Sumner, Petroc, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, Fred Boy, and Christopher D Chambers (2014). "The association between exaggeration in health related science news and academic press releases: retrospective observational study". In: *BMJ* 349, g7015. DOI: 10.1136/bmj.g7015.

Cofield, Stacey S, Rachel V Corona, and David B Allison (2010). "Use of causal language in observational studies of obesity and nutrition". In: *Obes. Facts* 3, pp. 353–356. DOI: 10.1159/000322940.

Robinson, Daniel H, Joel R Levin, Greg D Thomas, Keenan A Pituch, and Sharon Vaughn (2007). "The Incidence of "Causal" Statements in Teaching-and-Learning Research Journals". In: *Am. Educ. Res. J.* 44, pp. 400–413. DOI: 10.3102/0002831207302174.

Lombrozo, Tania (2017). *The Dangers Of Hidden Jargon In Communicating Science*. https://www.npr.org/sections/13.7/2017/06/12/532554252/the-dangers-of-hidden-jargon-in-communicating-science.

Richardson, Sarah S, Cynthia R Daniels, Matthew W Gillman, Janet Golden, Rebecca Kukla, Christopher Kuzawa, and Janet Rich-Edwards (2014). "Society: Don't blame the mothers". In: *Nature* 512.7513, pp. 131–132. DOI: 10.1038/512131a.

Leek, Jeffery T and Roger D Peng (2015). "Statistics. What is the question?" In: *Science* 347, pp. 1314–1315. DOI: 10.1126/science.aaa6146.

Patil, Prasad, Roger D Peng, and Jeffrey Leek (2016). "A statistical definition for reproducibility and replicability".

Cleveland, W S and R McGill (1985). "Graphical perception and graphical methods for analyzing scientific data". In: *Science* 229, pp. 828–833. DOI: 10.1126/science.229.4716.828.

Bayeck, Rebecca Yvonne (2016). "Exploratory study of MOOC learners' demographics and motivation: The case of students involved in groups". In: *Open Praxis* 8, pp. 223–233. DOI: 10.5944/openpraxis.8.3.282.

Sloman, Steven A and Nathaniel Rabb (2016). "Your Understanding Is My Understanding: Evidence for a Community of Knowledge". In: *Psychol. Sci.* 27, pp. 1451–1460. DOI: 10.1177/0956797616662271.

Wickham, Hadley (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer.

Leek, Jeffrey (2016). *Why I don't use ggplot2*. https://simplystatistics.org/2016/02/11/why-i-dont-use-ggplot2/. Blog.

Robinson, David (2016). *Why I use ggplot2*. http://varianceexplained.org/r/why-I-use-ggplot2/. Blog.

Stander, Julian and Luciana Dalla Valle (2017). "On Enthusing Students About Big Data and Social Media Visualization and Analysis Using R, RStudio, and RMarkdown". In: *J. Stat. Educ.* 25.2, pp. 60–67. DOI: 10.1080/10691898.2017.1322474.

Chapter 5

Discussion and Conclusion

The work in this dissertation constitutes a push towards evidence-based decision making in three different areas. The intentions for using the phrase "evidence-based" are rooted in its contrast to what has been traditionally done in these areas.

In computational biology, methodological work frequently relies solely on results from simulations. Authors simulate biological assay data from models with a range of complexity and argue for the superiority of their methods based on their belief in the plausibility of their simulated data. Evidence-based decision making in computational biology, as presented in this dissertation, relies on real data as opposed to simulated data for method evaluation. Simulations can certainly be useful for verifying theoretical properties, but evaluations derived from real data provide more compelling justification for the use of methods in the wild.

For mass spectrometry-based metabolomics, we developed a preprocessing method that outperforms existing alternatives in terms of measurement

variability and statistical power on several real datasets. Similarly, for massively parallel reporter assay analysis, we show that our proposed methods have good error rate calibration and competitive statistical power. Our heavy use of real data evaluations in this area is noteworthy because many computational methods in the closely related sequencing literature rely strongly on simulated data for evaluation.

The act of studying the process of data analysis from a behavioral standpoint would seem natural to psychologists or cognitive scientists but is novel for statisticians. That statistics is rooted in mathematics and theory likely explains why the role of human behavior has been underappreciated. For this reason, data analysis practice that is informed by real data on behavior will be useful to the community. In this dissertation, we make inquiries about certain aspects of human behavior and judgment in learner populations. Through randomized experiments we learn about the effects of explanation in observational settings and about the differences in two popular statistical graphics systems. Further research is needed to understand the generalizability of these results, but for now we have some intuition regarding how we communicate scientific results and teach beginners certain skills.

CV

Date of birth: January 12, 1992

Location of birth: Philadelphia, PA

Leslie Myint

PhD candidate in Biostatistics

e: lmyint1@jhu.edu
w: lmyint.github.io

Education

Johns Hopkins Bloomberg School of Public Health

PhD candidate - Biostatistics

Expected graduation: May 2018

Johns Hopkins University

Bachelor of Science

May 2013

Majors: Biomedical Engineering, Applied Mathematics and Statistics
Minor: Computer Science

Research

Statistical Methods for High-Throughput Biology

June 2014 - present

JHSPH - Advisor: Dr. Kasper Daniel Hansen

Pre-processing methods for mass spectrometry data for metabolomics applications and statistical methods for analyzing massively parallel reporter assays

Evidence-Based Data Analysis

July 2015 - present

JHSPH - Advisors: Dr. Jeffrey Leek and Dr. Leah Jager

Conducted and analyzed randomized trials on the Coursera platform to understand data analyst behavior

Computational Biology Laboratory

September 2011 - May 2013

JHU - Advisor: Dr. Feilim Mac Gabhann

Studied peripheral arterial disease using computational models of VEGF distribution in mice and humans

Internship: Institute of Genetic Medicine

May - October 2012

JHU - Advisor: Dr. Steven Salzberg

Performed an in-depth comparison of two widely used sequence alignment programs: Bowtie2 and BWA

REU: Modeling and Simulation in Systems Biology

May - August 2011

Virginia Bioinformatics Institute - Advisors: Dr. Shernita Lee, Dr. Reinhard Laubenbacher

Worked with two other students to develop a computational model of iron metabolism in lung epithelial cells exposed to fungus

Summer Undergraduate Research Fellowship

May - July 2010

Fox Chase Cancer Center - Advisor: Dr. Warren Kruger

Studied Schizosaccharomyces pombe yeast genetics

Publications

Published

2. Kang, Joon Y., Amin H. Rabiei, **Leslie Myint**, and Maromi Nei. "Equivocal Significance of Post-Ictal Generalized EEG Suppression as a Marker of SUDEP Risk." Seizure: The Journal of the British Epilepsy Association. doi:10.1016/j.seizure.2017.03.017.
1. **Myint, Leslie**, Andre Kleensang, Liang Zhao, Thomas Hartung, and Kasper D. Hansen. 2017. "Joint Bounding of Peaks Across Samples Improves Differential Analysis in Mass Spectrometry-Based Metabolomics." Analytical Chemistry 89 (6): 3517–23. doi:10.1021/acs.analchem.6b04719.

Preprints

2. **Myint, Leslie**, Dimitrios G. Avramopoulos, Loyal A. Goff, and Kasper Hansen. 2017. "Linear Models Enable Powerful Differential Activity Analysis in Massively Parallel Reporter Assays." bioRxiv. doi:10.1101/196394.

1. **Myint, Leslie**, Jeffrey T. Leek, and Leah R. Jager. 2017. "Explanation Implies Causation?" bioRxiv. https://doi.org/10.1101/218784.

In press

1. Anne K. Monroe, **Leslie Myint**, Richard Rutstein, Stephen Boswell, Judith Aberg, Allison Agwu, Kelly Gebo, Richard Moore. 2017. "Factors Associated with Gaps in Medicaid Enrollment among People with HIV and the Effect of Gaps on Viral Suppression." Journal of Acquired Immunodeficiency Syndromes.

Presentations

Joint Preprocessing of Samples Improves Power in Differential Analysis for Mass Spectrometry-Based Metabolomics
Invited Talk: JHU Biophysics
December 2017

Shiny Applications for Teaching and Dungeons and Dragons
Invited Talk: Baltimore UseR Group
September 2017

A Method for Joint Processing of Mass Spectrometry-Based Metabolomics Data for Improved Differential Analysis
Poster: ENAR, Washington D.C.
March 2017

Software

yamss: Tools for the analysis of high-throughput metabolomics data. An R package released through the Bioconductor project.
https://www.bioconductor.org/packages/yamss

mpra: Tools for the analysis of data from massively parallel reporter assays. An R package released through the Bioconductor project.
https://www.bioconductor.org/packages/mpra

Teaching

Johns Hopkins Bloomberg School of Public Health

Instructor

- Statistical Thinking for Informed Decision Making (2 semesters)
  I developed this course as part of the Gordis Teaching Fellowship, a school-wide award that provides funds to design and teach an undergraduate class. A news article-motivated introduction to major biostatistical areas, including causal inference, survey sampling, and survival analysis.

Teaching Assistant

- Public Health Biostatistics (3 semesters)
- Introduction to R for Public Health Researchers (1 course)
- Statistical Methods in Public Health (3 quarters)
- Data Analysis Workshop (2 courses)
- Statistics for Genomics (1 quarter)
- Statistics for Laboratory Scientists (2 quarters)
- Summer Institute: Statistical Reasoning in Public Health (2 courses)

Tutor

- Statistical Methods in Public Health (2 quarters)
- Center for Talented Youth
  Mentored a high school CTY Cogito Research Award Recipient

Johns Hopkins University

Teaching Assistant

- Introduction to Java (1 semester)

Awards

Helen Abbey Award, JHSPH
May 2017
Excellence in teaching: website

Service

2017: Referee - Observational Studies

Other Experience

Johns Hopkins Biostatistics Center

July 2016 - August 2017

JHSPH - Advisor: Carol Thompson, MS

Consulting work for multiple groups within the Johns Hopkins Medical Institution

Siemens Competition

2016 - 2017

Served as a Stage I, II, and finalist judge to evaluate entries in Computer Science, Mathematics, Bioinformatics, Cell/Cancer Biology, and Genetics

---

Techincal Skills

Programming languages

- R
- Stata
- Python
- Java
- Matlab

Application development

- Shiny
- HTML
- CSS
- Javascript
- d3.js

Other

- Git
- RMarkdown
- Adobe Photoshop