

Integration of transcriptomics and metabonomics: improving diagnostics, biomarker identification and phenotyping in ulcerative colitis

Jacob Tveiten Bjerrum · Mattias Rantalainen · Yulan Wang · Jørgen Olsen · Ole Haagen Nielsen

Received: 6 May 2013 / Accepted: 10 August 2013 / Published online: 21 August 2013
© Springer Science+Business Media New York 2013

Abstract A systems biology approach to multi-faceted diseases has provided an opportunity to establish a holistic understanding of the processes at play. Thus, the current study merges transcriptomics and metabonomics data in order to improve diagnostics, biomarker identification and to explore the possibilities of a molecular phenotyping of ulcerative colitis (UC) patients. Biopsies were obtained from the descending colon of 43 UC patients (22 active UC and 21 quiescent UC) and 15 controls. Genome-wide gene expression analyses were performed using Affymetrix GeneChip Human Genome U133 Plus 2.0. Metabolic

profiles were generated using ^1H Nuclear magnetic resonance spectroscopy (Bruker 600 MHz, Bruker BioSpin, Rheinstetten, Germany). Data were analyzed with the use of orthogonal-projection to latent structure-discriminant analysis and a multivariate logistic regression model fitted by lasso. Prediction performance was evaluated using nested Monte Carlo cross-validation. The prediction performance of the merged data sets and that of relative small (<20 variables) multivariate biomarker panels suggest that it is possible to discriminate between active UC, quiescent UC, and controls; between patients with or without steroid dependency, as well as between early or late disease onset. Consequently, this study demonstrates that the novel approach of integrating metabonomics and transcriptomics combines the better of the two worlds, and provides us with clinical applicable candidate biomarker panels. These combined panels improve diagnostics and more importantly also the molecular phenotyping in UC and provide insight into the pathophysiological processes at play, making optimized and personalized medication a possibility.

Jacob Tveiten Bjerrum and Mattias Rantalainen have contributed equally to this study.

Electronic supplementary material The online version of this article (doi:10.1007/s11306-013-0580-3) contains supplementary material, which is available to authorized users.

J. T. Bjerrum · J. Olsen
Department of Cellular and Molecular Medicine, The Panum Institute, University of Copenhagen, Copenhagen, Denmark

J. T. Bjerrum (✉) · O. H. Nielsen
Department of Gastroenterology, Medical Section, Herlev Hospital, University of Copenhagen, Copenhagen, Denmark
e-mail: bjerrum.jacob@gmail.com

M. Rantalainen
Department of Statistics, Oxford University, Oxford, UK

M. Rantalainen · Y. Wang
State Key Laboratory of Magnetic Resonance and Atomic and Molecular Physics, Wuhan Centre for Magnetic Resonance, Wuhan Institute of Physics and Mathematics, The Chinese Academy of Sciences, Wuhan, People's Republic of China

Y. Wang
Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, Hangzhou, People's Republic of China

Keywords Colonic biopsy · Gene expression profiles · Inflammatory bowel disease · Metabolomics · Metabolic profiles · NMR spectroscopy

Abbreviations

AUC	Area under the curve
CD	Crohn's disease
CPMG	Carr–Purcell–Meiboom–Gill
IBD	Inflammatory bowel disease
K-OPLS	Kernel-based orthogonal partial least squares regression
NMR	Nuclear magnetic resonance
NOE	Nuclear Overhauser effect

OPLS-DA	Orthogonal partial least squares-discriminant analysis
PCA	Principal component analysis
RD	Recycle delay
ROC	Receiver operating characteristic
Sd	Steroid dependence
Si	Steroid independence
UC	Ulcerative colitis

1 Introduction

Ulcerative colitis (UC) (Ordás et al. 2012) and Crohn's disease (CD) (Baumgart and Sandborn 2012) are chronic, relapsing disorders representing the two main entities of inflammatory bowel disease (IBD) with an increasing incidence and prevalence globally (Molodecky et al. 2012). It is believed that these diseases arise in genetically susceptible individuals as a result of an environmental and microbial induced aberrant immunological response. However, the actual pathogenesis of UC and CD remain enigmatic, and it is clear that a range of subphenotypes exist. Consequently, the diagnostic procedure relies on a multidisciplinary approach (endoscopy, radiology, and histopathology), which is hampered by its time-consuming procedures and lack of differential power. Unfortunately, this approach gives rise to a considerable number of patients (~10%) with the diagnosis of "IBD unclassified" (Geboes et al. 2008), which has serious clinical consequences, as successful clinical management and the patients' quality of life depends on early and correct diagnosis.

With the intention to improve our understanding of the pathophysiology and refine the diagnostics, innovative approaches such as DNA microarray-based genome-wide gene expression analysis (transcriptomics) (Bjerrum et al. 2010a; von Stein et al. 2008; Costello et al. 2005; Planell et al. 2012; Wu et al. 2007; Olsen et al. 2009) and ^1H nuclear magnetic resonance (^1H NMR) spectroscopy-based metabolic profiling (metabonomics) (Bezabeh et al. 2001; Balasubramanian et al. 2009; Bjerrum et al. 2008, 2010b) have been used to characterise colonic mucosal pinch biopsies from IBD patients and combined with multivariate statistical analysis. These studies demonstrate that both transcriptomics and metabonomics hold promising differential power in terms of discrimination of active IBD and controls, and between active UC and CD as well. Interestingly, some of these studies (Bjerrum et al. 2010a; Olsen et al. 2009; Planell et al. 2012) are also at the transcriptomics level able to differentiate between active UC, inactive UC, and controls, indicating the presence of a possible continuous inflammatory state in quiescent UC—a continuous inflammatory state that so far has not been

established in any other types of IBD, irritable bowel syndrome (IBS) or infectious colitis.

Besides active and quiescent stages of the disease a range of phenotypes such as steroid dependence (Sd) or independence (Si), early or late disease onset, and duration of disease also clinically characterize UC. A molecular portrait of these clinical phenotypes would greatly advance the understanding of UC and potentially lead to optimized and personalized treatment regimes.

Traditionally, the different 'omics' methods (e.g. transcriptomics and metabonomics) have been applied independently with diagnostic purposes and in order to derive fundamental insight into the molecular mechanisms of human diseases. However, a range of recent studies (Borgan et al. 2010; Astrakas et al. 2011; Chen et al. 2012; Tzika et al. 2007) have approached a systems biology description by merging these parallel 'omics' platforms as integration is expected to be highly synergistic (Fiocchi 2012). Thus, Borgan et al. (2010) successfully combined transcriptomics and metabonomics data from breast carcinoma samples, and demonstrated a refinement of the subclassification of breast cancers due to the identification of different but diagnostically synergistic biomarkers found at the transcriptomics and metabonomics level, respectively. Similarly, Astrakas et al. (2011) combined metabolite and gene expression profiles and revealed a more accurate discernment of brain tumor categories and a better prediction of patient survival than either method alone. This novel approach that combines biomarker detection at different 'omics' levels consequently seems to be a promising tool for improvement of diagnostics, phenotyping, and our understanding of the pathophysiology of complex diseases. Metabolites are the final product of cellular biological processes, which are continuously influenced by genetic and environmental changes. The metabolic profile of any given biological sample is consequently expected to closely reflect the phenotype making the integration of transcriptomics and metabonomics an appealing approach.

With this in mind, the current study aimed at merging microarray-based transcriptomics data and NMR spectroscopy-based metabonomics data generated from colonic mucosal pinch biopsies from patients with active UC, inactive UC, and controls in order to improve diagnostics, biomarker identification and to explore the possibilities of a molecular phenotyping of UC patients.

2 Materials and methods

2.1 Patient population

Sixty-three patients, subjected to flexible colonoscopy at the Department of Gastroenterology, Medical Section,

Herlev Hospital, Denmark were consecutively enrolled into the study: 27 UC patients with active disease, 21 quiescent UC patients, and 15 controls (i.e. patients undergoing colonoscopy due to gastrointestinal symptoms but where all clinical and paraclinical investigations subsequently turned out normal). All eligible UC patients ($n = 48$) had their diagnosis verified by well-established criteria (Nikolaus and Schreiber 2007) and were at the time of enrolment graded in accordance with the Mayo-score (Schroeder et al. 1987) (a score of 0–1: inactive UC, 2–4: mild UC, 5–8: moderate UC, and 9–12: severe UC). The endoscopic assessment of activity was confirmed by histological examination and any discrepancy between the observations led to an exclusion of the patient. Hence, none of the included controls or patients with quiescent UC had any signs of inflammation. Exclusion criteria were age above 80 or below 18 years, clinical evidence of infection, recent (within 14 days) use of antibiotics or probiotics, pregnancy, and severe mental illness.

In order to replicate normal circumstances and make the study as clinically relevant as possible no exclusion criteria were imposed in relation to diet, alcohol intake, medication, abdominal surgery, or exercise as the objective was to develop a robust, generic classifier that do not depend on such factors.

All patients participated in this study in confirmation with the principals outlined in the Declaration of Helsinki and with the approval of the Scientific Ethics Committee of the Copenhagen Capital Region (H-KA-20060164). Informed written consent was obtained from each participating person.

2.2 Patient information

Five patients originally included as active UC had to be excluded due to the finding of microgranulomas on histopathological examination and Crohn's-like endoscopic lesions, cf. Table 1 for clinical details.

2.3 Sample collection

Three adjacent (no more than 5 mm apart) colonic mucosal pinch biopsies (approximately 10–20 mg each) were obtained endoscopically from the left side of the colon in each patient using routine endoscopic forceps. The left side was explicitly preferred to avoid any intersegmental variation in gene expression, and because this is the most frequent area of inflammation in UC patients. The first biopsy was used for histopathological evaluation conducted in an unblinded fashion by staff pathologists in accordance with well-established criteria (Bentley et al. 2002). The second biopsy was immediately stabilized in RNA-Later (Ambion, Austin, TX, USA) in order to minimize RNA degradation, and after 48 h of storage in RNA-Later at 4 °C the biopsy was kept at –80 °C until total RNA extraction was initiated. The third biopsy was snap-frozen in liquid nitrogen and stored at –80 °C for subsequent NMR spectroscopy (Beckonert et al. 2010).

2.4 RNA extraction, hybridization, detection, and quantification of signals

RNA was extracted with the NucleoSpin® RNA/Protein mini kit (Macherey–Nagel, Düren, Germany) in

Table 1 Clinical details

	Inactive $n = 21$	UC Active $n = 22$	UC Controls $n = 15$
Gender (male/female)	8/13	8/14	4/11
Age, years (mean, range)	53 (27–78)	40 (18–76)	43 (19–65)
Age at diagnosis (<25 years/>25 years)	6/15	6/16	–
Years with disease (<10 years/>10 years)	12/9	16/6	–
Mayo score (mean, range)	0.2 (0–1)	6 (2–9)	–
Smoker/non-smoker	1/20	3/19	2/13
Steroids: Si/Sd/unknown	10/0/11	9/8/5	–
Daily medication			
Systemic mesalazine (1.6–3.2 g)	18	19	–
Topical mesalazine (1,000 mg)	6	9	–
Systemic glucocorticoids (75 mg)	2 ^a	4	–
Topical glucocorticoids (100 mg)	1	2	–
Azathioprine (100–150 mg)	0	6	–
Infliximab (5 mg/kg/infusion)	0	1	–
None	2	2	15

Steroid dependency is defined as the need for re-introduction of systemic glucocorticoids during tapering or within 30 days of drug discontinuation to maintain symptoms control. None of the patients were steroid refractory

UC ulcerative colitis, Si steroid independence, Sd steroid dependence

^a Tapering dose of 5 mg/day

accordance with the manufactures protocol. Integrity (RNA integrity number above seven was accepted) and purity were verified with an Agilent Bioanalyzer (Palo Alto, CA, USA). In accordance with the Affymetrix protocol and the one-cycle eukaryotic target labelling assay biotin-labelled cRNA was produced. The cRNA was fragmented and a hybridization mix was prepared, which included the fragmented target, probe array controls, bovine serum albumin, and herring sperm DNA. In this experiment, the Affymetrix GeneChip Human Genome U133 Plus 2.0 was applied. The hybridized probe array was subsequently stained with fluorescent protein streptavidin-phycoerythrin (Wang et al. 2007a) and scanned with a GeneArray scanner at the excitation wavelength of 488 nm. The amount of light emitted at 570 nm was proportional to the bound target at each location on the probe array. Data were stored as image files for further analysis. These procedures were all carried out at the Microarray Core Facility, Rigshospitalet, Copenhagen, Denmark, and all data from this study has been submitted to Gene Expression Omnibus.

2.5 ^1H NMR spectroscopy

The frozen biopsies were thawed and packed into 4 mm in diameter zirconia rotors with deuterium water for field lock. ^1H NMR spectra were acquired on a Bruker DRX-600 spectrometer (Bruker, Rheinstetten, Germany) and a triple-resonance-high-resolution-MAS probe with a magic-angle gradient. Samples were spun at 5 kHz at the magic angle (54.7°) and regulated at 283 K to minimize temperature induced biochemical degradation. A total of 15 min was allowed for the temperature to reach equilibration for each sample before a spectrum was acquired. All samples were carefully treated in the same way to avoid any procedure-based variations, and previous investigations on liver tissue have indicated that only minor changes are observed in the NMR spectra of snap-frozen samples and if samples are kept at low temperatures for short periods of time prior to the acquisition of spectra (Waters et al. 2000).

The 90° pulse length ($\sim 10.0 \mu\text{s}$) was set for all the samples. A total of 128 transients were collected into 16 k data points for each tissue spectrum with a spectral width of 20 ppm, and a recycle delay (RD) of 2.0 s. Two ^1H NMR spectra were acquired: (1) A standard one-dimensional NMR spectrum was acquired using the first increment of the nuclear Overhauser effect (NOE) spectroscopy pulse sequence [RD-90- t_1 -90- t_m -90-acq] with irradiation of water peak during RD and t_m to achieve water presaturation (Nicholson et al. 1995), (2) A spin-spin relaxation edited NMR spectrum was acquired using Carr-Purcell-Meiboom-Gill (CPMG) pulse sequence [RD-90-(T-180-T)n-acq] (Meiboom and Gill 1958). For the CPMG experiment, spin-spin relaxation delays, $2n\tau$, of 200 ms

was used. For assignment and verification purposes, two-dimensional ^1H - ^1H correlation spectroscopy (Hurd 1990) and total correlation spectroscopy (Bax and Davis 1985) NMR spectra were also acquired for selected intestinal samples as previously described in detail (Wang et al. 2007b).

2.6 Data analysis

All data analyses were carried out in R statistical environment (www.r-project.org) (RDC Team 2012) unless otherwise indicated.

2.6.1 Transcriptomics data

The transcriptomics data was pre-processed by the robust multiarray analysis procedure (Irizarry et al. 2003) with quantile normalization and background correction implemented in the *affy* package for the R statistical environment (Gentleman et al. 2004). A single log₂ scale expression measure for each probe set was attained from the low-level data files (CEL files). To reduce influence of atypical observations and extreme values, each variable was inverse normal transformed, i.e. mapped to the quantiles of the standard Gaussian distribution (mean = 0 and SD = 1).

2.6.2 Metabonomics data

Free induction decays were multiplied by an exponential function equivalent to a 0.3 Hz line-broadening factor prior to Fourier transformation. Spectra were manually corrected for phase and baseline distortions using XWINNMR 3.5 (Bruker) and referenced to the chemical shift of the anomeric proton of α -glucose at δ 5.22. The spectra over the range δ 0.5–9.5 were digitized using a Matlab script. The region δ 4.79–5.06 was removed to avoid the effects of imperfect water suppression. Regions δ 1.07–1.21, δ 3.37–3.57, δ 3.62–3.67, and δ 3.84–3.91 were also removed due to the presence of peaks from ethanol and propylene glycol, which were introduced during sampling. Regions across the spectra, where peaks were located, were manually selected (see Supplementary Material). The selection of spectral regions was comprehensive, also including low intensity spectral features, in order to avoid exclusion of relevant features prior to statistical analysis. Within each such region representing a peak/spectral feature, the local maximum intensity for each spectrum was calculated and extracted to represent the peak intensity in downstream analyses, thus allowing for local misalignment within each peak region. In cases no local maximum was present, the mean of the spectrum (i.e. baseline) of the region was used instead of the local maximum. The resulting data matrix of size $N \times p$ for N observations and

p peaks was subsequently normalized using the probabilistic quotient normalization method (Dieterle et al. 2006). To reduce influence of atypical observations and extreme values, each variable was inverse normal transformed, i.e. mapped to the quantiles of the standard Gaussian distribution (mean = 0 and SD = 1) prior to further analyses.

2.6.3 Quality control and outlier removal

To determine if there were any strong outliers present in the data, principal component analysis (PCA) models were calculated from each individual data set. Observations were removed from further analysis if they were outside the 95 % Hotelling T^2 confidence bound, leading to one active UC and two inactive UC observations being excluded. Thus, 19 patients with quiescent disease, 21 patients with active disease, and 15 controls were included in the final analyses.

2.6.4 Predictive modelling of full data sets using OPLS-DA

In the predictive modelling (classification) of the full data sets orthogonal partial least squares-discriminant analysis (OPLS-DA) models (Trygg and Wold 2002; Bylesjö et al. 2006) were fitted using the kernel-based OPLS algorithm (Rantalainen et al. 2007; Bylesjö et al. 2008) with a linear kernel, leading to identical prediction results as to the conventional linear OPLS model, but with computational benefits. The OPLS-DA models were fitted with one single predictive component and zero Y-orthogonal components, which in the case of no Y-orthogonal components is equivalent to a conventional PLS-DA model with one component. To estimate the classification performance, the class-balanced Monte Carlo cross-validation was applied, where 75 % of the data was used as training set in each cross-validation round. To minimize stochastic effects due to the cross-validation splits and ensure reliable prediction performance estimates, 100 cross-validation rounds were used. Prediction performance was estimated from the cross-validation results in the form of receiver operating characteristic (ROC) curves, which reveal information about the trade-off between the true-positive rate and the false-positive rate in the classifier. The area under the ROC curve (AUC) provides a summarizing metric of the prediction performance of the model, which is used for comparison of prediction performances between different models and data sets.

2.6.5 Predictive modelling and variable selection using logistic regression fitted by lasso

To identify candidate biomarker panels, a variable selection problem, and at the same time evaluate prediction

performance, a logistic regression model fitted by lasso was used (Friedman et al. 2010) as implemented in the *glmnet* package for R. The lasso solves the l_1 penalized regression problem, effectively shrinking some of the coefficients so that they become exactly zero, i.e. performing variable selection. To minimize the risk of over-fitting in respect to estimation of prediction performance, model optimization was performed in a nested cross-validation procedure. In the nested cross-validation, the ‘outer’ cross-validation is used to estimate prediction performance and the ‘inner’ cross-validation is used to optimize model parameters based on only the training set from the ‘outer’ cross-validation. In the case of the logistic regression fitted by lasso we needed to optimize the value for the λ parameter (controlling the amount of shrinkage), which was done empirically based on AUC from the ‘inner’ cross-validation procedure. To ensure that minimal sets of variables are selected, we added a penalty on the AUC estimates for each evaluated λ value, and selected λ as the value maximizing $AUC_i - 0.001 \times p_{\text{selected}(i)}$, where $p_{\text{selected}(i)}$ represents the number of non-zero coefficients in the model for the i th evaluated λ , and AUC_i is the corresponding AUC for i th evaluated λ . Once the λ value is selected, the training model was refitted using the full training data set, and the ‘outer’ cross-validation test set was predicted. 80 % of the data was used as training set in both the ‘outer’ and ‘inner’ cross-validation procedures, with 100 ‘outer’ cross-validation rounds and 50 ‘inner’ cross-validation rounds. Final ROC curves and corresponding AUC estimates were calculated from the ‘outer’ cross-validation test set, which were never part of the optimizing or fitting of the model. To determine a global set of selected variables over all cross-validation rounds, since one set of variables would be selected in each cross-validation round, we reported those variables that were selected in more than 10 % of the cross-validation rounds.

3 Results

3.1 Prediction performance in individual and combined full data sets

We fitted multivariate predictive models to the transcriptomics and metabolomics data and assessed how well we could discriminate between active UC, quiescent UC, and controls and six different subphenotypes in the study (Table 2). Each data set (two metabolomics and one transcriptomics data sets) was modelled separately, including all measured variables, and subsequently data was modelled in concatenated form, combining all variables measured on all three data sets. Using OPLS-DA together with class-balanced Monte Carlo cross-validation,

Table 2 Samples sizes in different subgroups of data

Analysis name	Case name	<i>N</i> total	<i>N</i> cases	Proportion cases
ActiveControl	Active	36	21	0.58
InactiveControl	Inactive	34	19	0.56
InactiveActive	Active	40	21	0.53
SiSd	Si	24	16	0.67
DurationLess10	Duration	40	12	0.30
DebutLess25	DebutLess25	40	29	0.73

Analysis name subgroups of data, *Case name* group defined as cases, *N total* number of observations in total, *N cases* number of cases, *Proportion cases* fraction of total number of samples that belong to the case group, *Si* steroid independence, *Sd* steroid dependence, *DurationLess10* disease duration less than 10 years, *DebutLess25* age at diagnosis less than 25 years

we estimated prediction performance as AUC for each comparison and each individual and combined data set (Table 3). The best prediction performances were observed with active UC versus quiescent UC (AUC >0.95 across all data sets, Table 3), and active UC versus controls (AUC >0.92 across all data sets, Fig. 1). With quiescent UC versus controls the classification performance (Fig. 1) was lower and more variable between the data sets, ranging from AUC = 0.57–0.79, with the metabonomic (NOE) performing the best.

ROC curves corresponding to all evaluated models across all the data sets are found in Supplementary Material.

3.2 Prediction performance in small candidate biomarker panels selected from individual and combined data sets

We utilised a multivariate logistic regression model fitted by lasso, and applied a nested cross-validation procedure to minimize the risk of over-fitting in order to evaluate the prediction performance of relatively limited (<~20

variables) multivariate biomarker panels. Prediction performance under variable selection for each subpopulation and each data set were estimated (Table 4). The results suggest that it is possible to discriminate between active UC, quiescent UC, and controls; between Si and Sd, and between the ages of onset, i.e. before and after 25 years of age, based on relatively small subsets of transcriptomic and metabonomic predictors. Discrimination between active UC and controls (Fig. 1) is in the range of AUC = 0.93–0.97, and for quiescent UC and controls AUC estimates are found to be ranging between 0.68 and 0.76 across the data sets. ROC curves corresponding to all evaluated models across all the data sets are found in Supplementary Material.

The lasso based logistic regression model applied in this study provides us with relatively small candidate subsets of biomarkers for evaluation. Figure 2 shows the distribution of the number of selected variables over cross-validation rounds for each subpopulation in classification and variable selection for each data set. Table 5 gives an example of candidate biomarkers for the inactive UC versus controls in the combined data set. Tables of candidate biomarker panels (selected >10 % of the cross-validation rounds) are provided for each data set in the Supplementary Material for all data sets and models.

4 Discussion

The objective of the current study was to apply a novel approach in which transcriptomics and metabonomics data were merged with the intention to improve diagnostics, biomarker identification, and to explore the possibilities of a molecular phenotyping in UC patients. Thus, herein we clearly demonstrate for the first time that transcriptomics and metabonomics data, both as separate data sets and as merged omics data, hold substantial differential power with respect to active UC, quiescent UC, and controls, and with different UC subphenotypes as well.

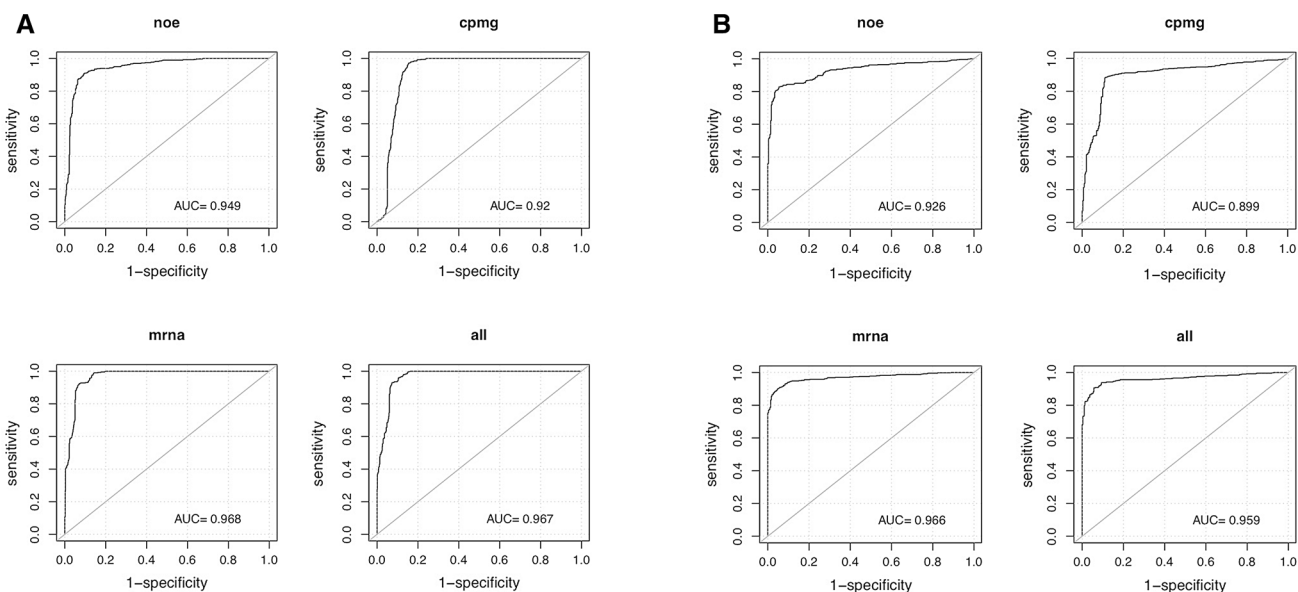
Table 3 Prediction performance

Analysis name	Metabonomics (NOE)	Metabonomics (CPMG)	Transcriptomics	Omics
ActiveControl	0.95	0.92	0.97	0.97
InactiveControl	0.79	0.65	0.57	0.58
InactiveActive	0.98	0.95	0.96	0.96
SiSd	0.73	0.76	0.80	0.78
DurationLess10	0.43	0.35	0.63	0.63
DebutLess25	0.43	0.34	0.38	0.36

Prediction performance estimates presented as area under the curve (AUC) for each sub-population and for each full data set, including combined data sets: Omics. AUC estimates are based on classification using OPLS-DA and Monte Carlo cross-validation

NOE nuclear Overhauser effect, *CPMG* Carr–Purcell–Meiboom–Gill, *Si* steroid independence, *Sd* steroid dependence, *DurationLess10* disease duration less than 10 years, *DebutLess25* age at diagnosis less than 25 years

Active UC vs. controls



Inactive UC vs. controls

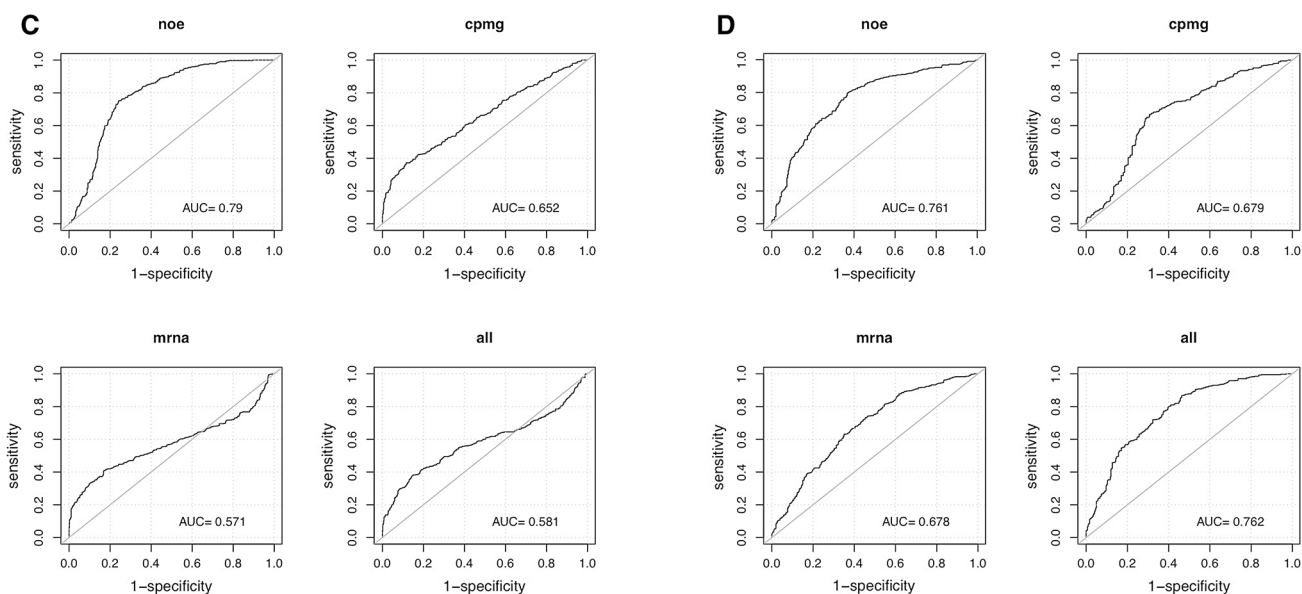


Fig. 1 Prediction performance estimates presented as area under the curve. Prediction performance estimates using OPLS-DA (**a, c**) or logistic regression fitted by lasso (**b, d**) on metabolomics (NOE and CPMG), transcriptomics (mrna), and omics (all) data sets

Previous attempts to classify quiescent UC based on the CPMG metabolomics data (Bjerrum et al. 2010b) has been unsuccessful, but with the use of the NOE metabolomics data and a predictive level approach the classification model with an $AUC = 0.79$ identifies quiescent UC as a distinct metabolic phenotype compared to controls, which is an original finding.

Intriguingly, the classification model was also able to differentiate between patients with Sd and Si; AUC ranging

from 0.73 to 0.80 (Table 3) using the full data set. This finding is of utmost importance, as early identification of patients with Sd will lead to optimized and personalized medication in terms of timely introduction of immunomodulators (e.g., thiopurines and methotrexate) or biologics (Ardizzone et al. 2012). Both adult (Faubion et al. 2001) and paediatric (Tung et al. 2006) studies emphasize that most UC patients initially respond to steroids, but already after 1 year a significant proportion (up to 43 % in

Table 4 Prediction performance of small candidate biomarker panels

Analysis name	Metabonomics (NOE)	Metabonomics (CPMG)	Transcriptomics	Omics
ActiveControl	0.93	0.90	0.97	0.96
InactiveControl	0.76	0.68	0.68	0.76
InactiveActive	0.94	0.89	0.93	0.93
SiSd	0.62	0.71	0.63	0.70
DurationLess10	0.54	0.50	0.38	0.38
DebutLess25	0.49	0.67	0.71	0.69

Prediction performance estimates presented as area under the curve (AUC) for each sub-population under selection of small candidate biomarker panels for each full data set and the combined data sets: Omics. AUC estimates are based on classification using logistic regression fitted by lasso and a nested Monte Carlo cross-validation procedure

NOE nuclear Overhauser effect, *CPMG* Carr–Purcell–Meiboom–Gill, *Si* steroid independence, *Sd* steroid dependence, *DurationLess10* disease duration less than 10 years, *DebutLess25* age at diagnosis less than 25 years

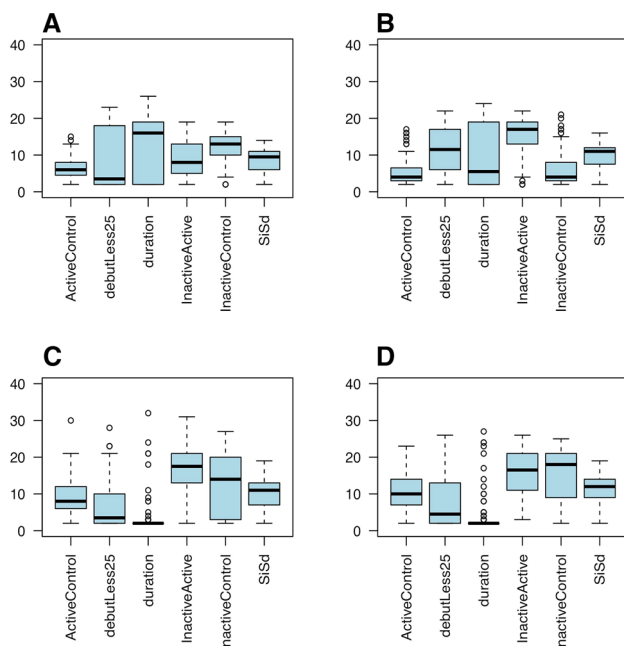


Fig. 2 Distribution of the size of the selected biomarker panels over 100 (outer) cross-validation rounds. **a** Metabonomics NOE, **b** Metabonomics CPMG, **c** transcriptomics, and **d** Omics data set. The *box border* represents the interquartile range and the *horizontal line* in the *box* is the median. The *whiskers* show the largest/smallest observation falling within a distance of 1.5 times the box size

paediatric patients) loses the response leading to Sd or surgery. Avoiding the latter would have a considerable impact on the patients' quality of life.

Classification models that contain thousands of variables are clinically non-applicable. Thus, besides the merging of 'omics' data, the overarching objective of the current study was to evaluate the prediction performance of relative small (<20 variables) multivariate biomarker panels. The results (Table 4) suggest that it is actually possible to discriminate between active UC, quiescent UC, and controls, and between Sd and Si with similar AUC estimates as for the full data set (Table 3). The fact that these biomarker

panels maintain and in some cases improve, i.e. InactiveControl and DebutLess25 increase from AUC of 0.58–0.76 and 0.36–0.69, respectively (Tables 3, 4), the discriminative power holds great promise for the development of clinically applicable similar discriminative power has to be demonstrated in extracts from colonic biopsies so that targeted biochemical assays can be developed, as NMR spectroscopy is not expected to be readily available in an everyday diagnostic lab.

The biomarker panels are also informative in terms of the dominant pathophysiological processes at play. In this respect it is interesting to note that biomarker panels related to Sd and Si contain transcripts (protein inhibitor of activated STAT 4, amino-terminal enhancer of split, and hairy/enhancer-of-split related with YRPW motif-like, see Supplementary Material, Table 14) that are known co-repressors or co-regulators of steroid hormone receptors (Lavery and McEwan 2005; Lavery et al. 2011). The same panel also contains the transcript SLC38A5, which encodes a Na-glutamine co-transporter found on the brush border membrane of enterocytes. During chronic intestinal inflammation Na-glutamine co-transporters are deregulated (Saha et al. 2012), and recent evidence indicates that this deregulation is due to an interaction with the glucocorticoid receptor (Arthur et al. 2012). Glutamine is a major respiratory fuel for gut-associated immune cells and enterocytes and plays an important part in the synthesis of the endogenous thiol antioxidant glutathione. This coincides with metabolites, i.e. glutamine and glutathione, found in the biomarker panels of the Sd/Si model (Supplementary Material, Table 15). Thus, besides providing a molecular snapshot at the different 'omics' levels, the integration of 'omics' also represents a systems biology approach in which entire pathways of importance to the pathophysiology can be revealed; alterations in co-repressors or co-regulators of steroid hormone receptors might lead to changes in Na-glutamine co-transporter and subsequently

Table 5 Most frequently selected variables in the inactive UC versus control

Name	HGNC	ppm	Data set	Selection freq.
ENSG00000243024_at	RPS11P6	3.176	Transcriptomics	0.81
			Metabonomics	0.77
ENSG00000165669_at	FAM204A		Transcriptomics	0.69
ENSG00000121351_at	IAPP		Transcriptomics	0.55
ENSG00000120675_at	DNAJC15		Transcriptomics	0.5
ENSG00000100664_at	EIF5		Transcriptomics	0.48
ENSG00000128881_at	TTBK2		Transcriptomics	0.47
ENSG00000116885_at	OSCP1		Transcriptomics	0.46
ENSG00000166323_at	C11orf65	3.093	Transcriptomics	0.31
			Metabonomics	0.29
ENSG00000159208_at	C1orf51		Transcriptomics	0.29
ENSG00000119487_at	MAPKAP1		Transcriptomics	0.27
ENSG00000164291_at	ARSK		Transcriptomics	0.26
ENSG00000171786_at	NHLH1		Transcriptomics	0.23
ENSG00000198590_at	C3orf35		Transcriptomics	0.17
ENSG00000051620_at	HEBP2		Transcriptomics	0.15
ENSG00000175550_at	DRAP1		Transcriptomics	0.14
ENSG00000122585_at	NPY		Transcriptomics	0.13
ENSG00000131697_at	NPHP4	2.8 (Aspartate)	Transcriptomics	0.12
			Metabonomics	0.11
ENSG00000111224_at	PARP11		Transcriptomics	0.11
ENSG00000189068_at	VSTM1		Transcriptomics	0.11
ENSG00000257726_at			Transcriptomics	0.11

Although there are fewer metabolites selected among the top predictors compared to transcripts, the metabolites carry a substantial amount of information in terms of classification, e.g. the NMR peak (unknown annotation) at 3.176 ppm is selected in 77 % of all cross-validation rounds. Even though the table is dominated by transcriptomics variables, the metabolomic data is indeed adding substantially impact to the classification performance

Name ENSEMBL id/metabolite name (if available), *HGNC* HGNC gene ID, *ppm* Peak position in NMR spectrum for metabolomic variables, *Data set* data set the variable belong to, *Selection freq.* frequency of which the variable is selected over (outer) cross-validation rounds

inadequate amounts of glutamine, which is known to become a conditionally essential amino acid during pathophysiological conditions.

Similarly, the biomarker panels related to early and late disease debut (see Supplementary Material, Table 22) contain transcripts [(CD5 molecule-like (CD5L) and autophagy related 10 (ATG10))] highly involved in phagocytosis and autophagy. Thus, CD5L also known as Api6 (apoptosis inhibitor 6) promotes macrophage survival in an inflammatory milieu, recognizes Gram-positive and negative bacteria, and enhances the phagocytic activity of macrophages (Haruta et al. 2001; Sarrias et al. 2005). In support of the later, ATG10 is a vital element of the autophagic process (Shao et al. 2007), which concurs with the dominant metabolites, i.e. lipids, choline, and inositols, in the associated biomarker panels of metabolites (see Supplementary Material, Table 23); lipids and choline are essential components of the phagocytic double membrane vesicles and inositols are a basic requirement for the initiation of autophagy (Codogno et al. 2012). Genome wide

association studies have previously identified IBD-associated genetic variants that have either genetic or functional links to autophagy, but the majority of these genes are thought to be specific CD susceptibility factors (Kabi et al. 2012). However, the current systems biology approach reveals that phagocytosis and autophagy might also be central pathophysiological elements in UC at a subphenotypical level, which molecularly characterizes early and late onset of the disease. In this context it is of significant interest that autophagy and especially Api6 is associated with the development of inflammation and adenocarcinoma (Qu et al. 2009), which is in line with the increased risk of CRC in patients with early onset of UC (Jess et al. 2012).

Finally, it should be noted that no single transcriptomics or metabolomics data set performs better across all populations investigated. Joint modelling of both data sets performs about as good as or marginally worse than the best of the metabolomics and transcriptomics data set. This is largely due to the fact that in the combined data, the number of transcript variables is vastly larger than the number of metabolic variables

(Table 3). In the logistic regression model fitted by lasso the larger number of transcripts is, however, less influential since this model performs variable selection (Table 4). Thus, transcriptomics and metabonomics are informative for classification, while the anticipated diagnostic advantages of combining ‘omics’ remain elusive. The later might be explained by a potentially low correlation between ‘omics’ and the moderately sized study (Schwanhäusser et al. 2011). Furthermore, despite the intention to build a robust, generic classifier that does not depend on factors such as diet, alcohol consumption, and medication, it is impossible to completely rule out a confounding effect from these elements. While the spectra from snap-frozen samples are highly reproducible, the minor changes induced by the snap-freezing process and the low temperature (10 °C) prior to the acquisition of spectra need to be acknowledged as potential confounders as well. Finally, the results also need to be interpreted in light of the cross-validation procedure. Although we apply rigorous cross-validation approaches to mitigate problems of bias and over-fitting as much as possible, gold standard evaluation of prediction performance should ideally be done through external data sets, which was not possible in the current setting.

In conclusion, this study demonstrates that the novel approach of integrating metabonomics and transcriptomics combines the better of the two worlds, and provides us with potentially clinical applicable biomarker panels. These panels improve diagnostics and more importantly also the molecular phenotyping in UC, making optimized and personalized medication a possibility. However, in order to achieve a more complete systems biology approach we want to stress the importance of integrating not just two but multiple biomolecular levels. Consequently, follow-up studies containing larger sample size, CD, IBS, infectious colitis patients, and a more complete ‘omics’ approach are motivated in order to validate the current results and to improve accuracy and robustness in the selection of the biomarker panels.

Acknowledgments This study was supported by grants from the Danish Agency for Science, Technology and Innovation. M.R. is supported by a Fellowship for Young International Scientists (Chinese Academy of Sciences, PR of China) and a MRC Biomedical Informatics Fellowship (Medical Research Council, UK, fellowship G0802460).

Conflict of interest The authors declared no competing interests.

References

- Ardizzone, S., Cassinotti, A., & de Franchis, R. (2012). Immunosuppressive and biologic therapy for ulcerative colitis. *Expert Opinion on Emerging Drugs*, 17(4), 449–467.
- Arthur, S., Saha, P., Sundaram, S., Kekuda, R., & Sundaram, U. (2012). Regulation of sodium-glutamine cotransport in villus and crypt cells by glucocorticoids during chronic enteritis. *Inflammatory Bowel Diseases*, 18(11), 2149–2157.
- Astrakas, L., Blekas, K. D., Constantinou, C., et al. (2011). Combining magnetic resonance spectroscopy and molecular genomics offers better accuracy in brain tumor typing and prediction of survival than either methodology alone. *International Journal of Oncology*, 38(4), 1113–1127.
- Balasubramanian, K., Kumar, S., Singh, R. R., et al. (2009). Metabolism of the colonic mucosa in patients with inflammatory bowel diseases: An in vitro proton magnetic resonance spectroscopy study. *Magnetic Resonance Imaging*, 27(1), 79–86.
- Baumgart, D. C., & Sandborn, W. J. (2012). Crohn’s disease. *Lancet*, 380(9853), 1590–1605.
- Bax, A., & Davis, D. G. (1985). MELV-17-based two-dimensional homonuclear magnetization transfer spectroscopy. *Journal of Magnetic Resonance*, 65, 355–360.
- Beckonert, O., Coen, M., Keun, H. C., et al. (2010). High-resolution magic-angle-spinning NMR spectroscopy for metabolic profiling of intact tissues. *Nature Protocols*, 5(6), 1019–1032.
- Bentley, E., Jenkins, D., Campbell, F., & Warren, B. (2002). How could pathologists improve the initial diagnosis of colitis? Evidence from an international workshop. *Journal of Clinical Pathology*, 55(12), 955–960.
- Bezabeh, T., Somorjai, R. L., Smith, I. C., Nikulin, A. E., Dolenko, B., & Bernstein, C. N. (2001). The use of 1H magnetic resonance spectroscopy in inflammatory bowel diseases: Distinguishing ulcerative colitis from Crohn’s disease. *American Journal of Gastroenterology*, 96(2), 442–448.
- Bjerrum, J. T., Hansen, M., Olsen, J., & Nielsen, O. H. (2010a). Genome-wide gene expression analysis of mucosal colonic biopsies and isolated colonocytes suggests a continuous inflammatory state in the lamina propria of patients with quiescent ulcerative colitis. *Inflammatory Bowel Diseases*, 16(6), 999–1007.
- Bjerrum, J. T., Nielsen, O. H., Hao, F., et al. (2010b). Metabonomics in ulcerative colitis: Diagnostics, biomarker identification, and insight into the pathophysiology. *Journal of Proteome Research*, 9(2), 954–962.
- Bjerrum, J. T., Nielsen, O. H., Wang, Y. L., & Olsen, J. (2008). Technology insight: Metabonomics in gastroenterology—basic principles and potential clinical applications. *Nature Clinical Practice Gastroenterology & Hepatology*, 5(6), 332–343.
- Borgan, E., Sitter, B., Lingjærde, O. C., et al. (2010). Merging transcriptomics and metabonomics—advances in breast cancer profiling. *BMC Cancer*, 10, 628.
- Bylesjö, M., Rantalainen, M., Cloarec, O., Nicholson, J. K., Holmes, E., & Trygg, J. (2006). OPLS discriminant analysis: Combining the strengths of PLS-DA and SIMCA classification. *Journal of Chemometrics*, 20(8–10), 341–351.
- Bylesjö, M., Rantalainen, M., Nicholson, J. K., Holmes, E., & Trygg, J. (2008). K-OPLS package: Kernel-based orthogonal projections to latent structures for prediction and interpretation in feature space. *BMC Bioinformatics*, 9, 106.
- Chen, R., Mias, G. I., Li-Pook-Than, J., et al. (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, 148(6), 1293–1307.
- Codogno, P., Mehrpour, M., & Proikas-Cezanne, T. (2012). Canonical and non-canonical autophagy: Variations on a common theme of self-eating? *Nature Reviews Molecular Cell Biology*, 13(1), 7–12.
- Costello, C. M., Mah, N., Häslér, R., et al. (2005). Dissection of the inflammatory bowel disease transcriptome using genome-wide cDNA microarrays. *PLoS Medicine*, 2(8), e199.
- Dieterle, F., Ross, A., Schlotterbeck, G., & Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Analytical Chemistry*, 78(13), 4281–4290.

- Faubion, W. A., Loftus, E. V., Harmsen, W. S., Zinsmeister, A. R., & Sandborn, W. J. (2001). The natural history of corticosteroid therapy for inflammatory bowel disease: A population-based study. *Gastroenterology*, *121*(2), 255–260.
- Fiocchi, C. (2012). Towards a ‘cure’ for IBD. *Digestive Diseases*, *30*(4), 428–433.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22.
- Geboes, K., Colombel, J.-F., Greenstein, A., et al. (2008). Indeterminate colitis: A review of the concept—What’s in a name? *Inflammatory Bowel Diseases*, *14*(6), 850–857.
- Gentleman, R. C., Carey, V. J., Bates, D. M., et al. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, *5*(10), R80.
- Haruta, I., Kato, Y., Hashimoto, E., et al. (2001). Association of AIM, a novel apoptosis inhibitory factor, with hepatitis via supporting macrophage survival and enhancing phagocytotic function of macrophages. *Journal of Biological Chemistry*, *276*(25), 22910–22914.
- Hurd, R. E. (1990). Gradient-enhanced spectroscopy. *Journal of Magnetic Resonance*, *87*, 422–428.
- Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., & Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, *31*(4), e15.
- Jess, T., Simonsen, J., Jørgensen, K. T., Pedersen, B. V., Nielsen, N. M., & Frisch, M. (2012). Decreasing risk of colorectal cancer in patients with inflammatory bowel disease over 30 years. *Gastroenterology*, *143*(2), 375–381.
- Kabi, A., Nickerson, K. P., Homer, C. R., & McDonald, C. (2012). Digesting the genetics of inflammatory bowel disease: Insights from studies of autophagy risk genes. *Inflammatory Bowel Diseases*, *18*(4), 782–792.
- Lavery, D. N., & McEwan, I. J. (2005). Structure and function of steroid receptor AF1 transactivation domains: Induction of active conformations. *Biochemical Journal*, *391*(Pt 3), 449–464.
- Lavery, D. N., Villaronga, M. A., Walker, M. M., Patel, A., Belandia, B., & Bevan, C. L. (2011). Repression of androgen receptor activity by HEYL, a third member of the Hairy/Enhancer-of-split-related family of Notch effectors. *Journal of Biological Chemistry*, *286*(20), 17796–17808.
- Meiboom, S., & Gill, D. (1958). Modified spin-echo method for measuring nuclear relaxation times. *Review of Scientific Instruments*, *29*(8), 688–691.
- Molodecky, N. A., Soon, I. S., Rabi, D. M., et al. (2012). Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. *Gastroenterology*, *142*(1), 46–54.
- Nicholson, J. K., Foxall, P. J., Spraul, M., Farrant, R. D., & Lindon, J. C. (1995). 750 MHz ¹H and ¹H-¹³C NMR spectroscopy of human blood plasma. *Analytical Chemistry*, *67*(5), 793–811.
- Nikolaus, S., & Schreiber, S. (2007). Diagnostics of inflammatory bowel disease. *Gastroenterology*, *133*(5), 1670–1689.
- Olsen, J., Gerds, T. A., Seidelin, J. B., et al. (2009). Diagnosis of ulcerative colitis before onset of inflammation by multivariate modeling of genome-wide gene expression data. *Inflammatory Bowel Diseases*, *15*(7), 1032–1038.
- Ordás, I., Eckmann, L., Talamini, M., Baumgart, D. C., & Sandborn, W. J. (2012). Ulcerative colitis. *Lancet*, *380*(9853), 1606–1619.
- Planell, N., Lozano, J. J., Mora-Buch, R., et al. (2012). Transcriptional analysis of the intestinal mucosa of patients with ulcerative colitis in remission reveals lasting epithelial cell alterations. *Gut*, *62*(7), 967–976.
- Qu, P., Du, H., Li, Y., & Yan, C. (2009). Myeloid-specific expression of Api6/AIM/Sp alpha induces systemic inflammation and adenocarcinoma in the lung. *The Journal of Immunology*, *182*(3), 1648–1659.
- Rantalainen, M., Bylesjö, M., Cloarec, O., Nicholson, J. K., Holmes, E., & Trygg, J. (2007). Kernel-based orthogonal projections to latent structures (K-OPLS). *Journal of Chemometrics*, *21*(7–9), 376–385.
- RDC Team. (2012). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0.
- Saha, P., Arthur, S., Kekuda, R., & Sundaram, U. (2012). N-glutamine co-transporters B(0)AT1 in villus and SN2 in crypts are differentially altered in chronically inflamed rabbit intestine. *Biochimica et Biophysica Acta*, *1818*(3), 434–442.
- Sarrias, M.-R., Roselló, S., Sánchez-Barbero, F., et al. (2005). A role for human Sp alpha as a pattern recognition receptor. *Journal of Biological Chemistry*, *280*(42), 35391–35398.
- Schroeder, K. W., Tremaine, W. J., & Ilstrup, D. M. (1987). Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis. A randomized study. *New England Journal of Medicine*, *317*(26), 1625–1629.
- Schwahnäusser, B., Busse, D., Li, N., et al. (2011). Global quantification of mammalian gene expression control. *Nature*, *473*(7347), 337–342.
- Shao, Y., Gao, Z., Feldman, T., & Jiang, X. (2007). Stimulation of ATG12-ATG5 conjugation by ribonucleic acid. *Autophagy*, *3*(1), 10–16.
- Trygg, J., & Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, *16*(3), 119–128.
- Tung, J., Loftus, E. V., Freese, D. K., et al. (2006). A population-based study of the frequency of corticosteroid resistance and dependence in pediatric patients with Crohn’s disease and ulcerative colitis. *Inflammatory Bowel Diseases*, *12*(12), 1093–1100.
- Tzika, A. A., Astrakas, L., Cao, H., et al. (2007). Combination of high-resolution magic angle spinning proton magnetic resonance spectroscopy and microscale genomics to type brain tumor biopsies. *International Journal of Molecular Medicine*, *20*(2), 199–208.
- von Stein, P., Lofberg, R., Kuznetsov, N. V., et al. (2008). Multigene analysis can discriminate between ulcerative colitis, Crohn’s disease, and irritable bowel syndrome. *Gastroenterology*, *134*(7), 1869–1881.
- Wang, Y., Holmes, E., Comelli, E. M., et al. (2007a). Topographical variation in metabolic signatures of human gastrointestinal biopsies revealed by high-resolution magic-angle spinning ¹H NMR spectroscopy. *Journal of Proteome Research*, *6*(10), 3944–3951.
- Wang, L., Lofton, C., Popp, M., & Tan, W. (2007b). Using luminescent nanoparticles as staining probes for Affymetrix GeneChips. *Bioconjugate Chemistry*, *18*(3), 610–613.
- Waters, N. J., Garrod, S., Farrant, R. D., et al. (2000). High-resolution magic angle spinning (¹H) NMR spectroscopy of intact liver and kidney: Optimization of sample preparation procedures and biochemical stability of tissue during spectral acquisition. *Analytical Biochemistry*, *282*(1), 16–23.
- Wu, F., Dassopoulos, T., Cope, L., et al. (2007). Genome-wide gene expression differences in Crohn’s disease and ulcerative colitis from endoscopic pinch biopsies: Insights into distinctive pathogenesis. *Inflammatory Bowel Diseases*, *13*(7), 807–821.