Published in final edited form as:

J Proteome Res. 2016 December 02; 15(12): 4742–4746. doi:10.1021/acs.jproteome.6b00554.

Tag-Count Analysis of Large-Scale Proteomic Data

Owen E. Branson, Michael A. Freitas*

The Ohio State Biochemistry Graduate Program, The Ohio State University, Columbus, Ohio 43210, United States

Department of Molecular Virology, Immunology and Medical Genetics, The Ohio State University, Columbus, Ohio 43210, United States

Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio 43210, United States

Abstract

Label-free quantitative methods are advantageous in bottom-up (shotgun) proteomics because they are robust and can easily be applied to different workflows without additional cost. Both label-based and label-free approaches are routinely applied to discovery-based proteomics experiments and are widely accepted as semiquantitative. Label-free quantitation approaches are segregated into two distinct approaches: peak-abundance-based approaches and spectral counting (SpC). Peak abundance approaches like MaxLFQ, which is integrated into the MaxQuant environment, require precursor peak alignment that is computationally intensive and cannot be routinely applied to low-resolution data. Not limited by these constraints, SpC approaches simply use the number of peptide identifications corresponding to a given protein as a measurement of protein abundance. We show here that spectral counts from multidimensional proteomic data sets have a mean-dispersion relationship that can be modeled in edgeR. Furthermore, by simulating spectral counts, we show that this approach can routinely be applied to large-scale discovery proteomics data sets to determine differential protein expression.

Graphical Abstract

*Corresponding Author: Phone: (614) 688-8432. Fax: (614) 688-8675. freitas.5@osu.edu.

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.6b00554.

Example R script to generate simulated count data. (TXT)

Example R script to generate simulated count data. (PDF)

The authors declare no competing financial interest.

Keywords

mass spectrometry; proteomics; spectral counting; tag-count; edgeR; R; bioconductor; label-free

INTRODUCTION

Two separate but related challenges in discovery-based proteomics are the identification of a complete proteome inventory and the determination of differential expression (DE) between the proteomes corresponding to different biological conditions. Proteome composition is ultimately inferred through the identification of peptide-spectrum matches (PSMs). Protein quantitation by label-free approaches are segregated into two distinct approaches: peak-abundance-based and spectral counting (SpC). Peak abundance approaches like MaxLFQ,¹ which is integrated into the MaxQuant environment, require precursor peak alignment that is computationally intensive but yield less dispersion for low abundant species. SpC approaches simply use the number of peptide identifications corresponding to a given protein as a measurement of protein abundance. Numerous studies have compared both approaches, with mixed conclusions regarding which yields the best results.^{1–5} Both approaches are advantageous and are used in the analysis pipelines of commercial (Scaffold-Proteome Software, Proteome Software, Progenesis Qi - Nonlinear Dynamics/Waters, Proteome Discoverer - Thermo Fisher Scientific, etc.) and academic (MaxQuant,¹ OpenMS,⁶ QSpec/QPROT,^{7,8} etc.) software packages.

SpC-based analysis has the advantage that it is significantly less computationally intensive than peak abundance analysis because there is no need to detect and align features in the data. For this reason, SpC scales well with increasing number of samples. There are a multitude of algorithms that have been used to determine DE from spectral counts including but not limited to Protein Abundance Index (PAI/emPAI)^{9,10} Normalized Spectral Abundance Factors (NASF/uNASF),^{11,12} Normalized Spectral Index (SIN),¹³ Absolute Protein Expression Measurements (APEX),¹⁴ QSpec,⁷ and QuasiTel.¹⁵ While each of these

aforementioned approaches is valid, they tend to be tailored for use in specific proteomic pipelines.¹⁶

SpC data from shotgun proteomics are multivariate and often derived from few biological replicates. These characteristics, which are also valid for RNA sequencing data, prompted the adoption of edgeR to determine differential expression. EdgeR is an approach used to determine DE from multivariate RNA sequencing count data and is available via a Bioconductor package in the R statistical computing environment. EdgeR is a hypothesis test-based approach that leverages the link between sample mean and sample variance by modeling data with the negative binomial distribution. It was originally developed to determine differential expression from Serial Analysis of Gene Expression data sets.¹⁷ While there are different tag-count approaches used to analyze RNA sequencing data, that is, DESeq¹⁸ and baySeq,¹⁹ the use of edgeR is advantageous because SAGE data sets are similar to proteomics data sets in terms of depth, diversity, and count frequencies. As such, edgeR has seen application to SpC analysis by the proteomics community.^{20–25}

In edgeR, to adjust for unequal count frequencies across samples, normalization is completed by means of a weighted trimmed mean of M-values approach (TMM normalization).^{17,26–28} Once TMM normalization is completed, a “common” dispersion parameter is estimated using a quantile-adjusted conditional maximum likelihood approach.^{17,28} The common dispersion parameter may fit the data well if the SpC are similar for all proteins in a given sample or if the number of biological replicates in the data set is small ($n < 3$), and determining a tag-wise dispersion estimate, an estimate for every individual protein cannot be calculated. When more than three biological replicates are available for each condition, a tag-wise dispersion estimate is often most appropriate. To calculate the tagwise dispersion estimate, an empirical Bayesian approach is used.²⁸ Once the dispersion estimates are calculated, differential expression is determined with an “exact-like test”,¹⁷ which is similar to the conditional Fisher’s exact test but utilizes the negative binomial distribution.¹⁷ Finally, the family-wise error rate is controlled by computing a Benjamini–Hochberg FDR approximation²⁹ or a q value.^{30,31} Both the Benjamini–Hochberg FDR approximation and the q value were investigated in regards to identifying DE proteins.

EXPERIMENTAL METHODS

To illustrate the utility of edgeR for DE analysis of large-scale discovery proteomics the “Proteome-wide Benchmark Dataset” was downloaded from the ProteomeXchange Consortium (ID: PXD000279). This “two proteome analysis” consisted of six dual proteome samples composed of 60 μg of HeLa S3 lysate, where three samples were spiked with 10 μg *E. coli* K12 lysate and three samples were spiked with 30 μg lysate. Each sample was digested and separated into 24 fractions by off-gel isoelectric focusing, as described by Cox et al.¹ One raw file was missing from this analysis so the remaining 143 raw files were searched with MyriMatch (v 2.2.140)³² and peptide identifications were assembled to appropriate proteins using the peptide parsimony approach³³ integrated into IDPicker software (v 3.1.593).³⁴ The spectral counts were exported and analyzed with edgeR as described above.

RESULTS AND DISCUSSION

Following an edgeR statistical analysis it is imperative to evaluate how well the “two proteome analysis” data fit the model. The edgeR biological coefficient of variation (BCV), defined as the square root of the edgeR dispersion parameter, is plotted as a relationship of $\log_2$ counts per million (CPM), a normalized measure of proteomic SpC. The BCV plot indicates both the common and tagwise dispersion estimates (Figure 1A). Because the sample variance is defined as a quadratic function of the sample mean and both the edgeR common and tag-wise dispersion parameters (see edgeR vignette for details), it is expected that proteins represented by low CPM values would produce elevated dispersion estimates. Furthermore, it is expected that “two-proteome analysis” SpC data have unequal variance, where the sample variance is overdispersed specifically at higher CPM. The edgeR mean-variance relationship supports this notion and suggests that a negative binomial model is appropriately suited to model the overdispersed data (Figure 1B). Once it has been determined that the “two proteome analysis” SpC data can be appropriately modeled by edgeR, the “two proteome analysis” SpC data were used to evaluate how well edgeR was able to estimate the fold changes and identify statistically DE proteins. TMM normalization is used over library normalization or raw counts analysis because it is robust to technical and biological variability that effects the library size.²⁷ Thus it would reveal the intrinsic relative protein changes regardless of the absolute amount of protein loaded onto the mass spectrometer. Therefore, the anticipated HeLa lysate fold change was $[60 \mu\text{g HeLa}/(30 \mu\text{g } E. coli + 60 \mu\text{g HeLa}) \div 60 \mu\text{g HeLa}/(10 \mu\text{g } E. coli + 60 \mu\text{g HeLa})]$ or $\log_2(0.78) \approx -0.36$, and the anticipated *E. coli* lysate fold change was $[30 \mu\text{g } E. coli/(30 \mu\text{g } E. coli + 60 \mu\text{g HeLa}) \div 10 \mu\text{g } E. coli/(10 \mu\text{g } E. coli + 60 \mu\text{g HeLa})]$ or $\log_2(2.33) \approx 1.22$. Figure 2 shows the Kernel density plots of the $\log_2$ fold changes of these two proteomes and indicates two distributions with some overlap. The proteomic analysis identified a total of 6045 proteins of human or *E. coli* origin. Of these protein identifications, 1683 were unique to *E. coli* and 4362 were unique to human. Using the statistical cutoff of 0.05, as determined by the Benjamini–Hochberg FDR approximation, 992/1683 ($\approx 58.9\%$) *E. coli* proteins and 12/4362 ($\approx 0.3\%$) human proteins were identified as DE. It has been well accepted in the “omics” community that tag-count approaches can gain statistical power when utilizing a q value rather than the Benjamini–Hochberg FDR approximation to control for the family wise error rate.³⁵ By implementing a q value statistical cutoff of 0.05, as depicted in Figure 2, 1047/1683 ($\approx 62.2\%$) *E. coli* proteins and 239/4362 ($\approx 5.5\%$) human proteins were found as statistically DE. The q value approach allowed for an additional 55 protein identifications from the *E. coli* lysate and 184 identification from HeLa lysate to meet the statistical cutoff. The anticipated $\log_2$ fold change associated with HeLa proteins is -0.36 , where a value of 0 would indicate no fold change. So the incorporation of an additional 55 *E. coli* protein identifications came at the cost of 184 HeLa protein identifications, which may logically be considered false-positives. In total, 126/239 ($\approx 52.7\%$) of the statistically significant HeLa protein identifications have CPM values <7.0 (Figure 2, right). The calculation of $\log_2$ fold changes for proteins with low CPM values are highly questionable, and the implementation of effect size filtering may be beneficial for certain experiments.²⁵

The utility of SpC for large-scale label-free analysis was first demonstrated by the Zhang et al. in the proteogenomic analysis of colon and rectal tumors from The Cancer Genome Atlas (TCGA).³⁶ To evaluate the efficiency of the edgeR algorithm for SpC DE analysis of large-scale proteomic data sets, a series of SpC data sets were simulated. The median SpC was calculated from the 4362 HeLa proteins from the six biological replicates in the “two-proteome analysis”. Using these values to represent λ in the Poisson distribution, 4362 SpC values were generated. This process was repeated to produce non-DE data sets consisting of 6, 10, 20, 30, 40, 50, 100, 200, and 500 simulated biological replicates.

In a similar fashion, a set of SpCs associated with DE proteins was generated. In total, 436 values of λ were randomly chosen with replacement from the list of 4362. For half of the simulated biological replicates, this value was directly used to represent λ . For the other half of the biological replicates, the chosen value was multiplied by a $\log_2$ fold change that was selected from a uniform distribution with a range [1.5, 4.0]. This value was rounded down to the nearest whole number and used as λ . Finally, the first 218 samples generated were chosen to be up-regulated, and the remaining 218 samples were chosen to be down-regulated. Overall, this produced a total of 4798 simulated proteins where 10% of these are expected to be DE.

Each of the eight data sets was analyzed five times with the edgeR script (Supporting Information) and the script runtime tracked (1.7 GHz Intel Core i7 processor, 8 GB 1600 MHz DDR3 memory, R (v 3.3.0) 64 bit) (Figure 3A). The average execution time for the six biological replicate set and 500 biological replicate data set set was 2.1 and 99.6 s, respectively, illustrating the capability to perform efficient DE analysis on large data sets. The mean-variance relationship from the simulated six biological replicate data set (Figure 3B) is nearly linear, as one could expect from data that fit a Poisson distribution. This observation supports the notion that SpC data derived from discovery proteomics experiments are indeed overdispersed, as illustrated in Figures 1B and 3C. To support the notion large-scale proteomic data can easily be efficiently analyzed, edgeR was used to analyze the data from 95 colon and rectal carcinomas.^{36–38} The pairwise analysis of this 7212 protein $\times$ 95 biological replicate data matrix was analyzed with edgeR in a mere 16.6 s (Figure 3D).

CONCLUSIONS

In conclusion, this technical note highlights the efficient application of negative binomial based tag-count DE approaches for the analysis of large-scale proteomics. While database search can still be a significant bottleneck in the informatics pipeline, SpC quantitation is an efficient approach to perform protein DE analysis across a large number of samples. Not hindered by computationally intensive precursor peak alignment and the prerequisite for high-resolution data, edgeR DE analysis of SpCs is robust, fast, and easily implemented for both large and small routine analyses in discovery proteomics. Utilizing similar methods for transcriptomics and proteomics also has distinct advantages for streamlining the development and application of integrated tools for proteogenomic analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was supported by NIH awards CA016058, OD018056 and support from the Ohio State University.

ABBREVIATIONS

SpC	spectral counts, spectral counting
DE	differential expression, differentially expressed
PSM(s)	peptide-spectrum match(es)
TMM	trimmed mean of M-values
BCV	biological coefficient of variation
CPM	counts per million

REFERENCES

- (1). Cox J; Hein MY; Luber CA; Paron I; Nagaraj N; Mann M Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* 2014, 13 (9), 2513–26. [PubMed: 24942700]
- (2). Old WM; Meyer-Arendt K; Aveline-Wolf L; Pierce KG; Mendoza A; Sevinsky JR; Resing KA; Ahn NG Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics* 2005, 4 (10), 1487–502. [PubMed: 15979981]
- (3). Zybailov B; Coleman MK; Florens L; Washburn MP Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. *Anal. Chem* 2005, 77 (19), 6218–24. [PubMed: 16194081]
- (4). Milac TI; Randolph TW; Wang P Analyzing LC-MS/MS data by spectral count and ion abundance: two case studies. *Stat Interface* 2012, 5 (1), 75–87. [PubMed: 24163717]
- (5). Weisser H; Nahnsen S; Grossmann J; Nilse L; Quandt A; Brauer H; Sturm M; Kenar E; Kohlbacher O; Aebersold R; Malmstrom L An automated pipeline for high-throughput label-free quantitative proteomics. *J. Proteome Res* 2013, 12 (4), 1628–44. [PubMed: 23391308]
- (6). Rost HL; Sachsenberg T; Aiche S; Bielow C; Weisser H; Aicheler F; Andreotti S; Ehrlich HC; Gutenbrunner P; Kenar E; Liang X; Nahnsen S; Nilse L; Pfeuffer J; Rosenberger G; Rurik M; Schmitt U; Veit J; Walzer M; Wojnar D; Wolski WE; Schilling O; Choudhary JS; Malmstrom L; Aebersold R; Reinert K; Kohlbacher O OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods* 2016, 13 (9), 741–8. [PubMed: 27575624]
- (7). Choi H; Fermin D; Nesvizhskii AI Significance analysis of spectral count data in label-free shotgun proteomics. *Mol. Cell. Proteomics* 2008, 7 (12), 2373–85. [PubMed: 18644780]
- (8). Choi H; Kim S; Fermin D; Tsou CC; Nesvizhskii AI QPROT: Statistical method for testing differential expression using protein-level intensity data in label-free quantitative proteomics. *J. Proteomics* 2015, 129, 121–6. [PubMed: 26254008]
- (9). Rappsilber J; Ryder U; Lamond AI; Mann M Large-scale proteomic analysis of the human spliceosome. *Genome Res* 2002, 12 (8), 1231–45. [PubMed: 12176931]
- (10). Ishihama Y; Oda Y; Tabata T; Sato T; Nagasu T; Rappsilber J; Mann M Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* 2005, 4 (9), 1265–72. [PubMed: 15958392]

- (11). Zybaylov B; Mosley AL; Sardu ME; Coleman MK; Florens L; Washburn MP Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J. Proteome Res* 2006, 5 (9), 2339–47. [PubMed: 16944946]
- (12). Zhang Y; Wen Z; Washburn MP; Florens L Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. *Anal. Chem* 2010, 82 (6), 2272–81. [PubMed: 20166708]
- (13). Griffin NM; Yu J; Long F; Oh P; Shore S; Li Y; Koziol JA; Schnitzer JE Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat. Biotechnol* 2010, 28 (1), 83–9. [PubMed: 20010810]
- (14). Lu P; Vogel C; Wang R; Yao X; Marcotte EM Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol* 2007, 25 (1), 117–24. [PubMed: 17187058]
- (15). Li M; Gray W; Zhang H; Chung CH; Billheimer D; Yarbrough WG; Liebler DC; Shyr Y; Slebos RJ Comparative shotgun proteomics using spectral count data and quasi-likelihood modeling. *J. Proteome Res* 2010, 9 (8), 4295–305. [PubMed: 20586475]
- (16). Langley SR; Mayr M Comparative analysis of statistical methods used for detecting differential expression in label-free mass spectrometry proteomics. *J. Proteomics* 2015, 129, 83–92. [PubMed: 26193490]
- (17). Robinson MD; Smyth GK Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* 2007, 9 (2), 321–32. [PubMed: 17728317]
- (18). Anders S; Huber W Differential expression analysis for sequence count data. *Genome Biol* 2010, 11 (10), R106. [PubMed: 20979621]
- (19). Hardcastle TJ; Kelly KA baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinf* 2010, 11, 422.
- (20). Johnson EK; Zhang L; Adams ME; Phillips A; Freitas MA; Froehner SC; Green-Church KB; Montanaro F Proteomic analysis reveals new cardiac-specific dystrophin-associated proteins. *PLoS One* 2012, 7 (8), e43515. [PubMed: 22937058]
- (21). Shapiro JP; Biswas S; Merchant AS; Satoskar A; Taslim C; Lin S; Rovin BH; Sen CK; Roy S; Freitas MA A quantitative proteomic workflow for characterization of frozen clinical biopsies: laser capture microdissection coupled with label-free mass spectrometry. *J. Proteomics* 2012, 77, 433–40. [PubMed: 23022584]
- (22). Fei SS; Wilmarth PA; Hitzemann RJ; McWeeney SK; Belknap JK; David LL Protein database and quantitative analysis considerations when integrating genetics and proteomics to compare mouse strains. *J. Proteome Res* 2011, 10 (7), 2905–12. [PubMed: 21553863]
- (23). Harshman SW; Canella A; Ciarlariello PD; Rocci A; Agarwal K; Smith EM; Talabere T; Efebera YA; Hofmeister CC; Benson DM Jr.; Paulaitis ME; Freitas MA; Pichiorri F Characterization of multiple myeloma vesicles by label-free relative quantitation. *Proteomics* 2013, 13 (20), 3013–29. [PubMed: 23983189]
- (24). Branson OE; Freitas MA A multi-model statistical approach for proteomic spectral count quantitation. *J. Proteomics* 2016, 144, 23–32. [PubMed: 27260494]
- (25). Gregori J; Villarreal L; Sanchez A; Baselga J; Villanueva J An effect size filter improves the reproducibility in spectral counting-based comparative proteomics. *J. Proteomics* 2013, 95, 55–65. [PubMed: 23770383]
- (26). Bullard JH; Purdom E; Hansen KD; Dudoit S Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinf* 2010, 11, 94.
- (27). Robinson MD; Oshlack A A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 2010, 11 (3), R25. [PubMed: 20196867]
- (28). Robinson MD; Smyth GK Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* 2007, 23 (21), 2881–7. [PubMed: 17881408]
- (29). Benjamini Y; Hochberg Y Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc* 1995, 57 (1), 289–300.
- (30). Dabney A; Storey JD; Warnes GR qvalue: Q-value estimation for false discovery rate control; R package, version 1.24.20, 2010.

- (31). Storey JD The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value. *Annals of Statistics* 2003, 31 (6), 2013–2035.
- (32). Tabb DL; Fernando CG; Chambers MC MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res* 2007, 6 (2), 654–61. [PubMed: 17269722]
- (33). Zhang B; Chambers MC; Tabb DL Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J. Proteome Res* 2007, 6 (9), 3549–57. [PubMed: 17676885]
- (34). Ma ZQ; Dasari S; Chambers MC; Litton MD; Sobecki SM; Zimmerman LJ; Halvey PJ; Schilling B; Drake PM; Gibson BW; Tabb DL IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J. Proteome Res* 2009, 8 (8), 3872–81. [PubMed: 19522537]
- (35). Storey JD A direct approach to false discovery rates. *Journal of the Royal Statistical Society* 2002, 64 (3), 479–498.
- (36). Zhang B; Wang J; Wang X; Zhu J; Liu Q; Shi Z; Chambers MC; Zimmerman LJ; Shaddox KF; Kim S; Davies SR; Wang S; Wang P; Kinsinger CR; Rivers RC; Rodriguez H; Townsend RR; Ellis MJ; Carr SA; Tabb DL; Coffey RJ; Slebos RJ; Liebler DC; Nci C; et al. Proteogenomic characterization of human colon and rectal cancer. *Nature* 2014, 513 (7518), 382–7. [PubMed: 25043054]
- (37). Slebos RJ; Wang X; Wang X; Zhang B; Tabb DL; Liebler DC Corrigendum: Proteomic analysis of colon and rectal carcinoma using standard and customized databases. *Sci. Data* 2015, 2, 150037. [PubMed: 26217491]
- (38). Slebos RJ; Wang X; Wang X; Zhang B; Tabb DL; Liebler DC Proteomic analysis of colon and rectal carcinoma using standard and customized databases. *Sci. Data* 2015, 2, 150022. [PubMed: 26110064]

Figure 1. edgeR diagnostic plots to evaluate the model fit of the “Two Proteome Analysis”. (A) Plot depicting the common and tag-wise biological coefficients of variation as they relate to the average Log₂ CPM values as calculated by edgeR. (B) edgeR mean-variance plot from the “Two Proteome Analysis”. Note the deviation from the linear line at high spectral counts indicating overdispersion of the data set.

Figure 2.

Evaluation of spectral count data from the “two proteome analysis”. (Main) Kernel density plot of HeLa (blue) and *E. coli* (green) $\log_2$ -fold-changes overlaid with a scatter plot of their respective $\log_2$ fold changes and corresponding $\log_2$ CPM values. Note two distinct distributions for HeLa and *E. coli* proteins. (Top) Cumulative distribution line plot tracking total *E. coli* proteins, differentially expressed *E. coli* proteins, and differentially expressed HeLa proteins as a function of $\log_2$ fold change. (Right) Cumulative distribution line plot tracking total *E. coli* proteins, differentially expressed *E. coli* proteins, and differentially expressed HeLa proteins as a function of $\log_2$ CPM.

Figure 3.

(A) Plot of edgeR runtimes for simulated data sets. (B) Magnified plot of edgeR runtimes for simulated data sets. (C) edgeR Mean-Variance plot from simulated data set. There is a strong linear relationship between the pooled gene variance and mean spectral counts. The red “X” indicates the running average of 50 data points. (D) edgeR Mean-Variance plot from 95 samples of colon/rectal tumors. Note the deviation from linearity at high spectral counts indicating overdispersion of the data set.