Phenotype prediction based on microRNA expression profiles: a novel diagnostic tool for inflammatory bowel disease

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von
Matthias Hübenthal

Kiel, 2019

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation eigenständig und nur mit den von mir angegebenen Quellen und Hilfsmitteln sowie der wissenschaftlichen Beratung durch meine Betreuer angefertigt habe. Diese Abhandlung hat weder ganz noch in Teilen bereits an anderer Stelle im Rahmen eines Prüfungsverfahrens vorgelegen. Es handelt bei dieser Arbeit um eine kumulative Dissertation. Entsprechend wurde sie in Teilen bereits als Beitrag in Fachzeitschriften, Herausgeberschriften oder in Form einer Patentanmeldung/Patentschrift veröffentlicht. Ich erkläre, dass die vorliegende Dissertation unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft entstanden ist. Weiterhin versichere ich, dass mir bisher kein akademischer Grad entzogen wurde.

Kiel, 28.03.2019   Matthias Hübenthal
(Ort, Datum)   (Unterschrift)

Zusammenfassung

Bei entzündlichen Darmerkrankungen (inflammatory bowel disease, IBD) handelt es sich um eine Gruppe chronischer, rezidivierender Krankheiten des Verdauungstraktes, die die Hauptformen Morbus Crohn (Crohn's disease, CD) und Colitis ulcerosa (ulcerative colitis, UC) umfasst. Als eine zunehmend häufige Diagnose belastet IBD weltweit Gesellschaften und Gesundheitssysteme in erheblichem Maße. Nach wie vor stellt die Diagnostik von IBD eine klinische Herausforderung dar und schließt die Bewertung zahlreicher endoskopischer, histologischer, radiologischer und/oder biochemischer Parameter ein. Als Ergänzung des diagnostischen Prozesses wurden zahlreiche Biomarker vorgeschlagen. Für die routinemäßige klinische Praxis jedoch kann keiner dieser Biomarker empfohlen werden.

Bei microRNAs (miRNAs) handelt es sich um eine Klasse kurzer, nicht-kodierender RNAs, die als post-transcriptionelle Regulatoren der Genexpression fungieren. Variabilität innerhalb der miRNA-Biogenese resultiert aus der alternativen Prozessierung unreifer (precursor cropping, precursor dicing), aber auch reifer miRNAs (sequence editing, sequence trimming, nucleotide addition) und bringt miRNA-Isoformen (isomiRs) hervor. Aufgrund ihrer bemerkenswerten Widerstandsfähigkeit gegenüber physikalischer und enzymatischer Degradation, wurden im Krankheitsfall deregulierte miRNAs als neuartige Biomarkers für eine Vielzahl von Erkrankungen vorgeschlagen. Unter Verwendung globaler Expressionprofile sowie zeitgemäßer Methoden des maschinellen Lernens, wird im Rahmen der vorliegenden Arbeit die Eignung dieser Moleküle als nicht-invasives Werkzeug für die Diagnose von IBD evaluiert.

In einer ersten Studie verwendeten wir Microarraytechnologie zur Quantifizierung blutbasierter Expressionslevel von 863 miRNAs, generiert auf Grundlage einer 314 deutsche Individuen umfassenden Kohorte. Diese resultierenden Expressionsprofile bildeten die Grundlage für die Etablierung von miRNA-Signaturen, hilfreich für die hochgenaue Unterscheidung zwischen CD und UC untereinander, sowie deren Abgrenzung von gesunden und inflammatorischer Kontrollen. Basierend auf diesen Signaturen generierten wir binomiale sowie multinomiale Modelle zur Lösung klinisch relevanter Klassifikationsprobleme. Die Evaluation der Klassifikatoren erfolgte unter Verwendung umfangreichen Samplings und ergab mediane balanced accuracies von bis zu 91.70% (CD vs. UC) bzw. 98.10% (CD oder UC vs. HC). In einer weiteren Studie wurde dieser Ansatz um die zusätzliche Berücksichtigung von miRNA-Varianten erweitert. Wir verwendeten Sequenziertechnologie der nächsten Generation zur Bestimmung von isomiR-Expressionsprofilen, generiert auf Grundlage einer Kohorte von 515 Individuen, einschließlich IBD-Patienten sowie gesunder und symptomatischer Kontrollen. Unter Verwendung charakteristischer isomiR-Signaturen generierten wir Klassifikatoren, deren Klassifikationgüte wir mit medianen balanced accuracies von 78.57%/78.83% (CD vs. UC, unbehandelt/behandelt) bzw. 100.00%/98.28% (CD oder UC vs. HC, unbehandelt/behandelt) bewerten konnten. Damit liefern wir statisch belastbare Belege für eine Überlegenheit unserer Modelle bezüglich etablierter Biomarker.

Im Weiteren schlagen wir eine zusätzliche Verbesserung unserer Modelle durch die Berücksichtigung von isomiR-Signaturen, spezifisch für die zellulären (NK-Zellen, B-Lymphozyten, cytotoxische T-Lymphozyten, T-Helferzellen, Monozyten, Neutrophile, Erythrozyten) und nicht-zellulären Komponenten peripheren Bluts (Serum, Exosomen) vor. In einer abschließenden Studie etablierten wir dafür einen umfangreichen Katalog zelltyp-spezifischer microRNA-Expression. Basierend auf dieser Ressource identifizierten wir Gruppen von miRNAs/isomiRs, spezifisch differenziell exprimiert in jeder der untersuchten Blutkomponenten. Schließlich liefern wir Belege für die Zelllinien-Spezifität bestimmter miRNA-Sequenzvarianten (3' trimming, 3' non-templated nucleotide addition).

Summary

Inflammatory bowel disease (IBD) is a chronic relapsing disorder of the alimentary tract, encompassing Crohn's disease (CD) and ulcerative colitis (UC) as its two major subtypes. Recognized as an increasingly common condition, it renders a significant burden to societies and healthcare systems around the world. The diagnosis of inflammatory bowel disease remains a clinical challenge and involves the assessment of numerous endoscopic, histological, radiological and/or biochemical parameters. A multitude of biomarkers has been proposed to complement the diagnostic process. However, none of them can be recommended for routine clinical practice.

MicroRNAs (miRNAs) represent a class of short, non-coding RNAs that act as post-transcriptional regulators of gene expression. Variability within the miRNA biogenesis is introduced by alternative precursor cropping and dicing, sequence editing or trimming as well as addition of nucleotides and gives rise to miRNA isoforms (isomiRs). Due to their remarkable resistance against physical and enzymatic degradation, miRNAs exhibiting deregulation in disease have been implicated as novel biomarkers for a variety of clinical conditions. In the scope of this work we evaluate whether systematic miRNA or miRNA variant expression profiling, in conjunction with state-of-the-art machine learning techniques, is suitable as a non-invasive tool for diagnostics of IBD.

In a first study we employed microarray technology to determine expression levels of 863 miRNAs for whole blood samples drawn from a cohort comprising 314 German individuals, to establish miRNA signature being informative for the highly accurate distinction of CD and UC among each other as well as from healthy and inflammatory controls. Utilizing these signatures we generated binomial as well as multinomial models, solving clinically relevant classification tasks. We evaluated these classifiers employing a comprehensive sampling strategy, obtaining median balanced accuracies of up to 91.70% (CD vs. UC) and 98.10% (CD or UC vs. HC), respectively. In another study we extended this approach by additionally incorporating expression profiles of miRNA variants. We employed next generation sequencing technology to examine isomiR expression profiles drawn from a cohort of 515 individuals, comprising IBD cases as well as healthy and symptomatic controls. Incorporating distinctive isomiR signatures, we generated classifiers performing with median balanced accuracies of 78.57%/78.83% (untreated/treated CD vs. UC) and 100.00%/98.28% (untreated/treated CD or UC vs. HC), respectively. Hence, we provide sampling-based evidence for our models' superiority over established biomarkers.

We propose further improvement of our models by incorporating isomiR signatures dinstinctive for the cellular (NK cells, B lymphocytes, cytotoxic T lymphocytes, T helper cells, monocytes, neutrophils and erythrocytes) and non-cellular components of human peripheral blood (serum, exosomes). In a final study we elaborate on this idea by establishing a comprehensive, cell-type-specific microRNA expression catalogue required for this purpose. Based on this resource we identified sets of miRNAs being specifically DE in each of the investigated blood compounds. Furthermore, we provide evidence for cell-lineage-specificity of particular types of miRNA sequence variation (3' trimming and 3' non-templated nucleotide addition).

Acknowledgements

Contents

1 Introduction

1.1 Inflammatory bowel disease: epidemiology and etiology

Inflammatory bowel disease (IBD) is a chronic relapsing disorder of the alimentary tract. It encompasses two major subtypes, namely Crohn's disease (CD) and ulcerative colitis (UC), each showing extensive heterogeneity in terms of clinical and histological presentation as well as response to treatment.

In the course of the last decades IBD has been recognized as an increasingly common diagnosis in the industrialized countries. Recent studies of western populations estimate incidence and prevalence of IBD to be as high as 57.90 per 100,000 person-years and 505.00 per 100,000. However, the estimates vary considerably with respect to IBD subtype as well as geographic region. Thus, incidence of CD (UC) in North America, Europe and Oceania is reported with up to 23.82 (23.14), 15.40 (57.90) and 29.30 (17.40) per 100,000 person-years, respectively. Prevalence of CD (UC) stratified by the afore-mentioned regions is estimated as up to 318.50 (286.30), 322.00 (505.00) and 197.30 (196.00) per 100,000. This corresponds to 0.32% (0.29%), 0.32% (0.51%) and 0.20% (0.20%) of the respective population size and total numbers of approximately 1.16 (1.04), 2.39 (3.75) and 0.08 (0.08) million affected individuals. Whereas there is evidence for the incidence and prevalence to stabilize in the western world, estimates of both epidemiological measures are on the rise in newly industrialized countries. Currently, incidence of CD (UC) in South America, Africa and Asia is reported to equal to up to 3.50 (6.76), 5.87 (3.29) and 8.40 (6.50) per 100,000 person-years. Prevalence of CD (UC) for these regions is estimated as up to 41.40 (44.30), 19.02 (10.57) and 53.10 (106.20) per 100,000. This corresponds to 0.04% (0.04%), 0.02% (0.01%) and 0.05% (0.11%) of the respective population size and approximate total numbers of 0.18 (0.19), 0.24 (0.14) and 2.41 (4.83) million affected individuals. This provides evidence for IBD no longer being a problem of western countries alone. It rather renders a significant burden to societies and healthcare systems around the world. For details see Ng et al. [1].

A growing body of evidence suggests the disease being a result of dysregulated inflammatory response to commensal microbes in a genetically susceptible host. In a series of genome-wide association studies (GWAS, including [2, 3]) 241 genetic loci have been identified to be associated with IBD. While these GWAS findings added to the understanding of the genetic architecture of the disease, its pathogenesis is still barely understood and linking SNP associations to functional mechanisms is a major challenge to the field. Consequently, the different aspects of gene regulation, such as transcription [4], RNA processing [5], mRNA trafficking and metabolism [6] as well as protein degradation [7], also have been investigated in IBD.

1.2 Clinical presentation and diagnostic features

Inflammatory bowel disease can occur at any age. However, age of onset varies with respect to age group and exhibits a peak at an age of 20–30. IBD affects males and females to the same extent. Its clinical features are heterogeneous but typically comprise gastro-intestinal symptoms, such as abdominal pain and chronic diarrhea or systemic symptoms, such as fever, malaise and anorexia. Notably, rectal bleeding and fecal incontinence as well as tenesmus are more common in UC than in CD [8, 9].

UC is characterized by diffuse, continuous inflammation which extends proximally from the rectum and does not involve the ileum. Commonly neither granulomas nor fissures or fistulas are observed. Histologically inflammation is limited to the mucosa [8]. Hallmark of CD is discontinuous inflammation affecting the entire gastro-intestinal tract. Common features include the involvement of the ileum as well as the development of granulomas, fissures and fistulas. Histologically the disease is characterized by transmural inflammation [9]. With a frequency of almost 50%, IBD patients develop both, intestinal and extraintestinal symptoms. The most common manifestations include musculoskeletal (peripheral and axial arthropathies) and mucocutaneous (erythema nodosum, pyoderma gangrenosum, aphthous stomatitis) as well as ophthalmological (conjunctivitis, scleritis, uveitis) and hepatic diseases (primary sclerosing cholangitis, granulomatous hepatitis) [10].

According to the current evidence-based consensus published by European (ECCO, European Crohn's and colitis organisation) [8, 9] as well as American gastroenterological societies (ACG, American college of gastroenterology) [11, 12] there is still no gold standard for diagnosing Crohn's disease and ulcerative colitis, repectively. Accordingly, the diagnostic procedure for both IBD subtypes involves a combination of clinical, endoscopic, histological and radiological parameters. At present, neither genetic nor serological testing is recommended for routine diagnosis.

1.3 Serological biomarkers

A growing body of evidence suggests that IBD results from abberant immune response to commensal microbes. This hypothesis is supported by the detection of anti-glycan antibodies recognizing glycosylated cell wall components of microbial pathogens. Consequently, anti-glycan antibodies are among the most extensively investigated serological biomarkers for diagnosis and prognosis of IBD. Examples include anti-Saccharomyces cerevisiae antibodies (ASCA, targeting mannan), the anti-mannobioside carbohydrate antibodies (AMCA, targeting mannan), anti-laminaribioside carbohydrate antibodies (ALCA, targeting laminarin), anti-chitobioside carbohydrate antibodies (ACCA, targeting chitin), anti-chitin carbohydrate antibodies (anti-C, targeting chitin) as well as anti-laminarin carbohydrate antibodies (anti-L, targeting laminarin). In addition to anti-carbohydrate antibodies, antibodies targeting protein-based antigens are a common group of biomarkers, comprising perinuclear atypical neutrophil cytoplasmic antibodies (pANCAs) as well as antibodies against the bacterial outer membrane porin C (OmpC) and the bacterial flagellin CBir1 [13, 14]. Differences in the prevalence of the antigens in IBD in comparison to healthy controls or IBD subtypes among each other have been shown repeatedly. However, performance of derived serological tests as measured by sensitivity, specificity and balanced accuracy vary considerably (see Table 1.1).

With specificities ranging from 86.0 to 100.0% the distinction between IBD and HC based on these antibodies appears to perform well. However, all of them show considerably smaller sensitivities (ranging from 6.0 to 45.0%) leading to only modest balanced accuracies (ranging from 52.5 to 72.5%). The distinction between CD and HC exhibits higher variability with regard to specificities (ranging from 68.0 to 100.0%) as well as an increase of sensitivities (ranging from 11.0 to 59.0%) leading to only moderately improved balanced accuracies (ranging from 50.0 to 79.0%). Accordingly, even the best performing serological biomarker (ASCA) appears to be insufficient to replace standard diagnostic methods, such as endoscopy, histology and radiology to distinguish between IBD and HC (BAC of 60.5 to 72.5%) as well as between CD and HC (BAC of 60.5 to 79.0%). Of note, the same biomarkers allow a distinction from symptomatic controls, even without a drop in accuracy (BAC of 51.0 to 71.5% comparing IBD and SC, BAC of 51.0 to 79.5% comparing CD and SC). The even more challenging diagnostic problem of distinguishing between the IBD subtypes among each other also has been addressed by employing the afore-mentioned serological biomarkers. With balanced accuracies ranging from 46.5 to 86.0% they show comparable diagnostic performance. Due to the

high variability of the performance estimates the current applicability of these biomarkers is clearly limited. However, they might complement standard diagnostics, particularly in case of diagnostic uncertainty.

problem	marker	BAC (%)	SN (%)	SP (%)	citation
IBD vs. HC	ASCA	60.5–72.5	31–45	90–100	[14, 15]
	AMCA	50.5–61.5	9–26	92–97	[14, 15]
	anti-OmpC	60.5	27	94	[15]
	ACCA	46.0–58.0	6–19	86–97	[14, 15]
	pANCA	57.5	17	98	[15]
	anti-L	55.5	12	99	[14]
	ALCA	54.5	15	94–99	[14, 15]
	anti-C	52.5	7	98	[14]
IBD vs. SC	ASCA	66.0–71.5	41–45	91–98	[15, 16]
	ACCA	51.5–64.5	19–40	84–89	[15, 16]
	AMCA	59.5	26	93	[15]
	ALCA	54.0–59.5	15–20	93–99	[15]
	Anti-OmpC	51.0	27	75	[15]
CD vs. HC	ASCA	60.5–79.0	37–59	84–99	[17–20]
	AMCA	51.5–63.5	12–27	91–100	[17, 18]
	ALCA	51.5–60.0	15–26	88–94	[17, 18]
	anti-L	58.0	26	90	[18]
	anti-C	56.5	25	88	[18]
	ACCA	50.0–56.5	11–17	89–96	[17, 18]
CD vs. SC	ASCA	64.0–79.5	51–63	77–96	[16–18]
	AMCA	52.0–63.5	12–27	92–100	[17, 18]
	ALCA	52.0–62.0	15–26	89–98	[16–18]
	anti-L	57.5	26	89	[18]
	ACCA	45.5–57.5	11–17	80–98	[16–18]
	anti-C	51.0	25	77	[18]
CD vs. UC	ASCA+	59.5–86.0	37–72	82–100	[14–19, 21–26]
	ASCA+/ANCA−	72.0–79.0	52–64	92–94	[19, 21–24, 27]
	pANCA−	71.5	52	91	[17]
	AMCA	47.0–62.5	12–28	82–97	[14, 15, 17, 18]
	anti-L	55.5–61.5	18–26	93–97	[14, 18]
	anti-C	50.0–61.5	10–25	90–98	[14, 18]
	ALCA	53.5–61.0	15–26	92–96	[14–18]
	ACCA	46.5–59.0	9–21	84–97	[14–18]
	anti-OmpC	55.0	29	81	[15]

Table 1.1: **Serological biomarkers in IBD [28, modified from Table 2 and 3].** Stratified by diagnostic problem the table summarizes previously studied serological biomarkers and reports their performance in terms of balanced accuracy (BAC), sensitivity (SN) and specificity (SP). ASCA = anti-saccharomyces cerevisiae antibody, ACCA = anti-chitobioside carbohydrate antibody, ALCA = anti-laminaribioside carbohydrate antibody, AMCA = anti-mannobioside carbohydrate antibody, anti-OmpC = anti-outer membrane porin C antibody, anti-C = anti-chitin carbohydrate antibody, anti-L = anti-laminarin carbohydrate antibody, pANCA = atypical perinuclear anti-neutrophil cytoplasmic antibody.

1.4 Fecal biomarkers

Mucosal immune response is associated with the migration of leukocytes (predominantly neutrophils) along a chemotactic gradient and their subsequent degradation at the site of inflammation, leading to the release of proteins into the intestinal lumen. Accordingly, fecal levels of those proteins (predominantly calprotectin and lactoferrin) are assumed to be elevated in response to inflammatory signals and have been extensively studied

as biomarkers distinguishing IBD from functional gastrointestinal disorders (including IBS) and health or predicting of disease activity. Of note, there is currently no fecal biomarker being able to distinguish between the substypes of IBD with high sensitivity and sprecificity, respectively.

Calprotectin, a heterodimer of calcium-binding S100 proteins (S100A8 and S100A9), allows a distinction between IBD and IBS with balanced accuracies ranging from 79.4 to 98.5%. Distinguishing between UC or CD and IBS, in turn, has been estimated to perform with a balanced accuracy of 91.5%. Lactoferrin, a monomeric member of the transferrin family of iron-binding glycoproteins, allows a discrimination between IBD and IBS with a balanced accuracy of 93.0%. Discriminating active IBD (or CD and UC, respectively) from inactive IBD (or CD and UC, respectively), functional bowel disorders or health, however, has been estimated to perform with balanced accuracies ranging from 70.9 to 82.0%. Accordingly, fecal calprotectin as well as lactoferrin serve as reliable indicators for intestinal inflammation (see Table 1.2). Nonetheless, neither the European Crohn's and Colitis Organisation (ECCO) [9] nor the American College of Gastroenterology (ACG) [11] recommend sole reliance of IBD diagnosis on these fecal biomarkers.

problem	marker	BAC (%)	SN (%)	SP (%)	citation
IBD vs. IBS	calprotectin	98.5	100.0	97.0	[29]
IBD vs. IBS	calprotectin	97.3	94.5	100.0	[30]
IBD vs. IBS	calprotectin	96.5	93.0	100.0	[31]
IBD vs. IBS	calprotectin	91.5	83.0	100.0	[32]
IBD vs. IBS	calprotectin	79.4	94.4	64.3	[33]
CD vs. IBS	calprotectin	91.5	83.0	100.0	[32]
UC vs. IBS	calprotectin	91.5	83.0	100.0	[32]
IBD vs. IBS	lactoferrin	93.0	86.0	100.0	[34]
active IBD vs. inactive IBD/IBS	lactoferrin	82.0	86.7	77.2	[35]
active CD vs. inactive CD/IBS/BIBD/HC	lactoferrin	86.0	92.0	80.0	[36]
active CD vs. inactive CD/IBS	lactoferrin	70.9	81.8	60.0	[35]
active UC vs. inactive UC/IBS/BIBD/HC	lactoferrin	89,0	90.0	88.0	[36]
active UC vs. inactive UC/IBS	lactoferrin	79.7	92.6	66.7	[35]

Table 1.2: **Fecal biomarkers in IBD [37, modified from Table 1 and 2].** Stratified by diagnostic problem the table summarizes previously studied fecal biomarkers and reports their performance in terms of balanced accuracy (BAC), sensitivity (SN) and specificity (SP). BIBD = bacteria infectious bowel disease.

1.5 MicroRNAs

MicroRNAs (miRNAs) represent a class of short, non-coding RNAs that act as post-transcriptional regulators of gene expression. Individual miRNAs regulate specific mRNA targets recognized by the RNA-induced silencing complex (RISC), whose formation corresponds to the end point of a multi-step biogenesis [38]. Variability in the biogenerative process is introduced by alternative precursor cropping and dicing, sequence editing or trimming as well as addition of nucleotides and gives rise to miRNA isoforms (isomiRs) [39].

miRNAs first have been described in Caenorhabditis elegans. Originally referred to as small temporal RNAs (stRNAs) due their role as key regulators of developmental timing [40, 41], they later have been recognized as a general class of post-transcriptional regulators being evolutionary conserved in a multitude of species [42–44]. In the course of the last years, thousands of miRNAs have been identified. Accordingly, miRBase, the primary miRNA sequence repository, now lists 48885 mature miRNAs (3982 with high confidence) from 271 species (20 with high confidence), including nematodes, insects, plants and mammals. At present, 2656 miRNAs have been identified in humans (1198 with high confidence) [45].

miRNAs have been recognized as important determinants of innate as well as adaptive immunity. They are involved in several steps of myelopoiesis, such as the development of MEPs (megakaryocyte-erythroid progenitors) and GMPs (granulocyte-macrophage progenitors) from CMPs (common myeloid progenitors) as well as their differentiation into terminal myeloid cells (dendritic cells, macrophages, platelets and erythrocytes). In the latter, miRNAs regulate innate immune response, e.g. via PAMP-dependent activation of NF-κB and MAPKs and subsequent induction of pro-inflammatory cytokines and antigen presentation genes. Moreover, miRNAs are associated with a multitude of developmental steps being part of lymphopoiesis, comprising the differentiation of B cells as well as different types of T cells from CLPs (common lymphoid progenitors). B cells further differentiate into plasma cells (mediating humoral immune response by producing high-affinity antibodies). CD4+ T cells differentiate into T helper cells (T_H, regulating cellular response to pathogens and tumors) or regulatory T cells (T_{reg}, mediating immune tolerance by interference with the activity of T_H cells). Driven by antigen exposure and inflammation, CD8+ T cells in turn, differentiate into cytotoxic T cells (T_C, inducing apoptosis via the caspase or perforin/granzyme pathway). Development, activation and function of all lymphoid cell types depend on complex regulatory networks comprising a multitude of miRNAs [46].

In accordance to their regulatory role in immune cell development and response, miRNAs are critical modulators of intestinal homeostasis. In line with this, they regulate autophagy (lysosomal degradation of cytoplasmic content) by direct or indirect impairment of the process of vesicle elongation [47]. Moreover, they interfere with the formation of tight junctions (affecting paracellular permeability), adherent junctions (affecting calcium-dependent adhesion of adjacent epithelial cells and spatial organization actin filaments) as well as desmosomes (affecting calcium-independent hyperadhesion of adjacent epithelial cells) and are therefore likewise contributing to intestinal homeostasis by regulation of epithelial integrity and barrier function [48].

Due to their remarkable resistance against physical and enzymatic degradation, miRNAs exhibiting deregulation in disease have been implicated as putative biomarkers [49]. Accordingly, a multitude of studies has been conducted to assess the utility of miRNAs with respect to the distinction of IBD from functional gastrointestinal disorders and health, or in the prediction of disease activity. As a result numerous candidates have been proposed to complement traditional serological and fecal biomarkers such as ASCAs, pANCAs and lactoferrin. However, the translatability of these findings into routine clinical practice remains to be proven [50].

The evident relationship between dysregulation of miRNAs and human diseases suggests a therapeutic application of these molecules. Accordingly, numerous (pre-)clinical studies have been conducted to investigate the directed manipulation of miRNA expression utilizing miRNA mimics or miRNA inhibitors [51]. Exemplary, MesomiR-1 (EnGeneIC), a miR-16 mimic encapsulated in an EGFR-targeted delivery vehicle, has been shown to induce substantial dose-dependent inhibition of tumor growth in human mesothelioma xenograft. A first clinical trial suggested comparable effects in patients with pleural mesothelioma (NCT02369198). Furthermore, MRG-201 (miRagen Therapeutics), a cholesterol-conjugated miR-29 mimic, has been reported to induce reversal of bleomycin-induced pulmonary fibrosis in mice. Likewise, a clinical study confirmed the reduction of scar tissue in patients with systemic sclerosis (NCT02603224). The treatment of Hepatitis C by Mirvirasen (Roche), a LNA-modified antimiR-122 represents one of the currently most promising applications of miRNA inhibitors. By impairing the mir-122-induced replication of HCV, the drug has been shown to reduce viral titres in mice and primates. With recent clinical trails this effect likewise could be confirmed in humans (NCT01872936). Finally, MRG-106 (miRagen Therapeutics), a LNA-modified antimiR-155, has been reported to induce tumor regression and apoptosis in mouse models of lymphoid malignancies. An ongoing clinical study in patients with various lymphomas and leukemia aims for the evaluation of safety, tolerability, pharmacokinetics and potential efficacy of the investigational drug (NCT02580552).

1.6 Summary of Article I

M. Hübenthal, S. Lipinski, A. Franke et al. "Molecular Genetics of Inflammatory Bowel Disease". In: ed. by M. D'Amato, C. Hedin, and J. D. Rioux. Springer Science+Business Media New York, 2019. Chap. MiRNAs in inflammatory bowel disease. (in press).

With this article we provide an in-depth review of current knowledge about miRNAs in the context of IBD. We describe the pivotal role of these molecules for the maintenance of intestinal homeostasis, focusing on their regulatory impact on mucosal immunity as well as epithelial autophagy and barrier integrity. We review recent biomarker research proposing candidate miRNAs as diagnostic tools predicting IBD-related traits. Ultimately, we report estimates of these candidates' performance as a foundation for qualitative classification of biomarkers and models being proposed in the course of this thesis.

1.7 Summary of Article II and related patents

M. Hübenthal, G. Hemmrich-Stanisak, F. Degenhardt et al. "Sparse Modeling Reveals miRNA Signatures for Diagnostics of Inflammatory Bowel Disease". In: PLOS ONE 10.10 (2015), e0140155.

In the course of this study we evaluated whether systematic miRNA expression profiling, in conjunction with machine learning techniques, is suitable as a non-invasive test for the major IBD phenotypes (Crohn's disease (CD) and ulcerative colitis (UC)). Based on microarray technology, expression levels of 863 miRNAs were determined for whole blood samples from 40 CD and 36 UC patients and compared to data from 38 healthy controls (HC). To further discriminate between disease-specific and general inflammation we included miRNA expression data from other inflammatory diseases (inflammation controls (IC): 24 chronic obstructive pulmonary disease (COPD), 23 multiple sclerosis, 38 pancreatitis and 45 sarcoidosis cases) as well as 70 healthy controls from previous studies. Multiclass classification problems considering 2, 3 or 4 of the afore-mentioned groups were solved employing different types of state-of-the-art machine learning methods (support vector machines, SVMs). The resulting models were assessed regarding sparsity and performance and a subset was selected for further investigation. Measured by the area under the ROC curve (AUC) the corresponding median holdout-validated accuracy was estimated as ranging from 0.75 to 1.00 (including IC) and 0.89 to 0.98 (excluding IC), respectively. In combination, the corresponding models provide tools for the distinction of CD and UC as well as CD, UC and HC with expected classification error rates of 3.1 and 3.3%, respectively. These results were obtained by incorporating not more than 16 distinct miRNAs. Validated target genes of these miRNAs have been previously described as being related to IBD. For others we observed significant enrichment for IBD susceptibility loci identified in earlier GWAS, suggesting that the proposed miRNA signature is of relevance for the etiology of IBD.

T. Brefort, A. Franke, G. Hemmrich-Stanisak et al. "MiRNAs as non-invasive biomarkers for inflammatory bowel disease". EP Patent 3212803B1 (granted). 2018.

T. Brefort, A. Franke, G. Hemmrich-Stanisak et al. "MiRNAs as non-invasive biomarkers for inflammatory bowel disease". US Patent 20170306407A1 (published). 2017.

T. Brefort, A. Franke, G. Hemmrich-Stanisak et al. "MiRNAs as non-invasive biomarkers for inflammatory bowel disease". WO Patent 2016066288A1 (published). 2016.

The miRNA signatures defined in the course of the afore-mentioned microarray study were utilized to complement/extend a set of patents filed at the European Patent Office (EPO), USPTO (United States Patent

and Trademark Office) and World Intellectual Property Organization (WIPO) to protect the application of miRNAs as non-invasive biomarkers for inflammatory bowel disease. The described inventions relate to the whole workflow of sampling/purifying miRNAs, quantifying their expression (employing hybridisation-, amplification-, sequencing-, cytometry- and/or spectroscopy-based approaches), deriving a mathematical model (employing regression, clustering and/or classification techniques) and applying this model to obtain a diagnostic prediction.

1.8 Summary of Article III

M. Hübenthal, S. Juzėnas, S. Nikolaus et al. "Comprehensive discriminative analysis of sequencing-based isomiR expression profiles reveals highly accurate tools for diagnostics of inflammatory bowel disease". (in submission).

With this study we provide the to-date most comprehensive discriminative analysis of sequencing-based isomiR expression profiles from whole blood, contributing to a more comprehensive understanding of inflammatory bowel disease. We examined isomiR expression profiles of 515 individuals, including 75 cases of untreated IBD (denoted as CD− and UC−, respectively), 271 cases of treated IBD (denoted as CD+ and UC+, respectively), 124 healthy and 45 symptomatic controls (denoted as HC and SC, respectively). We identified specific signatures of differentially expressed isomiRs, providing the foundation to derive mathematical classification models for discrimination of arbitrary combinations of these traits. For the first time, we do not only focus on 2-nomial classifiers (UC− vs. CD−, SC vs. CD−, SC vs. UC−, HC vs. CD−, HC vs. UC− and SC vs. HC as well as UC+ vs. CD+, HC vs. CD+ and HC vs. UC+) but likewise provide tools to solve 3-nomial (CD+ vs. UC+ vs. HC, CD− vs. UC− vs. HC and CD− vs. UC− vs. SC) as well as 4-nomial classification problems (CD− vs. UC− vs. SC vs. HC). To ensure biological interpretability we thereby account for unwanted variability of both biological and technical origin. Extensive sampling has been performed to obtain generalizable models but also a reliable evaluation of their clinical utility. In this way we do not only provide strong evidence for sequencing-based isomiR expression profiles being a valuable tools for diagnosing IBD but also for their superiority over established biomarkers. We furthermore detect treatment-dependent changes of the isomiR expression profiles, leading to improved classifier performance for treated patients. However, the translatability of the findings into routine clinical practice remains to be proven.

1.9 Summary of Article IV

S. Juzėnas, G. Venkatesh, M. Hübenthal et al. "A comprehensive, cell specific microRNA catalogue of human peripheral blood". In: Nucleic Acids Research 45.16 (2017), pp. 9290–9301.

In the course of this study we investigated the transcriptional profiles of miRNAs from seven distinct peripheral blood cell populations of 43 healthy donors employing next generation sequencing. These measurements were complemented by profiles of circulating miRNAs from whole blood (n = 77) as well as serum (n = 38) and blood-borne exosomes (n = 38). The resulting dataset represents our (still non-exhaustive, but substantial) contribution towards a reference of miRNA and isomiR expression in the cellular and non-cellular components of human peripheral blood. Pairwise differential expression analysis between the different cell types revealed 224 miRNA arms being significantly differentially expressed between the two major blood cell lineages (lymphoid versus myeloid). Furthermore, by focusing on the subtypes of lymphoid and myeloid cells, we found that 149 miRNAs were significantly deregulated between natural killer cells (CD56+) and

small lymphocytes (CD4+, CD8+ and CD19+), 119 miRNAs between T-lymphocytes (CD4+ and CD8+) and B-lymphocytes (CD19+) and 193 between CD14+ and CD15+ cells. Furthermore, we showed that cell type specificity of miRNA profiles can even be increased by considering miRNA variants (isomiRs). Thus, a comparison between lymphoid and myeloid lineages revealed 1862 differentially expressed isomiRs (from 281 miRNA arms). Analogous comparisons of the subtypes of lymphoid and myeloid cells displayed 1204 differentially expressed isomiRs (derived from 211 arms) between natural killer cells (CD56+) and small lymphocytes (CD4+, CD8+ and CD19+), 773 isomiRs (from 179 arms) between T-lymphocytes (CD4+ and CD8+) and B-lymphocytes (CD19+) and 1420 (from 246 arms) between CD14+ and CD15+ cells.

With this work we provide the most comprehensive contribution to date toward a complete miRNA inventory of healthy human blood. We identified blood cell specific transcription signatures at both the miRNA as well as the isomiR level. It might be reasonably assumed that these signatures are altered in the diseased state and that this alteration represents a valuable discriminative tool. Its impact on the predictive performance of IBD biomarkers might be investigated in the course of future studies.

References

[1]   S. C. Ng, H. Y. Shi, N. Hamidi, et al. "Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies." In: *Lancet (London, England)* 390.10114 (2018), pp. 2769–2778.

[2]   D. Ellinghaus, L. Jostins, S. L. Spain, et al. "Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci". In: *Nature Genetics* 48.5 (2016), pp. 510–518.

[3]   K. M. de Lange, L. Moutsianas, J. C. Lee, et al. "Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease". In: *Nature Genetics* 49.2 (2017), pp. 256–261.

[4]   M. Pierdomenico, L. Stronati, M. Costanzo, et al. "New Insights Into the Pathogenesis of Inflammatory Bowel Disease: Transcription Factors Analysis in Bioptic Tissues From Pediatric Patients". In: *Journal of Pediatric Gastroenterology and Nutrition* 52.3 (2011), pp. 271–279.

[5]   R. Häsler, M. Kerick, N. Mah, et al. "Alterations of pre-mRNA splicing in human inflammatory bowel disease". In: *European Journal of Cell Biology* 90.6 (2011), pp. 603–611.

[6]   P. Anderson, K. Phillips, G. Stoecklin, et al. "Post-transcriptional regulation of proinflammatory proteins". In: *Journal of Leukocyte Biology* 76.1 (2004), pp. 42–47.

[7]   S. S. Cao. "Epithelial ER Stress in Crohn's Disease and Ulcerative Colitis". In: *Inflammatory Bowel Diseases* 22.4 (2016), pp. 984–993.

[8]   F. Magro, P. Gionchetti, R. Eliakim, et al. "Third European Evidence-based Consensus on Diagnosis and Management of Ulcerative Colitis. Part 1: Definitions, Diagnosis, Extra-intestinal Manifestations, Pregnancy, Cancer Surveillance, Surgery, and Ileo-anal Pouch Disorders". In: *Journal of Crohn's and Colitis* 11.6 (2017), pp. 649–670.

[9]   F. Gomollón, A. Dignass, V. Annese, et al. "3rd European Evidence-based Consensus on the Diagnosis and Management of Crohn's Disease 2016: Part 1: Diagnosis and Medical Management". In: *Journal of Crohn's and Colitis* 11.1 (2017), pp. 3–25.

[10] C. Ott and J. Schölmerich. "Extraintestinal manifestations and complications in IBD". In: *Nature Reviews Gastroenterology & Hepatology* 10.10 (2013), pp. 585–595.

[11] G. R. Lichtenstein, E. V. Loftus, K. L. Isaacs, et al. "ACG Clinical Guideline: Management of Crohn's Disease in Adults". In: *The American Journal of Gastroenterology* 113.4 (2018), pp. 481–517.

[12] A. Kornbluth and D. B. Sachar. "Ulcerative Colitis Practice Guidelines in Adults: American College of Gastroenterology, Practice Parameters Committee". In: *The American Journal of Gastroenterology* 105.3 (2010), pp. 501–523.

[13] N. Dotan, R. Altstock, M. Schwarz, et al. "Anti-Glycan Antibodies as Biomarkers for Diagnosis and Prognosis". In: *Lupus* 15.7 (2006), pp. 442–450.

[14] C. H. Seow, J. M. Stempak, W. Xu, et al. "Novel anti-glycan antibodies related to inflammatory bowel disease diagnosis and phenotype." In: *The American journal of gastroenterology* 104.6 (2009), pp. 1426–34.

[15] M. Ferrante, L. Henckaerts, M. Joossens, et al. "New serological markers in inflammatory bowel disease are associated with complicated disease behaviour." In: *Gut* 56.10 (2007), pp. 1394–403.

[16] D. Simondi, G. Mengozzi, S. Betteto, et al. "Antiglycan antibodies as serological markers in the differential diagnosis of inflammatory bowel disease." In: *Inflammatory bowel diseases* 14.5 (2008), pp. 645–51.

[17] M. Papp, I. Altorjay, N. Dotan, et al. "New serological markers for inflammatory bowel disease are associated with earlier age at onset, complicated disease behavior, risk for surgery, and NOD2/CARD15 genotype in a Hungarian IBD cohort." In: *The American journal of gastroenterology* 103.3 (2008), pp. 665–81.

[18] F. Rieder, S. Schleder, A. Wolf, et al. "Association of the novel serologic anti-glycan antibodies anti-laminarin and anti-chitin with complicated Crohn's disease behavior." In: *Inflammatory bowel diseases* 16.2 (2010), pp. 263–74.

[19] M. Papp, I. Altorjay, G. L. Norman, et al. "Seroreactivity to microbial components in Crohn's disease is associated with ileal involvement, noninflammatory disease behavior and NOD2/CARD15 genotype, but not with risk for surgery in a Hungarian cohort of IBD patients." In: *Inflammatory bowel diseases* 13.8 (2007), pp. 984–92.

[20] A. G. Kaditis, J. Perrault, W. J. Sandborn, et al. "Antineutrophil cytoplasmic antibody subtypes in children and adolescents after ileal pouch-anal anastomosis for ulcerative colitis." In: *Journal of pediatric gastroenterology and nutrition* 26.4 (1998), pp. 386–92.

[21] R. K. Linskens, R. C. Mallant-Hent, Z. M. A. Groothuismink, et al. "Evaluation of serological markers to differentiate between ulcerative colitis and Crohn's disease: pANCA, ASCA and agglutinating antibodies to anaerobic coccoid rods." In: *European journal of gastroenterology & hepatology* 14.9 (2002), pp. 1013–8.

[22] K. Oberstadt, W. Schaedel, M. Weber, et al. "P-ANCA as a differential diagnostic marker in inflammatory bowel disease." In: *Advances in experimental medicine and biology* 371B (1995), pp. 1313–6.

[23] N. Yasuda, P. Thomas, H. Ellis, et al. "Perinuclear anti-neutrophil cytoplasmic antibodies in ulcerative colitis after restorative proctocolectomy do not correlate with the presence of pouchitis." In: *Scandinavian journal of gastroenterology* 33.5 (1998), pp. 509–13.

[24] M. Esteve, J. Mallolas, J. Klaassen, et al. "Factors related to the presence of IgA class antineutrophil cytoplasmic antibodies in ulcerative colitis." In: *The American journal of gastroenterology* 93.4 (1998), pp. 615–8.

[25] M. Esteve, J. Mallolas, J. Klaassen, et al. "Antineutrophil cytoplasmic antibodies in sera from colectomised ulcerative colitis patients and its relation to the presence of pouchitis." In: *Gut* 38.6 (1996), pp. 894–8.

[26] J. Aisenberg, J. Wagreich, J. Shim, et al. "Perinuclear anti-neutrophil cytoplasmic antibody and refractory pouchitis. A case-control study." In: *Digestive diseases and sciences* 40.9 (1995), pp. 1866–72.

[27]   P. L. Lakatos, I. Altorjay, T. Szamosi, et al. "Pancreatic autoantibodies are associated with reactivity to microbial antibodies, penetrating disease behavior, perianal disease, and extraintestinal manifestations, but not with NOD2/CARD15 or TLR4 genotype in a Hungarian IBD cohort." In: *Inflammatory bowel diseases* 15.3 (2009), pp. 365–74.

[28]   L. Prideaux, P. De Cruz, S. C. Ng, et al. "Serological Antibodies in Inflammatory Bowel Disease: A Systematic Review". In: *Inflammatory Bowel Diseases* 18.7 (2012), pp. 1340–1355.

[29]   J. A. Tibble, G. Sigthorsson, S. Bridger, et al. "Surrogate markers of intestinal inflammation are predictive of relapse in patients with inflammatory bowel disease." In: *Gastroenterology* 119.1 (2000), pp. 15–22.

[30]   C. M. T. Otten, L. Kok, B. J. M. Witteman, et al. "Diagnostic performance of rapid tests for detection of fecal calprotectin and lactoferrin and their ability to discriminate inflammatory from irritable bowel syndrome." In: *Clinical chemistry and laboratory medicine* 46.9 (2008), pp. 1275–80.

[31]   O. Schröder, M. Naumann, Y. Shastri, et al. "Prospective evaluation of faecal neutrophil-derived proteins in identifying intestinal inflammation: combination of parameters does not improve diagnostic accuracy of calprotectin." In: *Alimentary pharmacology & therapeutics* 26.7 (2007), pp. 1035–42.

[32]   A. M. Schoepfer, M. Trummler, P. Seeholzer, et al. "Discriminating IBD from IBS: comparison of the test performance of fecal markers, blood leukocytes, CRP, and IBD antibodies." In: *Inflammatory bowel diseases* 14.1 (2008), pp. 32–9.

[33]   S. Wang, Z. Wang, H. Shi, et al. "Faecal calprotectin concentrations in gastrointestinal diseases." In: *The Journal of international medical research* 41.4 (2013), pp. 1357–61.

[34]   A. M. Schoepfer, M. Trummler, P. Seeholzer, et al. "Accuracy of four fecal assays in the diagnosis of colitis." In: *Diseases of the colon and rectum* 50.10 (2007), pp. 1697–706.

[35]   J. Langhorst, S. Elsenbruch, J. Koelzer, et al. "Noninvasive markers in the assessment of intestinal inflammation in inflammatory bowel diseases: performance of fecal lactoferrin, calprotectin, and PMN-elastase, CRP, and clinical indices." In: *The American journal of gastroenterology* 103.1 (2008), pp. 162–9.

[36]   J. Dai, W.-Z. Liu, Y.-P. Zhao, et al. "Relationship between fecal lactoferrin and inflammatory bowel disease." In: *Scandinavian journal of gastroenterology* 42.12 (2007), pp. 1440–4.

[37]   U. Kopylov, G. Rosenfeld, B. Bressler, et al. "Clinical Utility of Fecal Biomarkers for the Diagnosis and Management of Inflammatory Bowel Disease". In: *Inflammatory Bowel Diseases* 20.4 (2014), pp. 742–756.

[38]   D. P. Bartel. "Metazoan MicroRNAs". In: *Cell* 173.1 (2018), pp. 20–51.

[39]   L. F. Gebert and I. J. MacRae. "Regulation of microRNA function in animals". In: *Nature Reviews Molecular Cell Biology* (2018), p. 1.

[40]   R. C. Lee, R. L. Feinbaum, and V. Ambros. "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14". In: *Cell* 75.5 (1993), pp. 843–854.

[41]   B. J. Reinhart, F. J. Slack, M. Basson, et al. "The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans". In: *Nature* 403.6772 (2000), pp. 901–906.

[42]   M. Lagos-Quintana. "Identification of Novel Genes Coding for Small Expressed RNAs". In: *Science* 294.5543 (2001), pp. 853–858.

[43]   N. C. Lau. "An Abundant Class of Tiny RNAs with Probable Regulatory Roles in Caenorhabditis elegans". In: *Science* 294.5543 (2001), pp. 858–862.

[44]   R. C. Lee. "An Extensive Class of Small RNAs in Caenorhabditis elegans". In: *Science* 294.5543 (2001), pp. 862–864.

[45] A. Kozomara and S. Griffiths-Jones. "miRBase: annotating high confidence microRNAs using deep sequencing data." In: *Nucleic acids research* 42.Database issue (2014), pp. D68–73.

[46] A. Mehta and D. Baltimore. "MicroRNAs as regulatory elements in immune system logic". In: *Nature Reviews Immunology* 16.5 (2016), pp. 279–294.

[47] J. Füllgrabe, D. J. Klionsky, and B. Joseph. "The return of the nucleus: transcriptional and epigenetic control of autophagy". In: *Nature Reviews Molecular Cell Biology* 15.1 (2014), pp. 65–74.

[48] M. Vancamelbeke and S. Vermeire. "The intestinal barrier: a fundamental role in health and disease". In: *Expert Review of Gastroenterology & Hepatology* 11.9 (2017), pp. 821–834.

[49] P. S. Mitchell, R. K. Parkin, E. M. Kroh, et al. "Circulating microRNAs as stable blood-based markers for cancer detection". In: *Proceedings of the National Academy of Sciences* 105.30 (2008), pp. 10513–10518.

[50] A. Kappel and A. Keller. "miRNA assays in the clinical laboratory: workflow, detection technologies and automation aspects". In: *Clinical Chemistry and Laboratory Medicine (CCLM)* 55.5 (2017).

[51] R. Rupaimoole and F. J. Slack. "MicroRNA therapeutics: Towards a new era for the management of cancer and other diseases". In: *Nature Reviews Drug Discovery* 16.3 (2017), pp. 203–221.

2 Publications

MicroRNAs and inflammatory bowel disease

Matthias Hübenthal[1], Andre Franke[1], Simone Lipinski[1], Simonas Juzėnas[1,*]

[1] Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, Kiel, Germany
* To whom correspondence should be addressed

Introduction

MicroRNAs (miRNAs; non-coding RNAs, approximately 22 nucleotides in length) represent a class of post-transcriptional regulators of gene expression. Individual miRNAs regulate specific mRNA targets recognized by the RNA-induced silencing complex (RISC). The formation of this complex corresponds to the end point of a multi-step biogenerative process [1]. Variability in miRNA biogenesis introduced by alternative precursor cropping and dicing, sequence editing or trimming as well as addition of nucleotides, gives rise to miRNA isoforms (isomiRs) [2]. Being implicated in numerous physiologic processes such as immune cell development and regulatory circuits of cell functions, miRNAs (and their variants) are also believed to play a critical role in the etiology of a variety of diseases [3, 4]. Imbalance of the regulatory processes within immune system development and function lead to an increased production of cytokines being involved in the development of inflammation. Relapsing as well as persisting inflammation, in turn, is a characteristic feature of inflammatory bowel disease (IBD), including its subtypes Crohn's disease (CD) and ulcerative colitis (UC) [5]. The pivotal role of miRNAs in regard to the development of the involved innate and adaptive immune cells as well as their response to inflammatory signals are further key points of this chapter.

Recent studies provide evidence for miRNAs modulating intestinal homeostasis by calibrating immune cell responses, and also regulating autophagy (lysosomal degradation of cytoplasmic content) by direct or indirect impairment of the process of vesicle elongation [6]. Furthermore, they interfere with the formation of tight junctions (affecting paracellular permeability), adherent junctions (affecting calcium-dependent adhesion of adjacent epithelial cells and spatial organization actin filaments) as well as desmosomes (affecting calcium-

independent hyperadhesion of adjacent epithelial cells) and are therefore critical for maintaining epithelial integrity and barrier function [7].

Establishing IBD diagnostics according to the current consensus of the European Crohn's and Colitis Organization (ECCO) [8, 9] or the American College of Gastroenterology (ACG) [10, 11] incorporates the assessment of numerous endoscopic, histological, radiological and/or biochemical parameters. Nonetheless, a definite differential diagnosis cannot be established in approximately 10%–17% of colitis patients [12], and more than 10% of IBD patients change diagnosis during the first two years of the disease course [13]. In recent past, miRNAs emerged as a new class of putative biomarkers detectable in a wide range of tissues and body fluids [14]. In the context of IBD, a multitude of studies has been conducted to assess the utility of miRNAs with respect to the distinction of IBD from functional gastrointestinal disorders and health, or in the prediction of disease activity. Numerous candidates have been proposed to complement traditional serological and fecal biomarkers such as ASCAs, pANCAs and lactoferrin. However, further research is needed to validate these findings and translate them into clinical practice [15].

MicroRNA biogenesis and regulatory roles

The canonical miRNA biogenesis is a multi-step process initiated by RNA polymerase II (Pol II)-mediated transcription, which are encoded as individual genes (monocistronic), as gene clusters (polycistronic) or in introns of host genes (intronic) [16]. In mammals, canonical miRNAs are transcribed as a part of much longer RNAs called primary miRNA (pri-miRNA) transcripts which typically contain 5' cap, poly-A tail and hairpins wherein miRNA sequences are embedded [1, 17]. The pri-miRNA transcripts are recognized by a nuclear protein complex referred to as microprocessor, which processes the latter into single hairpins called precursor miRNAs (pre-miRNAs). The microprocessor is a heterotrimeric complex containing one Drosha endonuclease and two DGCR8 subunits [1, 16]. The DGCR8 subunits recognize the hairpin of pri-miRNA, while Drosha cuts the stem with a two-nucleotide 3' overhang and releases approximately 60 nucleotide-long, stem-loop shaped pre-miRNA. After export to the cytoplasm by Exportin 5, the pre-miRNA is further processed by the Dicer nuclease, which in co-operation with TARBP or PACT, based on distance from 5' end terminus of dsRNA, cleaves the hairpin loop to liberate the 21–24 base pair miRNA duplex. After the cleavage, the duplex has a two-nucleotide-long 3' overhang on each end, resulting from the offset cuts made by both Drosha and Dicer nucleases [1, 17]. Once generated, the small RNA duplex is subsequently loaded onto an Argonaute protein (AGO), where it is unwound to form a RNA-induced silencing complex (RISC), the effector unit that recognizes target mRNA molecules. The unwinding of the miRNA duplex is promoted by mismatches in the guide strand at nucleotide positions 2–8 and 12–15. The guide strand is selected based on relative thermodynamic stability of the

MiRNAs in inflammatory bowel disease   3

two ends of the small RNA duplex. Accordingly, the strand with the least stable 5′ terminus is typically selected as the guide strand [17]. Crystal structures of the human RISC with and without target RNAs suggest a stepwise mechanism, in which AGO2 primarily exposes guide bases 2–5 for initial target pairing, which then promotes conformational changes that expose not only the seed sequence (2–8), but also bases 13–16 for further target recognition [2].

Since the final maturation of miRNAs takes place in the cytoplasmic space, where miRNAs silence gene expression via translational repression of targeted mRNAs, cytosol contains the highest abundance of functional RISCs. However, mature miRNAs also reside in multiple subcellular locations, such as nucleus and mitochondria, and can be secreted out of cells via exosomes into various types of body fluids and even into the gut lumen [18, 19]. Recent studies have shown that mature miRNAs localized within the nucleus can be involved in epigenetic regulation, whereas miRNAs localized in mitochondria can directly enhance translation of mitochondrial transcripts [20, 21]. Interestingly, as in the case of mitochondria, a recent study reported that host miRNAs can directly modulate gut microbiome composition by selectively promoting bacterial growth [19].

Although public databases catalogue miRNAs as single sequences, an individual miRNA hairpin arm can give rise to multiple distinct isoforms (isomiRs) which are referred to as the mature miRNA transcripts, and which can differ in sequence length and composition [22]. Such sequence divergence occurs due to alternative precursor cropping and dicing, sequence editing, end trimming or the addition of non-templated nucleotides [23]. Depending on the type of variation, isomiRs are categorized into three main classes: 5′, 3′ and polymorphic (internal) isomiRs. Additionally, 5′ and 3′ isomiRs are sub-classified into templated (trimming) or non-templated (tailing) variants [2, 24]. The primary source of templated isomiRs is alternative cleavage during the Drosha and Dicer processing steps. Cleavage by Dicer is regulated by a mechanism based on binding competition between TARBP and PACT which, in the end, tune the length of miRNAs [25]. The alternative cleavage by Dicer directly modulates the seed sequence of the 3p miRNA arm, whereas alternative cleavage by Drosha shifts the seed site of the 5p arm [2]. Due to the altered seed sequence, 5′ templated variants can broaden the binding site repertoire of the respective reference miRNA and often together regulate distinct or shared targets involved in the same biological pathways [4]. Polymorphic isomiRs are usually formed via sequence editing by RNA editing enzymes such as ADARs, which catalyze hydrolytic conversion of adenosine (A) to inosine (I) in pre-miRNA molecule [23]. Templated 3′ isomiRs can also be formed after maturation by exoribonucleases, which trim nucleotides from the 3′ end and produce shorter isoforms [26]. Another post-maturation sequence modification of 3′ isomiRs is the non-templated nucleotide addition by terminal nucleotidyl transferases (TUTs), which extend mature sequences by one or more nucleotides. 3′ isomiRs are the most frequently observed types of isomiRs and potentially have an effect on miRNA cellular or extracellular localization and

turnover [2]. Very often these variants are expressed in a cell type specific manner [27, 28].

Role of miRNAs in immune system

miRNAs are important for cell fate decisions in the immune system. By providing positive or negative feedback of gene expression, miRNAs also regulate cell signaling pathways during innate and adaptive immune responses [3]. Imbalance within these regulatory processes may lead to increased pro-inflammatory cytokine production and T cell proliferation and/or recruitment in the affected tissue. This in turn may lead to a relapsing and remitting course and persisting low-grade inflammation, which characterizes chronic inflammatory disorders, including IBD [5]. In this section, we will describe and summarize recent findings about the involvement of miRNAs in the regulation of innate and adaptive immunity.

Innate Immunity and miRNAs

The innate immune system is predominantly composed of myeloid lineage cells and it represents the first line of defense against infection by external pathogens such as bacteria, fungi and viruses [3]. These pathogens are mainly recognized by tissue-residing macrophages and dendritic cells (DCs) via pattern recognition receptors (PRRs), which bind to pathogen-associated molecular patterns (PAMPs) [29]. The innate immune system is maintained by extensive regulatory networks, in which miRNAs play an important role as regulatory elements in the development and function of immune cells (**Figure 1**) [3].

Innate cell development

Myeloid lineage development begins with common myeloid progenitor (CMP) cells, which give rise to megakaryocyte-erythroid progenitors (MEPs) and granulocyte-macrophage progenitors (GMPs). In mammals, CMP transition to GMP or MEP depends on miR-29a, which negatively regulates Dnmt3a translation [30]. The decision whether the GMPs will become granulocytes or monocytes also remarkably relies on miRNAs, i.e. during the granulocyte differentiation highly expressed miR-223 inhibits MEF2C (myocyte-specific enhancer factor 2C), which is a suppressor of granulocyte differentiation [3]. Additional miRNAs such as miR-21 and miR-196b also influence GMP fate. Enforced expression of these two molecules blocks granulocyte differentiation, whereas enforced expression of miR-424 promotes monocytic differentiation [3, 31]. Monocytes can differentiate into DCs or into macrophages and this transition is influenced by miR-34a-5p and miR-

MiRNAs in inflammatory bowel disease   5

21-5p, which repress the mRNAs encoding WNT1 (Wnt Family Member 1) and JAG1 (Jagged 1), respectively, to promote DC differentiation [31].

In another myeloid branch, progenitors (MEPs) give rise to anucleate erythrocytes and megakaryocytes, the latter of which produce anucleate platelets. The core transcription factor of erythropoiesis GATA1 directly targets promoters of miR-451a and miR-486 genes, and initiates their Pol II-mediated transcription [32, 33]. A recent study utilizing knockout mouse strains showed that these two miRNAs are necessary for erythroid development and maturation [34]. The fate of megakaryocytes, in turn, depends on several miRNAs, including miR-130a, miR-150, miR-34a and miR-10a [3, 35, 36]. For example, miR-10a-5p and miR-150-5p drive MEP differentiation to the megakaryocyte lineage by suppressing transcripts of *HOXA1* [37] and *MYB* [38] genes, respectively. Although megakaryocyte-derived platelets are anucleate, like erythrocytes, they contain a substantial number of miRNA species and some of those, including miR-96-5p, miR-200b-3p, miR-495-3p, miR-107 and miR-223-3p are highly abundant and critically involved in platelet reactivity, aggregation, secretion and adhesion. [39].

Innate cell response

The activation of innate immune response starts once PAMPs bind to DC or macrophage PRRs such as Toll-like receptors (TLRs) and Nod-like receptors (NLRs), which bind extracellular and intracellular microbial molecules, respectively. These receptors, in turn, recruit adaptor proteins that activate NF-⍰B (nuclear factor kappa-light-chain-enhancer of activated B-cells) and MAPKs (mitogen-activated protein kinases) for the induction of inflammatory cytokine and antigen presentation (*MHC*; major histocompatibility complex) genes [40, 41]. The regulation of the aforementioned biological cascade highly relies on miR-155 and miR-146a molecules, which are two of the most relevant and most studied miRNAs in the field [42, 43]. Both miRNAs are induced in DCs and macrophages by NF-⍰B in response to TLR or pro-inflammatory cytokine signals [3]. In monocytes, miR-146a-5p functions as an anti-inflammatory regulator by providing a negative feedback loop to decrease NF-⍰B via suppression of TRAF (TNF receptor associated factor 1) and IRAK (interleukin 1 receptor associated kinase) translation [44]. On the other hand, miR-155-5p acts as pro-inflammatory regulator by repressing the production of SHIP1 (SH-2 containing inositol 5' polyphosphatase 1) and SOCS1 (suppressor of cytokine signaling-1), which leads to the increase of NF-⍰B activity and promotion of macrophage transition to the pro-inflammatory M1 phenotype [45, 46]. During macrophage activation, there is a temporal imbalance between *miR-155* and *miR-146a* gene expression. Activated by inflammatory stimulus, NF-⍰B immediately induces miR-155 expression, which leads to even stronger NF-⍰B activity. At the same time, miR-146a levels rise to negatively regulate NF-⍰B activity, leading to inhibition of miR-155 expression and resolution of the inflammatory response [46]. Besides these two molecules, there are several other

miRNAs, which increase or decrease during macrophage polarization from anti-inflammatory M2 to pro-inflammatory M1 phenotype. More specifically, miR-26a/b, miR-27a/b, miR-127, miR-101, let-7e and miR-125b have been shown to promote M1 polarization while miR-21, miR-124, miR-223, let-7c, miR-34a, miR-146b and miR-125a have been shown to induce M2 polarization in macrophages by targeting various transcription factors and adaptor proteins [45, 47–50]. As in macrophages, miR-155 is also involved in DC TLR-evoked polarization to a pro-inflammatory milieu. Other miRNAs such as miR-146a and miR-21 are involved in the negative regulation of DC activation [51]. After activation of the NOD2 (nucleotide-binding oligomerization domain-containing protein 2) pathway DCs upregulate miR-29 expression, which in turn acts as an anti-inflammatory miRNA by directly suppressing the production of the cytokine IL-23 [52]. Interestingly, miRNA let-7c which has been shown to promote the resting M2 phenotype of macrophages, was reported to act as a positive regulator of DC activation [53].

Adaptive immunity and microRNAs

The adaptive immune system is comprised of lymphoid lineage cells, including T and B lymphocytes. These cells provide a second line of immune defense against specific pathogens after priming from the innate immune system (in case of T cells) or from the primary pathogen-derived molecules (in case of B cells) [3]. PRRs of innate immune cells determine the origin of antigens recognized by T cell receptor (TCR), as well as resolve the type of detected infection, and guide T-lymphocytes to induce the proper effector class of the immune response [54]. B Cell Receptor (BCR)-activated B cells, on the other hand, provide humoral immunity against extracellular pathogens themselves, through antibody production [55]. Similar to the innate immune system, the adaptive immune system is under tight control by complex regulatory networks including a number of transcriptional factors, cytokines and miRNAs, the latter of which are regulating gene-expression programs during development and function of the adaptive immune cells (**Figure 2**).

Adaptive cell development

Adaptive immune system development is a hierarchical and lineage-restricted differentiation process starting with the common lymphoid progenitor (CLP), which resides in bone marrow and can give rise either to precursor-progenitors of B cells (pre-pro-B cells) or to progenitors of T cells called double negative (DN) cells [3, 31]. The fate determination whether the CLP will become pre-pro-B cell or DN involves of miR-181a and miR-126 molecules. The increased expression of these two miRNAs drive CLPs toward B-lymphopoiesis [56, 57]. The further development of B lymphocytes is a rather linear and stepwise process, which ensures that each B lymphocyte acquires a single, unique B cell antigen

MiRNAs in inflammatory bowel disease   7

receptor (BCR) [58]. The process involves multiple transitional stages, such as progenitor B (pro-B), precursor B (pre-B) and mature B cells, in which miRNAs have a remarkable impact. Lineage specification from pre-pro-B to pro-B cells is controlled by miR-212-3p and miR-132-3p through the suppression of SOX4 (SRY-box 4) translation [59]. Another checkpoint of quality control is introduced by miR-34a, miR-150 and the miR-17~92 cluster. Both miR-34a-5p and miR-150-5p act to inhibit pro-B cell to pre-B cell transition by targeting FOXP1 (forkhead box P1) and MYB, respectively, whereas the miR-17~92 cluster acts to promote the pro-B to pre-B cell conversion by negatively regulating the pro-apoptotic protein BIM (Bcl-2 interacting protein) [3]. A recent study has shown that epigenetic silencing of *miR-125b* is required for normal B-cell development and that the overexpression of this miRNA leads to inefficient appearance of immature B cells in peripheral blood and also induces pre-B-cell leukemia [60]. In contrast, miR-148a, which is highly expressed in pro- and pre-B cells, is gradually down-regulated following their transition into immature and then into mature B cells [61]. A study in mice has shown that miR-148a might be critical for the regulation of B cell apoptosis and survival by targeting Gadd45a (growth arrest and DNA damage inducible alpha), Bim, and Pten (phosphatase and tensin homolog) proteins [62].

As introduced above, T-lymphopoiesis starts when bone marrow-residing CLPs differentiate into early stage DN cells, which then migrate to the thymus, where they undergo multiple stages of differentiation, T cell receptor (TCR) gene rearrangement and selection. The different stages include double positive (DP) and single positive (SP), the latter of which give rise to naïve CD4+ or naïve CD8+ cells [63]. To make the transition from DN cells to DP cells happen, the miR-17~92 cluster has to be expressed above a certain threshold, because the loss of its expression leads to reduced surface expression of IL-7R, resulting in decreased DN to DP transition [64]. Another miRNA having an influence on this transition is miR-150-5p, which inhibits early T-lymphopoiesis by downregulating NOTCH3 [65]. Of note, these miRNAs were already described as being involved in B-lymphopoiesis, where they positively affect cell differentiation as in T-lymphopoiesis. However, this effect is achieved by targeting different mRNAs. A further step in the thymocyte differentiation is the transition from DP to SP cells, where miR-181a plays an important role by regulating TCR sensitivity in DP cells during positive and negative selection in the thymus [66].

Adaptive cell response

Once exposed to an antigen, mature B cells become activated and further proliferate in germinal centers and finally differentiate into plasma cells [58]. The plasma cells, in turn, are critical for mediating humoral immune responses through the production of high-affinity antibodies [67]. During the activation of mature B cells, miR-155 is upregulated and promotes further proliferation via direct regulation of PU.1 (hematopoietic transcription factor), which has an important

role in antigen-driven B-cell maturation [68]. Besides that, miR-155 together with miR-181b, miR-210 and miR-217 are involved in the regulation of BCR signaling, antibody production and proliferation of the activated B cells. miR-146a-5p, together with miR-142-3p, prevents uncontrolled B cell activation trough repression of NF-κB and BAFF-R (B-cell activating factor receptor), respectively [3]. During the terminal stage of B cell differentiation, in an incoherent feed-forward loop regulated by STAT3 (signal transducer and activator of transcription 3) and BLIMP-1 (PR domain zinc finger protein 1), miR-21 provides a temporally delayed repression of plasma cell differentiation to prevent dysregulated proliferation [69]. miR-155 also plays a role in plasma cell fate commitment, where it acts on BSAP (B-cell-specific activator protein) via the miR-155/PU.1 axis, thus promoting efficient terminal B cell differentiation [70].

The function of CD8+ T cells is to destroy cells presenting exogenous antigens on MHC I class molecules; for example, cytotoxic T (T_C) cells kill virally infected cells, preventing them from being the source of more viral pathogen [71]. Differentiation of naïve CD8+ T cells into T_C cells is driven by antigen exposure and inflammation, which leads to the induction of specific gene-expression programs that also include miRNAs [66]. One example is the IL-2 signaling network, which is induced during the differentiation of T_C cells. The IL-2 pathway induces the expression of perforin, granzymes, and effector cytokines, while down-regulating miR-139-5p and miR-150-5p, which target perforin and IL-2R, respectively [72]. In contrast, miR-155-5p is induced in effector CD8+ T cells and promotes cell proliferation and differentiation through the repression of SOCS1 [73].

After activation, CD4+ T cells can differentiate into different subsets of T helper (T_H) cells, such as T_H1, T_H2 and T_H17 cells, regulatory T (T_{reg}) cells and T follicular helper (T_{FH}) cells. Each of the subsets has specific functions and distinctive gene expression programs, and interestingly, they possess a fair amount of plasticity to differentiate into different T_H cell subsets [66]. T_H1 cells are responsible for immune responses against intracellular viral and bacterial pathogens and tumors. T_H2 cells target helminths and facilitate tissue repair, and their deregulated response can cause allergy or asthma. T_H17 cells mediate immune responses against extracellular bacteria and fungi, and their dysregulated response can induce many organ-specific autoimmune diseases [74]. T_{FH} are involved in the elimination of numerous pathogens such as viruses, bacteria and fungi, by forming immunological memory and providing help to B cells [75]. T_{reg} cells, in turn, comprise a distinct anti-inflammatory lineage subpopulation of T CD4+ cells. They are essential mediators of immune tolerance, which actively suppress the function of conventional T_H cells and limit inappropriate or excessive responses of the immune system [76]. The differentiation of IL-17-producing T_H17 cells is promoted by numerous miRNAs, including miR-155, miR-21, miR-181c, miR-301a, miR-326 as well as clusters miR-183~96~182 and miR-212~132 [77–83]. For example, the pro-inflammatory miR-155-5p together with miR-326 play an important role in the T_H17 commitment by targeting ETS1 [77, 84]. Likewise, miR-

155-5p alone regulates the T_H17/T_{reg} ratio by targeting SOCS1 [85]. T cell-specific miR-17~92 cluster deficiency reduces T_H17 differentiation and ameliorates symptoms of experimental autoimmune encephalomyelitis (EAE), thus showing the importance of the cluster in promoting T_H17 responses [86], even though individual miRNAs of the cluster might have opposing effects on T_H17 differentiation [87]. On the other hand, miR-146a has been reported as a negative regulator of T_H17 differentiation. A study of miR-146a-dificient mice has shown that 2D2 T cells produce higher levels of autocrine IL-6 and IL-21, contributing to enhanced T_H17 responses and development of severe EAE [88]. The transition towards T_H1 cells is also endorsed by miR-155 and miR-17~92 cluster [89, 90]. As in T_H17 cells, miR-155 promotes the commitment of T_H1 cells by repression of SOCS1 transcripts [90]. Similarly, the miR-17~92 cluster promotes T_H1 commitment by blocking apoptosis, increasing IFN-γ (interferon gamma) production, and obstructing inducible T_{reg} cell differentiation [89]. In contrast, miR-29 acts as a negative regulator of T_H1 cell differentiation by suppressing IFN-γ production [66]. The differentiation of IL-4-producing T_H2 cells is also promoted by the miR-17~92 cluster. It has been shown that one of the members, miR-19a, endorses T_H2 cytokine production by direct targeting of PTEN, SOCS1, and TNFAIP3 (TNF-alpha-induced protein 3) [91]. Conversely, miR-24 and miR-27a limit T_H2 cell differentiation through inhibition of IL-4 production [92]. The development and function of T_{reg} cells is highly supported by miR-155, miR-10a, miR-146a, miR-146b and the miR-17~92 cluster [93–97]. Highly expressed in T_{reg} cells, miR-155 has been reported to be involved in the differentiation of these cells [98], as well as in the maintenance of competitive fitness [93]. Another highly abundant molecule, miR-146a, has been shown to be critical for T_{reg} suppressor function [94]. Its family member, miR-146b, which is a part of TRAF6-NF-B-FoxP3 signaling pathway, has been demonstrated to repress T_{reg} cell survival, proliferation, and suppressor function [95]. miR-17, a member of the miR-17~92 cluster, has been reported to obstruct the differentiation of T_{reg} cells [89]. However, another study showed that T_{reg}-specific loss of miR-17~92 expression results in a failure of T_{reg} accumulation and function during an acute EAE *in vivo* [97], thus confirming its importance for the function of T_{reg} cells. The TGF-β (transforming growth factor beta) and retinoic acid inducible miR-10a plays a role in positively maintaining the stability of inducible T_{reg} cells and preventing their phenotypic conversion into follicular helper T cells [96]. T_{FH} cells themselves are also highly regulated by miRNAs, including the well-described miR-155, miR-146a and miR-17~92 cluster [99]. The miR-17~92 cluster is a critical regulator of T cell-dependent antibody responses, T_{FH} cell differentiation and the fidelity of the T_{FH} cell gene-expression program [100]. Pro-inflammatory miR-155 is required for the development and expansion of T_{FH} cells. In contrast to this, miR-146a has an opposing function to T_{FH} cell development and restricts spontaneous expansion of T_{FH} cells [99].

Interestingly, upon naïve CD4 T-cell activation, besides up- or down-regulation of particular miRNAs, all miRNA sequences bearing 3'-terminal uridines

10   Hübenthal, Franke, Lipinski, Juzėnas

are specifically degraded and this decrease is accompanied by the down-regulation of uridyltransferases, including TUT4 and TUT7 [101], thus again highlighting the importance of miRNA regulation in T cell development and function.

Role of miRNAs in Inflammatory Bowel Disease

IBD belongs to the group of multifactorial chronic inflammatory disorders and affects the gastrointestinal tract. Its main types CD and UC are associated with various pathogenic phenotypes, including broadly dysregulated immune responses, disturbed autophagy and increased permeability of the epithelial barrier [102]. In the following section, we summarize the available evidence regarding miRNA involvement in the aforementioned pathological processes.

Role of miRNAs in Disruption of Gut Immune Homeostasis

The human gut immune system consists of innate and adaptive immune cells as well as non-immune cells scattered all over the intestinal layers. These cells sustain a balance between immune tolerance and rapid activation [5]. Immune balance is maintained through the combination of activating and inhibitory signaling pathways, including secretion of cytokines and activation of signal transducers, where miRNAs play an important regulatory role [43]. Dysregulation of these processes leads to inappropriate host immune responses to commensal bacteria. This, in turn, causes a continuous inflammation in genetically susceptible [103, 104] and environmentally conditioned individuals [105], which is a hallmark of multiple complex diseases, including IBD (**Figure 3**).

In genetically predisposed individuals, environmental factors such as diet or smoking can alter the composition of the gut microbiota, reduce the mucus layer and change the permeability of the intestinal epithelium, resulting in the increased uptake of luminal antigens [105, 106]. Once exposed to these antigens, DCs or macrophages convert from a tolerogenic state to an activated phenotype. In addition to producing pro-inflammatory cytokines (such as TNF, IL-6, IL-12 and IL-23) [107], TLR-NF-⬚B pathway activated cells also induce the expression of miR-155 and miR-146a, which, as described in the previous section, create a combined positive and negative feedback loop to control NF-⬚B activity itself [46]. The up-regulation of miR-155 in macrophages and DCs leads to increased production of the aforementioned pro-inflammatory cytokines [108, 109], which are usually found to be elevated in these cells during IBD pathogenesis [107]. Recent studies report the up-regulation of miR-155 in the IBD-affected tissues [110, 111] and support its functional importance in macrophages under IBD conditions [112, 113]. Another TLR-induced miRNA, anti-inflammatory miR-146a is also commonly found to be up-regulated during active colon inflammation [110, 114], and, in contrast to miR-155, it acts as a suppressor of inflammation and conditions DCs as well as

MiRNAs in inflammatory bowel disease 11

macrophages towards tolerogenic phenotype by suppressing NF-⬚B activity [44]. However, it has been shown that enforced miR-155 expression overrides miR-146a-mediated repression of NF-⬚B activation, thus emphasizing the dominant pro-inflammatory function of miR-155 and possibly explaining why the active state of macrophages persists after stimulation, even in the presence of large quantities of miR-146a [46]. Importantly, mouse studies have shown that deletion of miR-146a leads to aberrant myeloproliferation [115–117], autoimmunity [116] and even chronic inflammatory stress-related tumorigenesis [115, 117], thus suggesting the miRNA's critical role in TNF signaling regulation. The evidence on miR-146a down-regulation in peripheral blood mononuclear cells during lupus erythematosus additionally supports these findings [118]. TLR-activated NF-⬚B also prompts the expression of another anti-inflammatory miRNA, miR-9, which in a negative feedback loop via NF-⬚B1 (nuclear factor kappa B subunit 1) suppresses NF-⬚B dependent inflammatory responses [119], and inhibits activation of the NLRP3 inflammasome, most likely through the JAK1/STAT1 signaling pathway [120]. Although miR-9 has been reported to be up-regulated under active CD and UC conditions [121, 122] (most likely because of TLR stimulation), its methylation levels have been shown to be increased in the rectal mucosa of UC patients and even more increased in CRC patients when compared with healthy individuals [123], thus showing its importance in the maintenance of inflammation and tumorigenesis. Another TLR-induced miRNA in macrophages, miR-124, negatively modulates TLR inflammatory activity itself, by targeting multiple components of the TLR signaling cascade [124]. Interestingly, miR-124 has been reported to be down-regulated in the tissues of pediatric UC due to promoter hypermethylation, but not in the cohort of adult UC patients, reflecting the mechanistic differences between adult and pediatric UC [125]. The stimulation of intestinal DCs through TLR signaling also leads to decreased expression of miR-10a, which was observed in the IBD inflamed mucosa [126]. In human DCs, anti-inflammatory miR-10a reduces the production of IL-12 and IL-23, which in turn suppress IBD T_H1 and T_H17 cell immune responses. Moreover, miR-10a directly targets NOD2, which has a central role in intracellular pathogen sensing [127], and is highly relevant for IBD pathogenesis [128]. The genetic association between CD and NOD2 has been consistently replicated at the genome-wide level [129]. It has been shown that in human DCs, NOD2 induces miR-29 expression to limit IL-23 release, and that DSS-induced colitis was more severe due to the elevated IL-23 and T_H17 cell signature cytokines in the intestinal mucosa of miR-29-deficient mice. Moreover, CD patient DCs expressing NOD2 polymorphisms failed to induce miR-29 upon PRR stimulation and showed enhanced release of IL-12p40 [52]. The differentiation and function of intestinal macrophages and DCs is also calibrated by anti-inflammatory miR-223, which directly targets pro-inflammatory transcription factor C/EBPß [130]. Another study has shown that in addition to miR-9, miR-223 plays an important role in regulation of the myeloid-specific NLRP3 inflammasome during intestinal inflammation [131]. Although miR-223 has been repeatedly reported to

be elevated in active IBD conditions [131, 132], its deficiency results in severe DSS-colitis in mice [130, 131].

After processing of antigens, macrophages and DCs present them to naïve CD4+ T-cells, promoting differentiation into effector T cells, including T_H1, T_H2, T_H17 and T_{reg} cells. The intestinal inflammatory infiltrate of CD patients contains elevated numbers of T_H1 and T_H17 cells, whereas the mucosa of UC patients exhibits disturbed homoeostatic balance towards T_H2 and T_H17 cells [133, 134]. The increase of T_H1 and T_H17 cells during CD pathogenesis might be affected by miR-155, miR-29, miR-326 as well as the miR-17~92 cluster. Overexpression of miR-155 forces naïve CD4+ T cells to differentiate into T_H1 and T_H17 cells [90]. The role of miR-155 in the IBD T_H1/T_H17 pathway has been investigated in more detail using knockout, indicating that the loss of miR-155 leads to protection from advanced DSS-induced colitis development, and results in decreased T_H1/T_H17, CD11b+, and CD11c+ cells. This decrease, in turn, correlated with reduced clinical scores and severity of disease [135]. Besides the function in DCs, miR-29 has been shown to inhibit IFN-γ production and T_H1 differentiation via direct repression of T-bet and Eomes [136, 137]. Therefore, the down-regulation of miR-29 in T cells might likewise play a role in the IBD pathogenesis via the T_H1/T_H17 axis, even though the expression data in the IBD affected mucosa and serum shows up-regulation in the active disease state [122, 138, 139]. The T_H17 cell-promoting miRNA, miR-326 was found to be up-regulated in inflamed mucosa, thus suggesting its possible role in IBD pathogenesis [140]. Other T_H1/T_H17 response-promoting and T_{reg} differentiation-preventing miRNAs include members of the miR-17~92 cluster, which are found to be commonly deregulated in IBD [122, 138], and might be involved in the pathogenesis via T_H1/T_H17 or T_H17/T_{reg} pathways. However, to date, there are no functional studies that provide direct evidence for miR-17~92 modulating the IBD T_H1/T_H17 response pathway.

The increase of T_H2 cells during UC pathogenesis might be affected by miR-21, which has been repeatedly shown to be up-regulated in IBD, especially in UC [122, 141, 142]. Interestingly, in the colonic T cells during the recovery stages of UC, the expression of miR-21 has been found to be down-regulated, thus suggesting that miR-21 may play a role in limiting pathogenic T-cell responses [143].

During the pathogenesis of IBD, intestinal T_{reg}/T_H17 cell balance is altered and shifted towards pro-inflammatory T_H17 cells [144]. A recent study employing a murine IBD model, showed that TNFα-mediated loss of T_{reg} suppressive function coincides with induction of miR-106a via NF-◻B promoter-binding to suppress post-transcriptional regulation of IL-10 release, and that selective inhibition of miR-106a leads to enhanced T_{reg} function *in vitro* [145]. Another miRNA that might play a role in IBD T_{reg}/T_H17 balance regulation is miR-31, which was found to be up-regulated in IBD [146]. A recent study that investigated knockout mice in the EAE model has shown that miR-31 targets RAI3 (retinoic acid-inducible protein 3) and conditional deletion of this miRNA results in enhanced induction of peripheral T_{reg}

MiRNAs in inflammatory bowel disease   13

cells [147], thus showing its importance in T_{reg} cell regulation. T_{reg} suppressor function is likewise controlled by miR-146a. As shown in a study by Lu et al., miR-146a contributes to T_{reg}-mediated control of T_H1 responses via suppression of Stat1 activation [94]. To date, there is no direct evidence that miR-146a might play a role in IBD T_{reg} function. However, findings supporting this role have been described in the context of rheumatoid arthritis [148].

The elevated number of T_H1, T_H2, T_H17 and reduced number of T_{reg} cells leads to an increased production of various pro-inflammatory cytokines, which in turn leads to activation of innate immune cells such as macrophages, and to the loss of immune tolerance to commensal bacteria, reflecting the typical phenotype of IBD [107]. Processes described in this section are summarized and visualized in the Figure 3.

miRNAs and dysregulation of cellular autophagy

Recent studies provide evidence for dysfunctional autophagy in the intestinal epithelial cells (IECs) being a factor contributing to the development of IBD. Autophagy thereby refers to a degradative pathway contributing to cellular homeostasis. It involves the delivery of cytoplasmic components being enclosed by the autophagosome to the lysosome. The biogenesis of the autophagy machinery is a multi-stage process comprising the formation of a double-membrane structure (phagophore) and its expansion to a vesicle (autophagosome). miRNAs have been shown to be involved in various stages of autophagy [6]. However, current knowledge suggests that IBD-related miRNAs mainly interfere with vesicle elongation [40, 149, 150]. TLR (toll-like receptor)-dependent activation of the NF-◻B pathway leads to increased expression of miR-30c and miR-130a [151]. These miRNAs suppress the transcription of ATG5 as well as ATG16L, thereby inducing a direct impairment of autophagy. Indirect autophagy impairment, in turn, involves NOD2 [152, 153]. For this intracellular PRR, inhibitory effects on TLR-based recognition of PAMPs as well as inductive effects on the assembly of the autophagy elongation complex have been shown. By transcriptionally inhibiting NOD2, miR-192, miR-495, miR-512 and miR-671 interfere with both pathways, leading to a disruption of autophagy [154] (see **Figure 4A**).

miRNAs and disruption of epithelial barrier integrity

Functional maintenance of the healthy gastrointestinal tract is provided by the intestinal epithelium, a physical barrier spatially segregating the environment and the host. Being selectively permeable, it allows protection against intraluminal antigens, microorganisms and their toxins, as well as the passage of dietary nutrients, electrolytes and water. Dysfunction of the intestinal

barrier has been reported to be associated with a variety of diseases, including IBD [7].

The intestinal epithelium consists of a continuous, polarized layer of IECs, adjoining to the mucus layer (apical, biochemical barrier) and the lamina propria (basolateral, immunological barrier) [155] . To maintain homeostasis of the epithelial barrier, adjacent IECs are interconnected by different types of junctional protein complexes. Tight junctions (TJs) are composed of a multitude of transmembrane proteins (including claudins, tricellulin, occulin, as well as junctional adhesion molecules) polymerized into strands. Extracellular domains of these strands establish close contact between adjacent cells. Intracellular domains of the strands are associated to a cytoplasmic plaque (formed by peripheral PDZ proteins, including ZO-1, ZO-2, ZO-3 as well as the Par3-Par6-aPKC polarity complex), which, in turn, is linked to the cytoskeleton [156]. Adherens junctions (AJs) establish cell adhesion by interaction of transmembrane cadherins (E-, N and P-cadherin) of neighboring cells. Catenins (such as plakoglobin, p120-catenin) link cadherins to the cytoskeleton incorporating adapter proteins (such as α-catenin and ß-catenin). Desmosomes, finally, show structural similarities to adherens junctions but are formed by distinct proteins. Thus, cell adhesion of neighboring cells is established by interaction of desmosomal cadherins (desmogleins 1–4 and desmocollins 1–3). Members of the catenin family of proteins (plakoglobin, p120-catenin) link these cadherins to the cytoskeleton by incorporating adapter proteins (such as desmoplakin 1/2) [157].

miRNAs have been shown to interfere with the maintenance of the gastrointestinal epithelial barrier employing different regulatory mechanisms (see **Figure 4B**). For example, miR-122, miR-155 and miR-223 have been shown to target occludin, claudin-1 and claudin-8, respectively, leading to the depletion of these transmembrane proteins and therewith to the impairment of intestinal barrier function [158]. Other mechanisms leading to the destabilization of the junctional complex involve scaffolding proteins of the ZO family. According to their pivotal role for cross-linking the tight junction strands to the actin cytoskeleton, these proteins are targets of a multitude of regulatory pathways. Hence, for example, miR-21 (via RhoB, ras homolog family member B; F-actin), miR-146a (via beta-catenin), miR-155 (via RhoA, ras homolog family member A) as well as miR-200b and miR-221 (via PTEN, phosphatase and tensin homolog; beta-catenin) have been shown to disrupt junctional complexes by impairment of ZO-1 [159, 160]. A last mechanism incorporates JAM-A (junctional adhesion molecule A). By recruiting Par3, this protein induces the formation of the Par3-Par6-aPKC polarity complex and therewith regulates permeability, assembly, and stabilization of tight junctions. This process is regulated through inhibition of aPKC (atypical protein kinase C, maintaining phosphorylation of JAM-A) by miR-155, or of PP2A (protein phosphatase 2A, antagonizing phosphorylation of JAM-A) by miR-221 [161]. Although not yet shown for human IECs, studies in various cell lines suggest that miRNAs are likewise important for regulating the function of adherent junction

MiRNAs in inflammatory bowel disease   15

and desmosomes. Thus, miR-200b induces E-cadherin expression by inhibiting its transcriptional repressor ZEB1. In this way, this miRNA promotes the formation of adherent junctions as shown in rat [162] and mouse [163] studies. By inhibiting a major desmosomal component (desmocollin-2), in turn, miR-29 controls desmosome function by impairing its hyperadhesiveness in human keratinocytes [164].

Role of microRNAs in Diagnosis of IBD

miRNAs have been implicated in numerous physiologic and pathologic processes and their expression has been shown to be deregulated in a multitude of diseases, including several types of cancer [165] as well as inflammatory diseases (such as multiple sclerosis, psoriasis, rheumatic arthritis and systemic lupus erythematosus) [166]. These features, along with their remarkable resistance against physical and enzymatic degradation implicate miRNAs being putative biomarkers [14]. Accordingly, recent studies report candidate biomarkers for diagnosis and prognosis of a multitude of clinical conditions. However, so far none of these candidates have been translated into routine clinical practice [15].

In recent years, deregulation of miRNA expression has also been investigated in the context of IBD. Notably, current literature often solely reports differential expression. Accordingly, studies evaluating the predictive performance of IBD-specific miRNA deregulation are underrepresented. Tables 1 and 2 list miRNAs that have been repeatedly detected to be differentially expressed in intestinal biopsies, peripheral blood, serum and/or saliva comparing IBD to healthy controls, and report point or range estimates of clinical performance as measured by sensitivity (SN), specificity (SP) and balanced accuracy (BAC). miRNAs being specifically deregulated in CD include miR-30e-5p, miR-484, miR-106a-5p, miR-195-5p, miR-140-5p, miR-20a-5p, let-7b-5p, miR-192-5p and miR-93-5p. Conducting receiver-operating characteristics (ROC) analysis, the distinction between CD and HC based on these miRNAs has been reported to perform with SN and SP ranging from 69.57 to 82.61% and 75.00 to 96.88%, respectively, and BAC ranging from 78.06 to 85.40% [167, 168]. Additional miRNAs, such as miR-19b-3p, miR-29a-3p, miR-107, miR-126-3p, miR-191-5p and miR-200c-3p, exhibit differential expression in CD but estimates of their clinical utility remain to be reported. miRNAs specifically deregulated in UC comprise miR-28-5p, miR-103a-2-5p, miR-142-3p, miR-151a-5p, miR-155-5p, miR-188-5p, miR-340-3p, miR-345-5p, miR-378a-3p, miR-422a, miR-500a-5p, miR-501-5p, miR-532-5p, miR-769-5p and miR-874-3p. The distinction of UC from HC based on these miRNAs shows SN and SP ranging from 75.86 to 89.50% and 66.67 to 96.20%, respectively, and BAC ranging from 71.27 to 92.85%. The studies reporting these estimates employed analyses based on ROC [132, 167] or (penalized) support vector machines (SVM) [169, 170]. Additional miRNAs, including miR-142-5p and miR-505-5p exhibit

differential expression in UC but have not yet been evaluated with respect to their clinical utility [168, 171–173].

Remarkably, the more interesting task of distinguishing the subtypes of IBD using miRNAs has not been extensively studied so far. However, Wu et al. [173] and Hübenthal et al. [170] report sets of miRNAs being differentially expressed in CD with respect to UC. Although not replicating each other, both studies indicate that miRNAs may serve as a diagnostic tool complementing classical serological markers such as ASCAs (anti-Saccharomyces cerevisiae antibodies) or pANCAs (perinuclear atypical neutrophil cytoplasmic antibodies). Estimated SN, SP and BAC of 100.00, 83.30 and 91.70%, respectively, suggest that miRNAs even may supersede currently used biomarkers in terms of clinical performance. Beyond diagnosis, miRNAs may be a valuable tool to assess and predict disease activity in IBD. In line with this assumption, Polytarchou et al. [132] and Wang et al. [174] report serum levels of miR-223-3p to be significantly correlated with the activity of CD (as measured by Crohn's disease activity index (CDAI)) as well as UC (as measured by (partial) Mayo score) and therefore allow prediction of an IBD patient's disease state. How the performance of these predictions compares to traditional fecal markers, such as lactoferrin [175, 176], still needs to be evaluated.

Despite these promising findings miRNAs have not been recognized as an emerging tool for IBD diagnosis by neither the European Crohn's and Colitis Organisation (ECCO) [8, 9] nor the American College of Gastroenterology (ACG) [10, 11] so far. Further research is needed to translate the findings into clinical practice.

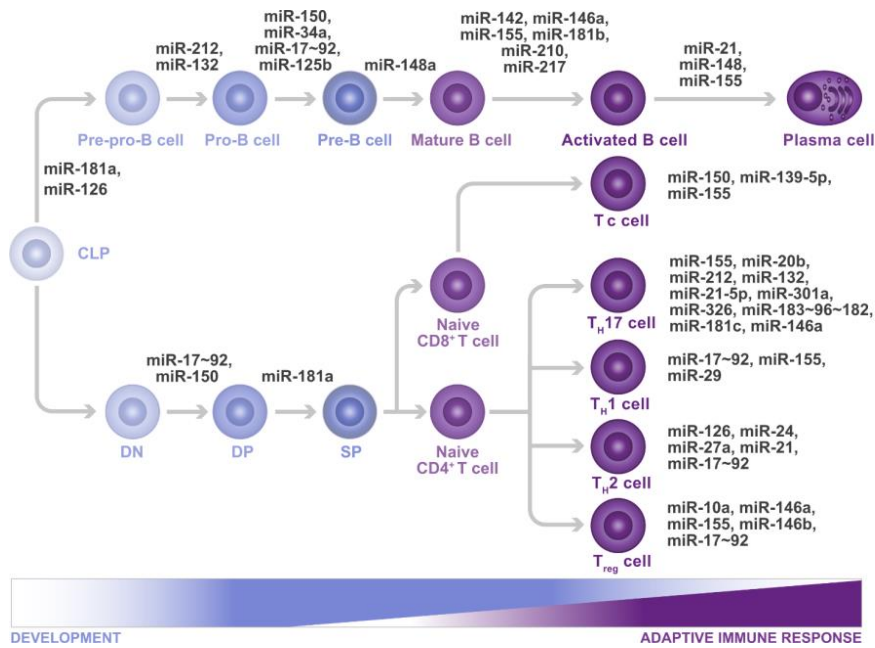MiRNAs in inflammatory bowel disease   17

Figure 1: miRNAs in development and response of innate immune system cells. The common myeloid progenitor (CMP) gives rise to the main two myeloid lineage branches, megakaryocyte-erythroid progenitor (MEP) and granulocyte-monocyte progenitor (GMP) cells, which in turn differentiate into terminal myeloid cells, responsible for the innate immune response. The figure depicts miRNAs involved in the regulation of myelopoiesis as well as in the innate immune response. The color gradient indicates relative time scale of developmental and functional stages of myeloid lineage cells. The figure has been adapted and updated from Mehta et al. [3].

18 Hübenthal, Franke, Lipinski, Juzėnas
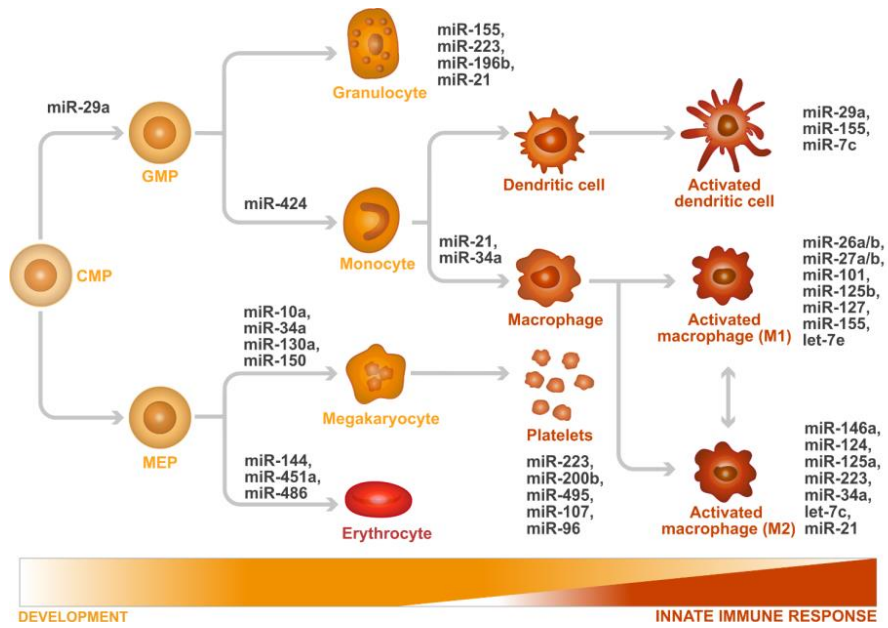
Figure 2: miRNAs in development and response of adaptive immune system cells. The common lymphoid progenitor (CLP) gives rise to the two main lymphoid lineage branches, T cell progenitors double negative (DN) cells and precursor progenitor B-cells (Pre-pro-B), which in turn differentiate into terminal effector cells, responsible for the adaptive immune response. The figure depicts miRNAs which are involved in the regulation of lymphoesis as well as in the adaptive immune response. The color gradient indicates relative time scale of developmental and functional stages of lymphoid lineage cells. DP, double positive; SP, single positive; Pro-B cell, progenitor B cell; Pre-B cell, precursor B cell; T_C cell, cytotoxic T cell; T_H1 cell, T helper 1 cell; T_H2, T helper 2 cell; T_H17, T helper 17 cell; T_{reg}, T regulatory cell. The figure has been modified and updated from Mehta et al. [3].

MiRNAs in inflammatory bowel disease 19

Figure 3: The pathophysiology of inflammatory bowel disease (IBD) in relation to miRNAs. The characteristic immunological features of IBD-affected mucosa are over-reaction of the innate immune cells, defective functionality of intestinal regulatory T (T$_{reg}$) cells, increased recruitment of T helper 17 (T$_H$17), T helper 1 (T$_H$1; in Crohn's disease) and T helper 2 (T$_H$2; in ulcerative colitis) cells in the gut. These features lead to an increased production of pro-inflammatory cytokines, including TNF, IL-6, IL-17, IL-12, IL-23, etc., which in turn lead to the failure of immune tolerance to commensal bacteria. The elevated expression of pro-inflammatory miRNAs (i.e. miR-155, miR-31, miR-17~92 cluster, etc.), as well as decreased expression of anti-inflammatory miRNAs (i.e. miR-146a, miR-10a, miR-29a/b, miR-124, etc.) may disturb gut homeostasis and might cause the aforementioned alterations. However, the increased expression of anti-inflammatory miRNAs (i.e. miR-146a, miR-223, miR-21, etc.), a common event in active IBD conditions, is more likely a consequence of the upstream inflammatory processes (such as TLR stimulation) rather than the cause of tolerance disruption. The figure (excluding miRNAs) has been adapted from Lutter et al. [5] with the additional information from Khalili et al. [105] and Neurath [107].

20 Hübenthal, Franke, Lipinski, Juzėnas
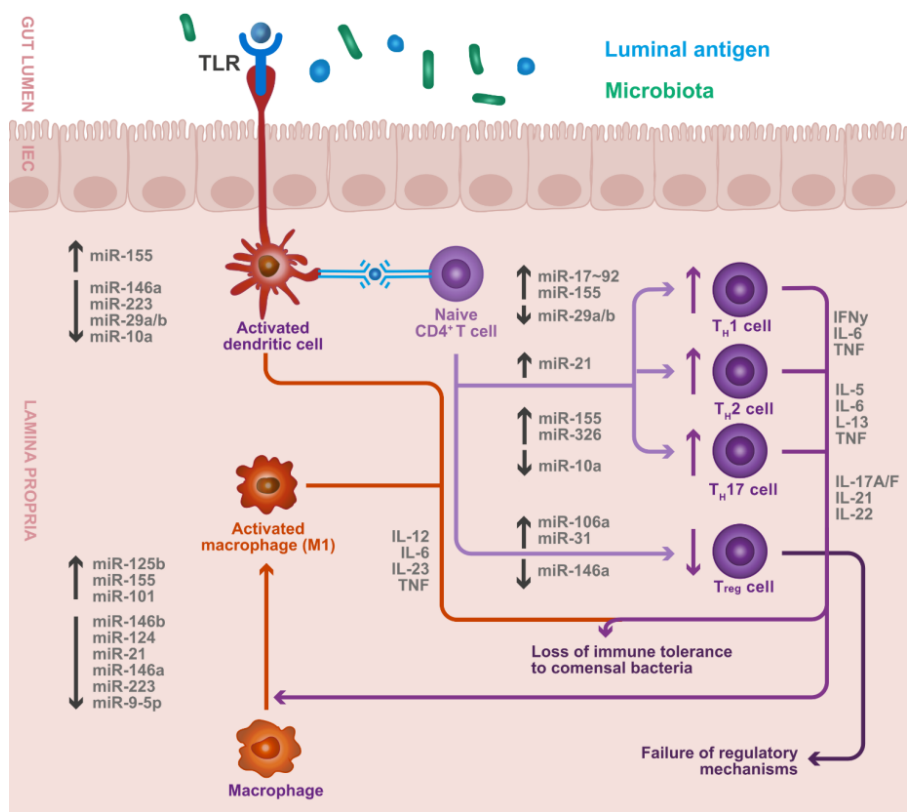
Figure 4: (A) IBD-related miRNAs affecting cellular autophagy. This figure illustrates how miRNAs interfere with vesicle elongation. Direct impairment of the Atg12-Atg5-Atg16-complex depending on TLR and the resulting elevated expression of miR-30c and miR-130a. Indirect impairment by inhibition of NOD2 results from elevated expression of miR-192, miR-495, miR-512 and miR-671. The figure has been modified from Füllgrabe et al. [6] with additions from [40, 149, 150]. **(B) IBD-related miRNA affecting epithelial barrier integrity.** This figure illustrates how transmembrane proteins (claudins CLDN; occludin, OCLN; JAM-A) forming tight junctions between adjacent intestinal epithelial cells (IEC, pink) are affected by miRNAs. Furthermore, adherent junctions (formed by cadherin, CDH; catenin, CAT; vinculin, VCL) as well as desmosomes (formed by desmocollin, DSC; desmoglein, DSG; plakoglobin, PKG; plakophilin, PKP; desmoplakin, DSP). The arrow head indicates the miRNA's mode of action (bar=inhibition, arrow=induction), the line style indicates whether the affection is of direct (solid line) or indirect type (dotted line). For further detail reference is made to the literature (miRNA and autophagy in general [155, 177–179] and specific to IBD [180])
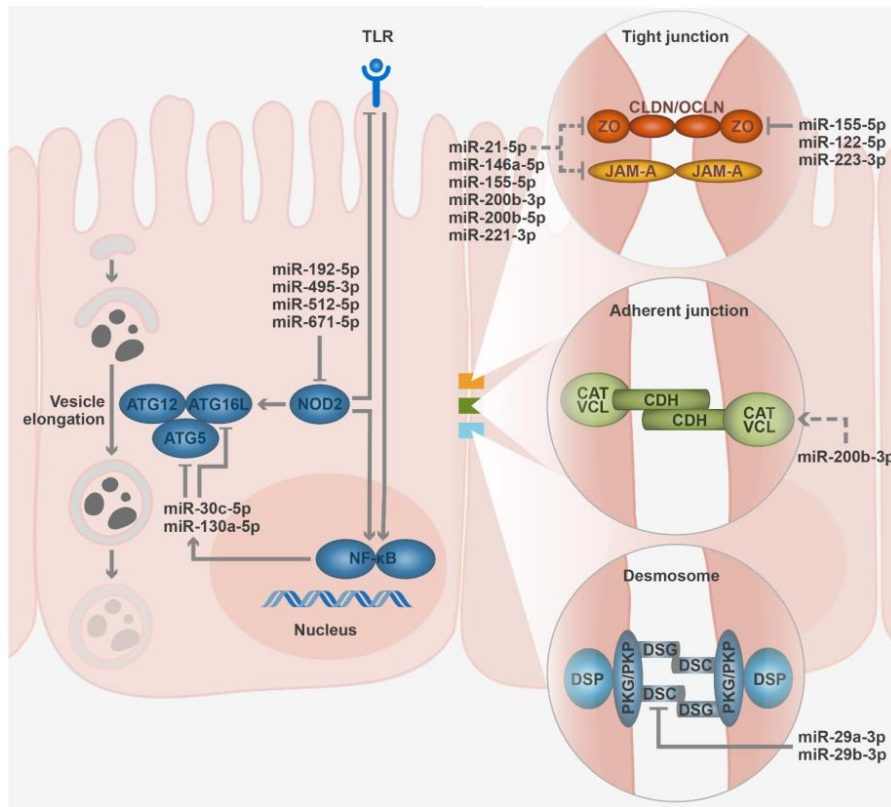
MiRNAs in inflammatory bowel disease   21

Table 1: miRNAs as biomarkers distinguishing CD from HC. The table summarizes miRNAs (named in accordance to miRBase 22) being repeatedly reported to be deregulated in CD with regard to health (direction of effect indicated by ↓ or ↑, respectively) as measured given different types of sample tissue (blood, serum, biopsy and/or saliva, data availability indicated by •). miRNAs for which inconsistent data has been reported are not shown. If available, point or range estimates of clinical performance in term of sensitivity (SN), specificity (SP) and balanced accuracy (BAC) are given. miRNAs whose deregulation in regard to HC is not specific to CD but observed both IBD subtypes are highlighted with a gray background (see **Table 2**).

miRNA	blood	serum	biopsy	saliva	SN (in %)	SP (in %)	BAC (in %)	study
miR-16-5p↑	•	•	•		73.91	100.00	86.96	[138, 139, 172], [168][1]
miR-21-5p↑	•				75.86–76.09	66.67–84.38	71.27–80.24	[138], [168][1]
miR-106a-5p↑	•	•	•		76.09	90.62	83.36	[122, 138, 139], [168][1]
miR-192-5p↑	•	•			78.26–79.31	77.78–78.12	78.19–78.55	[138], [167, 168][1]
let-7b-5p↑	•				82.61	75.00	78.81	[138], [168][1]
miR-20a-5p↑	•				73.91	87.50	80.71	[138], [168][1]
miR-23a-3p↑	•	•			79.00	68.00	73.50	[138], [168][1]
miR-93-5p↑	•				71.74	84.38	78.06	[138], [168][1]
miR-140-5p↑	•				73.91	87.50	80.71	[138], [168][1]
miR-195-5p↑	•				69.57	96.88	83.23	[138, 172], [168][1]
miR-30e-5p↑	•				73.91	96.88	85.40	[138], [168][1]
miR-484↑	•				82.61	84.38	83.50	[138], [168][1]
miR-199a-5p↑	•	•			—	—	—	[138, 168, 173]
miR-223-3p↑	•	•	•		—	—	—	[139, 174, 181]
miR-362-3p↑	•	•			—	—	—	[138, 168, 173]
miR-532-3p↑	•	•			—	—	—	[138, 168, 173]
miR-19b-3p↓	•	•	•		—	—	—	[182, 183]
miR-29a-3p↑	•	•			—	—	—	[138, 168]
miR-107↑	•	•			—	—	—	[138, 168]
miR-191-5p↑	•	•			—	—	—	[138, 168]
miR-126-3p↑	•	•			—	—	—	[138, 168]
miR-200c-3p↑	•	•			—	—	—	[138, 168]

[1]study reporting estimates of clinical performance

Table 2: miRNAs as biomarkers distinguishing UC from HC. The table summarizes miRNAs (named in accordance to miRBase 22) being repeatedly reported to be deregulated in UC with regard to health (direction of effect indicated by ↓ or ↑, respectively) as measured given different types of sample tissue (blood, serum, biopsy and/or saliva, data availability indicated by •). miRNAs for which inconsistent data has been reported are not shown. If available, point or range estimates of clinical performance in term of sensitivity (SN), specificity (SP) and balanced accuracy (BAC) are given. miRNAs whose deregulation in regard to HC is not specific to UC but observed both IBD subtypes are highlighted with a grey background (see **Table 1**).

miRNA	blood	serum	biopsy	saliva	SN (in %)	SP (in %)	BAC (in %)	study
miR-21-5p↑	•	•	•	•	75.86–89.50	66.67–96.20	71.27–92.85	[138, 141, 171, 173, 184–187], [167, 169][1,2]
miR-223-3p↑	•	•			79.00	72.00	75.50	[171, 172, 174, 188], [132][1]
miR-16-5p↑	•				89.50–100.00	90.00–96.20	92.86–95.00	[138, 173], [169, 170][1,2]
miR-199a-5p↑	•	•			89.50	96.20	92.85	[138, 173, 189], [169][1,2]
miR-362-3p↑	•				89.50	96.20	92.85	[138, 168, 173], [169][1,2]
miR-532-3p↑	•				89.50	96.20	92.85	[138, 168, 173], [169][1,2]
miR-155-5p↑	•	•	•		89.50	96.20	92.85	[110, 111, 138, 141, 173], [169][1,2]
miR-28-5p↑	•	•			89.50	96.20	92.85	[138, 173], [169][1,2]
miR-188-5p↑	•				89.50	96.20	92.85	[138, 173], [169][1,2]
miR-378a-3p↑	•				89.50	96.20	92.85	[138, 173], [169][1,2]
miR-340-3p↑	•				89.50	96.20	92.85	[138, 173], [169][1,2]
miR-422a↑	•				89.50	96.20	92.85	[138, 173], [169][1,2]
miR-501-5p↑	•				89.50	96.20	92.85	[138, 173], [169][1,2]
miR-769-5p↑	•				89.50	96.20	92.85	[138, 173], [169][1,2]
miR-151a-5p↑	•	•			89.50	96.20	92.85	[138, 173], [169][1,2]
miR-500a-5p↑	•				89.50	96.20	92.85	[138, 173], [169][1,2]
miR-874-3p↑	•				89.50	96.20	92.85	[138, 173], [169][1,2]
miR-103a-2-5p↑	•				89.50	96.20	92.85	[138, 173], [169][1,2]
miR-142-3p↑	•	•	•	•	75.86	66.67	71.27	[171], [167][1]
miR-345-5p↑	•				89.50	96.20	92.85	[172], [169][1,2]
miR-532-5p↑	•				89.50	96.20	92.85	[172], [169][1,2]
miR-142-5p↑	•	•			—	—	—	[171, 172]
miR-505-5p↓	•				—	—	—	[168, 173]

[1]study reporting estimates of clinical performance, [2]performance estimates based on combinations of miRNAs

MiRNAs in inflammatory bowel disease   23

References

1. Bartel DP (2018) Metazoan MicroRNAs. Cell 173:20–51. https://doi.org/10.1016/j.cell.2018.03.006
2. Gebert LFR, MacRae IJ (2018) Regulation of microRNA function in animals. Nat Rev Mol Cell Biol 1. https://doi.org/10.1038/s41580-018-0045-7
3. Mehta A, Baltimore D (2016) MicroRNAs as regulatory elements in immune system logic. Nat Rev Immunol 16:279–294. https://doi.org/10.1038/nri.2016.40
4. Cloonan N, Wani S, Xu Q, et al (2011) MicroRNAs and their isomiRs function cooperatively to target common biological pathways. Genome Biol 12:R126. https://doi.org/10.1186/gb-2011-12-12-r126
5. Lutter L, Hoytema van Konijnenburg DP, Brand EC, et al (2018) The elusive case of human intraepithelial T cells in gut homeostasis and inflammation. Nat Rev Gastroenterol Hepatol 15:637–649. https://doi.org/10.1038/s41575-018-0039-0
6. Füllgrabe J, Klionsky DJ, Joseph B (2014) The return of the nucleus: Transcriptional and epigenetic control of autophagy. Nat Rev Mol Cell Biol 15:65–74. https://doi.org/10.1038/nrm3716
7. Vancamelbeke M, Vermeire S (2017) The intestinal barrier: a fundamental role in health and disease. Expert Rev Gastroenterol Hepatol 11:821–834. https://doi.org/10.1080/17474124.2017.1343143
8. Gomollón F, Dignass A, Annese V, et al (2017) 3rd European evidence-based consensus on the diagnosis and management of Crohn's disease 2016: Part 1: Diagnosis and medical management. J Crohn's Colitis 11:3–25. https://doi.org/10.1093/ecco-jcc/jjw168
9. Magro F, Gionchetti P, Eliakim R, et al (2017) Third European evidence-based consensus on diagnosis and management of ulcerative colitis. Part 1: Definitions, diagnosis, extra-intestinal manifestations, pregnancy, cancer surveillance, surgery, and ileo-anal pouch disorders. J Crohn's Colitis 11:649–670. https://doi.org/10.1093/ecco-jcc/jjx008
10. Lichtenstein GR, Loftus E V., Isaacs KL, et al (2018) ACG Clinical Guideline: Management of Crohn's Disease in Adults. Am J Gastroenterol 113:481–517. https://doi.org/10.1038/ajg.2018.27
11. Kornbluth A, Sachar DB (2010) Ulcerative colitis practice guidelines in adults: American college of gastroenterology, practice parameters committee. Am J Gastroenterol 105:501–523. https://doi.org/10.1038/ajg.2009.727
12. Dotan I, Fishman S, Dgani Y, et al (2006) Antibodies against laminaribioside and chitobioside are novel serologic markers in Crohn's disease. Gastroenterology 131:366–78. https://doi.org/10.1053/j.gastro.2006.04.030
13. Moum B, Ekbom A, Vatn MH, et al (1997) Inflammatory bowel disease: re-evaluation of the diagnosis in a prospective population based study in south eastern Norway. Gut 40:328–332. https://doi.org/10.1136/gut.40.3.328
14. Mitchell PS, Parkin RK, Kroh EM, et al (2008) Circulating microRNAs as stable blood-based markers for cancer detection. Proc Natl Acad Sci 105:10513–10518. https://doi.org/10.1073/pnas.0804549105
15. Kappel A, Keller A (2017) MiRNA assays in the clinical laboratory: Workflow, detection technologies and automation aspects. Clin Chem Lab Med 55:636–647. https://doi.org/10.1515/cclm-2016-0467
16. Treiber T, Treiber N, Meister G (2018) Regulation of microRNA biogenesis and its crosstalk with other cellular pathways. Nat Rev Mol Cell Biol. https://doi.org/10.1038/s41580-018-0059-1
17. Ha M, Kim VN (2014) Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol

24	Hübenthal, Franke, Lipinski, Juzėnas

15:509–24. https://doi.org/10.1038/nrm3838

18. Leung AKL (2015) The Whereabouts of microRNA Actions: Cytoplasm and Beyond. Trends Cell Biol 25:601–610. https://doi.org/10.1016/j.tcb.2015.07.005

19. Liu S, Da Cunha AP, Rezende RM, et al (2016) The Host Shapes the Gut Microbiota via Fecal MicroRNA. Cell Host Microbe 19:32–43. https://doi.org/10.1016/j.chom.2015.12.005

20. Roberts TC (2014) The MicroRNA Biology of the Mammalian Nucleus. Mol Ther - Nucleic Acids 3:e188. https://doi.org/10.1038/mtna.2014.40

21. Zhang X, Zuo X, Yang B, et al (2014) MicroRNA directly enhances mitochondrial translation during muscle differentiation. Cell 158:607–619. https://doi.org/10.1016/j.cell.2014.05.047

22. Lee LW, Zhang S, Etheridge A, et al (2010) Complexity of the microRNA repertoire revealed by next-generation sequencing. Rna 16:2170–2180. https://doi.org/10.1261/rna.2225110

23. Ameres SL, Zamore PD (2013) Diversifying microRNA sequence and function. Nat Rev Mol Cell Biol 14:475–488. https://doi.org/10.1038/nrm3611

24. Neilsen CT, Goodall GJ, Bracken CP (2012) IsomiRs – the overlooked repertoire in the dynamic microRNAome. Trends Genet 28:544–549. https://doi.org/10.1016/j.tig.2012.07.005

25. Wilson RC, Tambe A, Kidwell MA, et al (2015) Dicer-TRBP complex formation ensures accurate mammalian MicroRNA biogenesis. Mol Cell 57:397–408. https://doi.org/10.1016/j.molcel.2014.11.030

26. Han BW, Hung JH, Weng Z, et al (2011) The 3'-to-5' exoribonuclease nibbler shapes the 3' ends of microRNAs bound to drosophila argonaute1. Curr Biol 21:1878–1887. https://doi.org/10.1016/j.cub.2011.09.034

27. Juzenas S, Venkatesh G, Hübenthal M, et al (2017) A comprehensive, cell specific microRNA catalogue of human peripheral blood. Nucleic Acids Res 45:9290–9301. https://doi.org/10.1093/nar/gkx706

28. McCall MN, Kim MS, Adil M, et al (2017) Toward the human cellular microRNAome. Genome Res 27:1769–1781. https://doi.org/10.1101/gr.222067.117

29. Tsitsiou E, Lindsay MA (2009) microRNAs and the immune response. Curr Opin Pharmacol 9:514–520. https://doi.org/10.1016/j.coph.2009.05.003

30. Hu W, Dooley J, Chung SS, et al (2015) MiR-29a maintains mouse hematopoietic stem cell self-renewal by regulating Dnmt3a. Blood 125:2206–2216. https://doi.org/10.1182/blood-2014-06-585273

31. O'Connell RM, Rao DS, Chaudhuri AA, Baltimore D (2010) Physiological and pathological roles for microRNAs in the immune system. Nat Rev Immunol 10:111–122. https://doi.org/10.1038/nri2708

32. Shaham L, Vendramini E, Ge Y, et al (2015) MicroRNA-486-5p is an erythroid oncomiR of the myeloid leukemias of down syndrome. Blood 125:1292–1301. https://doi.org/10.1182/blood-2014-06-581892

33. Dore LC, Amigo JD, dos Santos CO, et al (2008) A GATA-1-regulated microRNA locus essential for erythropoiesis. Proc Natl Acad Sci 105:3333–3338. https://doi.org/10.1073/pnas.0712312105

34. Jee D, Yang JS, Park SM, et al (2018) Dual Strategies for Argonaute2-Mediated Biogenesis of Erythroid miRNAs Underlie Conserved Requirements for Slicing in Mammals. Mol Cell 69:265–278.e6. https://doi.org/10.1016/j.molcel.2017.12.027

35. Paladini L, Fabris L, Bottai G, et al (2016) Targeting microRNAs as key modulators of tumor immune response. J Exp Clin Cancer Res 35:103. https://doi.org/10.1186/s13046-016-0375-2

36. Navarro F, Gutman D, Meire E, et al (2009) miR-34a contributes to megakaryocytic

MiRNAs in inflammatory bowel disease   25

differentiation of K562 cells independently of p53. Blood 114:2181–2192. https://doi.org/10.1182/blood-2009-02-205062

37. Garzon R, Pichiorri F, Palumbo T, et al (2006) MicroRNA fingerprints during human megakaryocytopoiesis. Proc Natl Acad Sci 103:5078–5083. https://doi.org/10.1073/pnas.0600587103

38. Lu J, Guo S, Ebert BL, et al (2008) MicroRNA-Mediated Control of Cell Fate in Megakaryocyte-Erythrocyte Progenitors. Dev Cell 14:843–853. https://doi.org/10.1016/j.devcel.2008.03.012

39. Gatsiou A, Boeckel J-N, Randriamboavonjy V, Stellos K (2012) MicroRNAs in Platelet Biogenesis and Function: Implications in Vascular Homeostasis and Inflammation. Curr Vasc Pharmacol 10:524–531. https://doi.org/10.2174/157016112801784611

40. Kalla R, Ventham NT, Kennedy NA, et al (2015) MicroRNAs: new players in IBD. Gut 64:504–513. https://doi.org/10.1136/gutjnl-2014-307891

41. Kawasaki T, Kawai T (2014) Toll-like receptor signaling pathways. Front Immunol 5:461. https://doi.org/10.3389/fimmu.2014.00461

42. Testa U, Pelosi E, Castelli G, Labbaye C (2017) miR-146 and miR-155: Two Key Modulators of Immune Response and Tumor Development. Non-Coding RNA 3:22. https://doi.org/10.3390/ncrna3030022

43. Tahamtan A, Teymoori-Rad M, Nakstad B, Salimi V (2018) Anti-inflammatory MicroRNAs and their potential for inflammatory diseases treatment. Front Immunol 9:1377. https://doi.org/10.3389/fimmu.2018.01377

44. Taganov KD, Boldin MP, Chang K-J, Baltimore D (2006) NF- B-dependent induction of microRNA miR-146, an inhibitor targeted to signaling proteins of innate immune responses. Proc Natl Acad Sci 103:12481–12486. https://doi.org/10.1073/pnas.0605298103

45. Essandoh K, Li Y, Huo J, Fan GC (2016) MiRNA-mediated macrophage polarization and its potential role in the regulation of inflammatory response. Shock 46:122–131. https://doi.org/10.1097/SHK.0000000000000604

46. Mann M, Mehta A, Zhao JL, et al (2017) An NF-κB-microRNA regulatory network tunes macrophage inflammatory responses. Nat Commun 8:851. https://doi.org/10.1038/s41467-017-00972-z

47. Wu X-Q, Dai Y, Yang Y, et al (2016) Emerging role of microRNAs in regulating macrophage activation and polarization in immune response and inflammation. Immunology 148:237–48. https://doi.org/10.1111/imm.12608

48. Roy S (2016) miRNA in Macrophage Development and Function. Antioxid Redox Signal 25:795–804. https://doi.org/10.1089/ars.2016.6728

49. Peng L, Zhang H, Hao Y, et al (2016) Reprogramming macrophage orientation by microRNA 146b targeting transcription factor IRF5. EBioMedicine 14:83–96. https://doi.org/10.1016/j.ebiom.2016.10.041

50. Li H, Jiang T, Li MQ, et al (2018) Transcriptional regulation of macrophages polarization by microRNAs. Front Immunol 9:1175. https://doi.org/10.3389/fimmu.2018.01175

51. Smyth LA, Boardman DA, Tung SL, et al (2015) MicroRNAs affect dendritic cell function and phenotype. Immunology 144:197–205. https://doi.org/10.1111/imm.12390

52. Brain O, Owens BMJ, Pichulik T, et al (2013) The intracellular sensor NOD2 induces microrna-29 expression in human dendritic cells to limit IL-23 release. Immunity 39:521–536. https://doi.org/10.1016/j.immuni.2013.08.035

53. Kim SJ, Gregersen PK, Diamond B (2013) Regulation of dendritic cell activation by microRNA let-7c and BLIMP1. J Clin Invest 123:823–833. https://doi.org/10.1172/JCI64712

54. Iwasaki A, Medzhitov R (2015) Control of adaptive immunity by the innate immune system. Nat Immunol 16:343–353. https://doi.org/10.1038/ni.3123

26 Hübenthal, Franke, Lipinski, Juzėnas

55. Cooper MD (2015) The early history of B cells. Nat Rev Immunol 15:191–197. https://doi.org/10.1038/nri3801
56. Okuyama K, Ikawa T, Gentner B, et al (2013) MicroRNA-126-mediated control of cell fate in B-cell myeloid progenitors as a potential alternative to transcriptional factors. Proc Natl Acad Sci 110:13410–13415. https://doi.org/10.1073/pnas.1220710110
57. Chen CZ, Li L, Lodish HF, Bartel DP (2004) MicroRNAs Modulate Hematopoietic Lineage Differentiation. Science (80-) 303:83–86. https://doi.org/10.1126/science.1091903
58. Coffre M, Koralov SB (2017) miRNAs in B Cell Development and Lymphomagenesis. Trends Mol Med 23:721–736. https://doi.org/10.1016/j.molmed.2017.06.001
59. Mehta A, Mann M, Zhao JL, et al (2015) The microRNA-212/132 cluster regulates B cell development by targeting Sox4. J Exp Med 212:1679–1692. https://doi.org/10.1084/jem.20150489
60. Li G, So AYL, Sookram R, et al (2018) Epigenetic silencing of miR-125b is required for normal B-cell development. Blood 131:1920–1930. https://doi.org/10.1182/blood-2018-01-824540
61. Kuchen S, Resch W, Yamane A, et al (2010) Regulation of MicroRNA expression and abundance during lymphopoiesis. Immunity 32:828–839. https://doi.org/10.1016/j.immuni.2010.05.009
62. Gonzalez-Martin A, Adams BD, Lai M, et al (2016) The microRNA miR-148a functions as a critical regulator of B cell tolerance and autoimmunity. Nat Immunol 17:433–440. https://doi.org/10.1038/ni.3385
63. Rothenberg E V., Moore JE, Yui MA (2008) Launching the T-cell-lineage developmental programme. Nat Rev Immunol 8:9–21. https://doi.org/10.1038/nri2232
64. Regelin M, Blume J, Pommerencke J, et al (2015) Responsiveness of Developing T Cells to IL-7 Signals Is Sustained by miR-17~92. J Immunol 195:4832–4840. https://doi.org/10.4049/jimmunol.1402248
65. Ghisi M, Corradin A, Basso K, et al (2011) Modulation of microRNA expression in human T-cell development: Targeting of NOTCH3 by miR-150. Blood 117:7053–7062. https://doi.org/10.1182/blood-2010-12-326629
66. Baumjohann D, Ansel KM (2013) MicroRNA-mediated regulation of T helper cell differentiation and plasticity. Nat Rev Immunol 13:666–678. https://doi.org/10.1038/nri3494
67. Nagasawa T (2006) Microenvironmental niches in the bone marrow required for B-cell development. Nat Rev Immunol 6:107–116. https://doi.org/10.1038/nri1780
68. Alivernini S, Kurowska-Stolarska M, Tolusso B, et al (2016) MicroRNA-155 influences B-cell function through PU.1 in rheumatoid arthritis. Nat Commun 7:12970. https://doi.org/10.1038/ncomms12970
69. Barnes NA, Stephenson S, Cocco M, et al (2012) BLIMP-1 and STAT3 Counterregulate MicroRNA-21 during Plasma Cell Differentiation. J Immunol 189:253–260. https://doi.org/10.4049/jimmunol.1101563
70. Lu D, Nakagawa R, Lazzaro S, et al (2014) The miR-155–PU.1 axis acts on Pax5 to enable efficient terminal B cell differentiation. J Exp Med 211:2183–2198. https://doi.org/10.1084/jem.20140338
71. Andersen MH, Schrama D, thor Straten P, Becker JC (2006) Cytotoxic T Cells. J Invest Dermatol 126:32–41. https://doi.org/10.1038/SJ.JID.5700001
72. Trifari S, Pipkin ME, Bandukwala HS, et al (2013) MicroRNA-directed program of cytotoxic CD8+ T-cell differentiation. Proc Natl Acad Sci 110:18608–18613. https://doi.org/10.1073/pnas.1317191110
73. Dudda JC, Salaun B, Ji Y, et al (2013) MicroRNA-155 is required for effector cd8+t cell responses to virus infection and cancer. Immunity 38:742–753. https://doi.org/10.1016/j.immuni.2012.12.006

MiRNAs in inflammatory bowel disease   27

74. Zhu J, Paul WE (2008) CD4 T cells: Fates, functions, and faults. Blood 112:1557–1569. https://doi.org/10.1182/blood-2008-05-078154

75. Baumjohann D, Heissmeyer V (2018) Posttranscriptional Gene Regulation of T Follicular Helper Cells by RNA-Binding Proteins and microRNAs. Front Immunol 9:1794. https://doi.org/10.3389/fimmu.2018.01794

76. Workman CJ, Szymczak-Workman AL, Collison LW, et al (2009) The development and function of regulatory T cells. Cell Mol Life Sci 66:2603–2622. https://doi.org/10.1007/s00018-009-0026-2

77. Du C, Liu C, Kang J, et al (2009) MicroRNA miR-326 regulates TH-17 differentiation and is associated with the pathogenesis of multiple sclerosis. Nat Immunol 10:1252–1259. https://doi.org/10.1038/ni.1798

78. Mycko MP, Cichalewska M, Machlanska A, et al (2012) microRNA-301a regulation of a T-helper 17 immune response controls autoimmune demyelination. Proc Natl Acad Sci 109:E1248–E1257. https://doi.org/10.1073/pnas.1114325109

79. Murugaiyan G, Da Cunha AP, Ajay AK, et al (2015) MicroRNA-21 promotes Th17 differentiation and mediates experimental autoimmune encephalomyelitis. J Clin Invest 125:1069–1080. https://doi.org/10.1172/JCI74347

80. Zhang Z, Xue Z, Liu Y, et al (2018) MicroRNA-181c promotes Th17 cell differentiation and mediates experimental autoimmune encephalomyelitis. Brain Behav Immun 70:305–314. https://doi.org/10.1016/j.bbi.2018.03.011

81. Ichiyama K, Gonzalez-Martin A, Kim BS, et al (2016) The MicroRNA-183-96-182 Cluster Promotes T Helper 17 Cell Pathogenicity by Negatively Regulating Transcription Factor Foxo1 Expression. Immunity 44:1284–1298. https://doi.org/10.1016/j.immuni.2016.05.015

82. Yao R, Ma YL, Liang W, et al (2012) MicroRNA-155 Modulates Treg and Th17 Cells Differentiation and Th17 Cell Function by Targeting SOCS1. PLoS One 7:e46082. https://doi.org/10.1371/journal.pone.0046082

83. Nakahama T, Hanieh H, Nguyen NT, et al (2013) Aryl hydrocarbon receptor-mediated induction of the microRNA-132/212 cluster promotes interleukin-17-producing T-helper cell differentiation. Proc Natl Acad Sci 110:11964–11969. https://doi.org/10.1073/pnas.1311087110

84. Hu R, Huffaker TB, Kagele DA, et al (2013) MicroRNA-155 Confers Encephalogenic Potential to Th17 Cells by Promoting Effector Gene Expression. J Immunol 190:5972–5980. https://doi.org/10.4049/jimmunol.1300351

85. Wang D, Tang M, Zong P, et al (2018) MiRNA-155 regulates the Th17/Treg ratio by targeting SOCS1 in severe acute pancreatitis. Front Physiol 9:686. https://doi.org/10.3389/fphys.2018.00686

86. Liu SQ, Jiang S, Li C, et al (2014) Mir-17-92 cluster targets phosphatase and tensin homology and ikaros family zinc finger 4 to promote th17-mediated inflammation. J Biol Chem 289:12446–12456. https://doi.org/10.1074/jbc.M114.550723

87. Montoya MM, Maul J, Singh PB, et al (2017) A Distinct Inhibitory Function for miR-18a in Th17 Cell Differentiation. J Immunol 199:559–569. https://doi.org/10.4049/jimmunol.1700170

88. Li B, Wang X, Choi IY, et al (2017) miR-146a modulates autoreactive Th17 cell differentiation and regulates organ-specific autoimmunity. J Clin Invest 127:3702–3716. https://doi.org/10.1172/JCI94012

89. Jiang S, Li C, Olive V, et al (2011) Molecular dissection of the miR-17-92 cluster's critical dual roles in promoting Th1 responses and preventing inducible Treg differentiation. Blood 118:5487–5497. https://doi.org/10.1182/blood-2011-05-355644

90. O'Connell RM, Kahn D, Gibson WSJ, et al (2010) MicroRNA-155 promotes autoimmune inflammation by enhancing inflammatory T cell development. Immunity 33:607–619.

28 Hübenthal, Franke, Lipinski, Juzėnas

https://doi.org/10.1016/j.immuni.2010.09.009

91. Singh PB, Pua HH, Happ HC, et al (2017) MicroRNA regulation of type 2 innate lymphoid cell homeostasis and function in allergic inflammation. J Exp Med 214:3627–3643. https://doi.org/10.1084/jem.20170545

92. Pua HH, Steiner DF, Patel S, et al (2016) MicroRNAs 24 and 27 Suppress Allergic Inflammation and Target a Network of Regulators of T Helper 2 Cell-Associated Cytokine Production. Immunity 44:821–832. https://doi.org/10.1016/j.immuni.2016.01.003

93. Lu LF, Thai TH, Calado DP, et al (2009) Foxp3-Dependent MicroRNA155 Confers Competitive Fitness to Regulatory T Cells by Targeting SOCS1 Protein. Immunity 30:80–91. https://doi.org/10.1016/j.immuni.2008.11.010

94. Lu LF, Boldin MP, Chaudhry A, et al (2010) Function of miR-146a in Controlling Treg Cell-Mediated Regulation of Th1 Responses. Cell 142:914–929. https://doi.org/10.1016/j.cell.2010.08.012

95. Lu Y, Hippen KL, Lemire AL, et al (2016) MiR-146b antagomir-treated human Tregs acquire increased GVHD inhibitory potency. Blood 128:1424–1435. https://doi.org/10.1182/blood-2016-05-714535

96. Takahashi H, Kanno T, Nakayamada S, et al (2012) TGF-β and retinoic acid induce the microRNA miR-10a, which targets Bcl-6 and constrains the plasticity of helper T cells. Nat Immunol 13:587–595. https://doi.org/10.1038/ni.2286

97. de Kouchkovsky D, Esensten JH, Rosenthal WL, et al (2013) microRNA-17-92 regulates IL-10 production by regulatory T cells and control of experimental autoimmune encephalomyelitis. J Immunol 191:1594–605. https://doi.org/10.4049/jimmunol.1203567

98. Kohlhaas S, Garden OA, Scudamore C, et al (2009) Cutting Edge: The Foxp3 Target miR-155 Contributes to the Development of Regulatory T Cells. J Immunol 182:2578–2582. https://doi.org/10.4049/jimmunol.0803162

99. Maul J, Baumjohann D (2016) Emerging Roles for MicroRNAs in T Follicular Helper Cell Differentiation. Trends Immunol 37:297–309. https://doi.org/10.1016/J.IT.2016.03.003

100. Baumjohann D, Kageyama R, Clingan JM, et al (2013) The microRNA cluster miR-17~92 promotes T FH cell differentiation and represses subset-inappropriate gene expression. Nat Immunol. https://doi.org/10.1038/ni.2642

101. Gutiérrez-Vázquez C, Enright AJ, Rodríguez-Galán A, et al (2017) 3' Uridylation controls mature microRNA turnover during CD4 T-cell activation. Rna 23:882–891. https://doi.org/10.1261/rna.060095.116

102. De Souza HSP, Fiocchi C (2016) Immunopathogenesis of IBD: Current state of the art. Nat Rev Gastroenterol Hepatol 13:13–27. https://doi.org/10.1038/nrgastro.2015.186

103. Franke A, McGovern DPB, Barrett JC, et al (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 42:1118–1125. https://doi.org/10.1038/ng.717

104. Franke A, Balschun T, Sina C, et al (2010) Genome-wide association study for ulcerative colitis identifies risk loci at 7q22 and 22q13 (IL17REL). Nat Genet 42:292–294. https://doi.org/10.1038/ng.553

105. Khalili H, Chan SSM, Lochhead P, et al (2018) The role of diet in the aetiopathogenesis of inflammatory bowel disease. Nat Rev Gastroenterol Hepatol 15:525–535. https://doi.org/10.1038/s41575-018-0022-9

106. Berkowitz L, Schultz BM, Salazar GA, et al (2018) Impact of Cigarette Smoking on the Gastrointestinal Tract Inflammation: Opposing Effects in Crohn's Disease and Ulcerative Colitis. Front Immunol 9:74. https://doi.org/10.3389/fimmu.2018.00074

107. Neurath MF (2014) Cytokines in inflammatory bowel disease. Nat Rev Immunol

MiRNAs in inflammatory bowel disease   29

14:329–342. https://doi.org/10.1038/nri3661

108. Ceppi M, Pereira PM, Dunand-Sauthier I, et al (2009) MicroRNA-155 modulates the interleukin-1 signaling pathway in activated human monocyte-derived dendritic cells. Proc Natl Acad Sci 106:2735–2740. https://doi.org/10.1073/pnas.0811073106

109. Bala S, Marcos M, Kodys K, et al (2011) Up-regulation of microRNA-155 in macrophages contributes to increased Tumor Necrosis Factor α (TNFα) production via increased mRNA half-life in alcoholic liver disease. J Biol Chem 286:1436–1444. https://doi.org/10.1074/jbc.M110.145870

110. Béres NJ, Szabó D, Kocsis D, et al (2016) Role of Altered Expression of miR-146a, miR-155, and miR-122 in Pediatric Patients with Inflammatory Bowel Disease. Inflamm Bowel Dis 22:327–35. https://doi.org/10.1097/MIB.0000000000000687

111. Min M, Peng L, Yang Y, et al (2014) Microrna-155 is involved in the pathogenesis of ulcerative colitis by targeting FOXO3a. Inflamm Bowel Dis 20:652–659. https://doi.org/10.1097/MIB.0000000000000009

112. Lu ZJ, Wu JJ, Jiang WL, et al (2017) MicroRNA-155 promotes the pathogenesis of experimental colitis by repressing SHIP-1 expression. World J Gastroenterol 23:976–985. https://doi.org/10.3748/wjg.v23.i6.976

113. Li J, Zhang J, Guo H, et al (2018) Critical role of alternative M2 skewing in miR-155 deletion-mediated protection of colitis. Front Immunol 9:904. https://doi.org/10.3389/fimmu.2018.00904

114. Van Der Goten J, Vanhove W, Lemaire K, et al (2014) Integrated miRNA and mRNA expression profiling in inflamed colon of patients with ulcerative colitis. PLoS One 9:e116117. https://doi.org/10.1371/journal.pone.0116117

115. Zhao JL, Rao DS, O'Connell RM, et al (2013) MicroRNA-146a acts as a guardian of the quality and longevity of hematopoietic stem cells in mice. Elife 2013:e00537. https://doi.org/10.7554/eLife.00537

116. Magilnick N, Reyes EY, Wang W-L, et al (2017) *miR-146a – Traf6* regulatory axis controls autoimmunity and myelopoiesis, but is dispensable for hematopoietic stem cell homeostasis and tumor suppression. Proc Natl Acad Sci 114:201706833. https://doi.org/10.1073/pnas.1706833114

117. Grants J, Wegrzyn J, Knapp D, et al (2017) Single Cell-Resolution Analysis of HSC Dysfunction in Mir-146a knockout Mice. Blood 130:

118. Tang Y, Luo X, Cui H, et al (2009) MicroRNA-146a contributes to abnormal activation of the type I interferon pathway in human lupus by targeting the key signaling proteins. Arthritis Rheum 60:1065–1075. https://doi.org/10.1002/art.24436

119. Bazzoni F, Rossato M, Fabbri M, et al (2009) Induction and regulatory function of miR-9 in human monocytes and neutrophils exposed to proinflammatory signals. Proc Natl Acad Sci 106:5282–5287. https://doi.org/10.1073/pnas.0810909106

120. Wang Y, Han Z, Fan Y, et al (2017) MicroRNA-9 inhibits NLRP3 inflammasome activation in human atherosclerosis inflammation cell models through the JAK1/STAT signaling pathway. Cell Physiol Biochem 41:1555–1571. https://doi.org/10.1159/000470822

121. Ben-Shachar S, Yanai H, Horev HS, et al (2016) MicroRNAs expression in the ileal pouch of patients with ulcerative colitis is robustly up-regulated and correlates with disease phenotypes. PLoS One 11:e0159956. https://doi.org/10.1371/journal.pone.0159956

122. Fasseu M, Tréton X, Guichard C, et al (2010) Identification of Restricted Subsets of Mature microRNA Abnormally Expressed in Inactive Colonic Mucosa of Patients with Inflammatory Bowel Disease. PLoS One 5:e13160. https://doi.org/10.1371/journal.pone.0013160

123. Toiyama Y, Okugawa Y, Tanaka K, et al (2017) A Panel of Methylated MicroRNA Biomarkers for Identifying High-Risk Patients With Ulcerative Colitis-Associated

30 Hübenthal, Franke, Lipinski, Juzėnas

Colorectal Cancer. Gastroenterology 153:1634–1646.e8. https://doi.org/10.1053/j.gastro.2017.08.037

124. Ma C, Li Y, Li M, et al (2014) microRNA-124 negatively regulates TLR signaling in alveolar macrophages in response to mycobacterial infection. Mol Immunol 62:150–8. https://doi.org/10.1016/j.molimm.2014.06.014

125. Koukos G, Polytarchou C, Kaplan JL, et al (2013) MicroRNA-124 Regulates STAT3 Expression and Is Down-regulated in Colon Tissues of Pediatric Patients With Ulcerative Colitis. Gastroenterology 145:842–852.e2. https://doi.org/10.1053/j.gastro.2013.07.001

126. Xue X, Feng T, Yao S, et al (2011) Microbiota Downregulates Dendritic Cell Expression of miR-10a, Which Targets IL-12/IL-23p40. J Immunol 187:5879–5886. https://doi.org/10.4049/jimmunol.1100535

127. Wu W, He C, Liu C, et al (2015) miR-10a inhibits dendritic cell activation and Th1/Th17 cell immune responses in IBD. Gut 64:1755–1764. https://doi.org/10.1136/gutjnl-2014-307980

128. Cuthbert AP, Fisher SA, Mirza MM, et al (2002) The contribution of NOD2 gene mutations to the risk and site of disease in inflammatory bowel disease. Gastroenterology 122:867–874. https://doi.org/10.1053/gast.2002.32415

129. Barrett JC, Hansoul S, Nicolae DL, et al (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40:955–962. https://doi.org/10.1038/ng.175

130. Zhou H, Xiao J, Wu N, et al (2015) MicroRNA-223 Regulates the Differentiation and Function of Intestinal Dendritic Cells and Macrophages by Targeting C/EBPβ. Cell Rep 13:1149–1160. https://doi.org/10.1016/j.celrep.2015.09.073

131. Neudecker V, Haneklaus M, Jensen O, et al (2017) Myeloid-derived miR-223 regulates intestinal inflammation via repression of the NLRP3 inflammasome. J Exp Med 214:1737–1752. https://doi.org/10.1046/j.1365-2559.1996.297345.x

132. Polytarchou C, Oikonomopoulos A, Mahurkar S, et al (2015) Assessment of Circulating MicroRNAs for the Diagnosis and Disease Activity Evaluation in Patients with Ulcerative Colitis by Using the Nanostring Technology. Inflamm Bowel Dis 21:2533–9. https://doi.org/10.1097/MIB.0000000000000547

133. Ungaro R, Mehandru S, Allen PB, et al (2017) Ulcerative colitis. Lancet (London, England) 389:1756–1770. https://doi.org/10.1016/S0140-6736(16)32126-2

134. Torres J, Mehandru S, Colombel J-F, Peyrin-Biroulet L (2017) Crohn's disease. Lancet (London, England) 389:1741–1755. https://doi.org/10.1016/S0140-6736(16)31711-1

135. Singh UP, Murphy AE, Enos RT, et al (2014) miR-155 deficiency protects mice from experimental colitis by reducing T helper type 1/type 17 responses. Immunology 143:478–489. https://doi.org/10.1111/imm.12328

136. Ma F, Xu S, Liu X, et al (2011) The microRNA miR-29 controls innate and adaptive immune responses to intracellular bacterial infection by targeting interferon-γ. Nat Immunol 12:861–869. https://doi.org/10.1038/ni.2073

137. Steiner DF, Thomas MF, Hu JK, et al (2011) MicroRNA-29 Regulates T-Box Transcription Factors and Interferon-γ Production in Helper T Cells. Immunity 35:169–181. https://doi.org/10.1016/j.immuni.2011.07.009

138. Paraskevi A, Theodoropoulos G, Papaconstantinou I, et al (2012) Circulating MicroRNA in inflammatory bowel disease. J Crohn's Colitis 6:900–904. https://doi.org/10.1016/j.crohns.2012.02.006

139. Wu F, Zhang S, Dassopoulos T, et al (2010) Identification of microRNAs associated with ileal and colonic Crohn's disease. Inflamm Bowel Dis 16:1729–38. https://doi.org/10.1002/ibd.21267

140. Schaefer JS, Montufar-Solis D, Vigneswaran N, Klein JR (2011) Selective Upregulation

MiRNAs in inflammatory bowel disease 31

of microRNA Expression in Peripheral Blood Leukocytes in IL-10-/- Mice Precedes Expression in the Colon. J Immunol 187:5834–5841. https://doi.org/10.4049/jimmunol.1100922

141. Takagi T, Naito Y, Mizushima K, et al (2010) Increased expression of microRNA in the inflamed colonic mucosa of patients with active ulcerative colitis. J Gastroenterol Hepatol 25 Suppl 1:S129-33. https://doi.org/10.1111/j.1440-1746.2009.06216.x

142. Thorlacius-Ussing G, Schnack Nielsen B, Andersen V, et al (2017) Expression and Localization of miR-21 and miR-126 in Mucosal Tissue from Patients with Inflammatory Bowel Disease. Inflamm Bowel Dis 23:739–752. https://doi.org/10.1097/MIB.0000000000001086

143. Ando Y, Mazzurana L, Forkel M, et al (2016) Downregulation of MicroRNA-21 in colonic CD3+T cells in UC remission. Inflamm Bowel Dis 22:2788–2793. https://doi.org/10.1097/MIB.0000000000000969

144. Eastaff-Leung N, Mabarrack N, Barbour A, et al (2010) Foxp3+ regulatory T cells, Th17 effector cells, and cytokine environment in inflammatory bowel disease. J Clin Immunol 30:80–89. https://doi.org/10.1007/s10875-009-9345-1

145. Sanctuary MR, Huang RH, Jones AA, et al (2018) miR-106a deficiency attenuates inflammation in murine IBD models. Mucosal Immunol 1. https://doi.org/10.1038/s41385-018-0091-7

146. Keith BP, Barrow JB, Toyonaga T, et al (2018) Colonic epithelial miR-31 associates with the development of Crohn's phenotypes. JCI Insight 3:. https://doi.org/10.1172/jci.insight.122788

147. Zhou W, Pal AS, Hsu AYH, et al (2018) MicroRNA-223 Suppresses the Canonical NF-κB Pathway in Basal Keratinocytes to Dampen Neutrophilic Inflammation. Cell Rep 22:1810–1823. https://doi.org/10.1016/j.celrep.2018.01.058

148. Zhou Q, Haupt S, Kreuzer JT, et al (2015) Decreased expression of miR-146a and miR-155 contributes to an abnormal Treg phenotype in patients with rheumatoid arthritis. Ann Rheum Dis 74:1265–1274. https://doi.org/10.1136/annrheumdis-2013-204377

149. Chapman CG, Pekow J (2015) The emerging role of miRNAs in inflammatory bowel disease: A review. Therap Adv Gastroenterol 8:4–22. https://doi.org/10.1177/1756283X14547360

150. Cao B, Zhou X, Ma J, et al (2017) Role of MiRNAs in Inflammatory Bowel Disease. Dig Dis Sci 62:1426–1438. https://doi.org/10.1007/s10620-017-4567-1

151. Nguyen HTT, Dalmasso G, Müller S, et al (2014) Crohn's disease-associated adherent invasive escherichia coli modulate levels of microRNAs in intestinal epithelial cells to reduce autophagy. Gastroenterology 146:508–519. https://doi.org/10.1053/j.gastro.2013.10.021

152. Cooney R, Baker J, Brain O, et al (2010) NOD2 stimulation induces autophagy in dendritic cells influencing bacterial handling and antigen presentation. Nat Med 16:90–7. https://doi.org/10.1038/nm.2069

153. Travassos LH, Carneiro LAM, Ramjeet M, et al (2010) Nod1 and Nod2 direct autophagy by recruiting ATG16L1 to the plasma membrane at the site of bacterial entry. Nat Immunol 11:55–62. https://doi.org/10.1038/ni.1823

154. Chuang AY, Chuang JC, Zhai Z, et al (2014) NOD2 expression is regulated by microRNAs in colonic epithelial HCT116 cells. Inflamm Bowel Dis 20:126–135. https://doi.org/10.1097/01.MIB.0000436954.70596.9b

155. Neunlist M, Van Landeghem L, Mahé MM, et al (2013) The digestive neuronal-glial-epithelial unit: A new actor in gut health and disease. Nat Rev Gastroenterol Hepatol 10:90–100. https://doi.org/10.1038/nrgastro.2012.221

156. Anderson JM, Van Itallie CM (2009) Physiology and function of the tight junction. Cold Spring Harb Perspect Biol 1:a002584. https://doi.org/10.1101/cshperspect.a002584

32   Hübenthal, Franke, Lipinski, Juzėnas

157. Nekrasova O, Green KJ (2013) Desmosome assembly and dynamics. Trends Cell Biol 23:537–546. https://doi.org/10.1016/j.tcb.2013.06.004

158. Wang H, Chao K, Ng SC, et al (2016) Pro-inflammatory miR-223 mediates the cross-talk between the IL23 pathway and the intestinal barrier in inflammatory bowel disease. Genome Biol 17:58. https://doi.org/10.1186/s13059-016-0901-8

159. Liu M, Tang Q, Qiu M, et al (2011) MiR-21 targets the tumor suppressor RhoB and regulates proliferation, invasion and apoptosis in colorectal cancer cells. FEBS Lett 585:2998–3005. https://doi.org/10.1016/j.febslet.2011.08.014

160. Xue Q, Sun K, Deng HJ, et al (2013) Anti-miRNA-221 sensitizes human colorectal carcinoma cells to radiation by upregulating PTEN. World J Gastroenterol 19:9307–9317. https://doi.org/10.3748/wjg.v19.i48.9307

161. Ebnet K (2017) Junctional Adhesion Molecules (JAMs): Cell Adhesion Receptors With Pleiotropic Functions in Cell Physiology and Development. Physiol Rev 97:1529–1554. https://doi.org/10.1152/physrev.00004.2017

162. Chen Y, Xiao Y, Ge W, et al (2013) MiR-200b inhibits TGF-β1-induced epithelial-mesenchymal transition and promotes growth of intestinal epithelial cells. Cell Death Dis 4:e541. https://doi.org/10.1038/cddis.2013.22

163. Korpal M, Lee ES, Hu G, Kang Y (2008) The miR-200 family inhibits epithelial-mesenchymal transition and cancer cell migration by direct targeting of E-cadherin transcriptional repressors ZEB1 and ZEB2. J Biol Chem 283:14910–14914. https://doi.org/10.1074/jbc.C800074200

164. Kurinna S, Schäfer M, Ostano P, et al (2014) A novel Nrf2-miR-29-desmocollin-2 axis regulates desmosome function in keratinocytes. Nat Commun 5:5099. https://doi.org/10.1038/ncomms6099

165. Lu J, Getz G, Miska EA, et al (2005) MicroRNA expression profiles classify human cancers. Nature 435:834–838. https://doi.org/10.1038/nature03702

166. Iborra M, Bernuzzi F, Invernizzi P, Danese S (2012) MicroRNAs in autoimmunity and inflammatory bowel disease: Crucial regulators in immune response. Autoimmun Rev 11:305–314. https://doi.org/10.1016/j.autrev.2010.07.002

167. Zahm AM, Hand NJ, Tsoucas DM, et al (2014) Rectal microRNAs are perturbed in pediatric inflammatory bowel disease of the colon. J Crohn's Colitis 8:1108–1117. https://doi.org/10.1016/j.crohns.2014.02.012

168. Zahm AM, Thayu M, Hand NJ, et al (2011) Circulating microRNA is a biomarker of pediatric crohn disease. J Pediatr Gastroenterol Nutr 53:26–33. https://doi.org/10.1097/MPG.0b013e31822200cc

169. Duttagupta R, DiRienzo S, Jiang R, et al (2012) Genome-wide maps of circulating miRNA biomarkers for Ulcerative Colitis. PLoS One 7:e31241. https://doi.org/10.1371/journal.pone.0031241

170. Hübenthal M, Hemmrich-Stanisak G, Degenhardt F, et al (2015) Sparse modeling reveals miRNA signatures for diagnostics of inflammatory bowel disease. PLoS One 10:1–20. https://doi.org/10.1371/journal.pone.0140155

171. Schaefer JS, Attumi T, Opekun AR, et al (2015) MicroRNA signatures differentiate Crohn's disease from ulcerative colitis. BMC Immunol 16:1–13. https://doi.org/10.1186/s12865-015-0069-0

172. Iborra M, Bernuzzi F, Correale C, et al (2013) Identification of serum and tissue micro-RNA expression profiles in different stages of inflammatory bowel disease. Clin Exp Immunol 173:250–258. https://doi.org/10.1111/cei.12104

173. Wu F, Guo NJ, Tian H, et al (2011) Peripheral blood MicroRNAs distinguish active ulcerative colitis and Crohn's disease. Inflamm Bowel Dis 17:241–250. https://doi.org/10.1002/ibd.21450

174. Wang H, Zhang S, Yu Q, et al (2016) Circulating MicroRNA223 is a new biomarker for

inflammatory bowel disease. Med (United States) 95:e2703.
https://doi.org/10.1097/MD.0000000000002703

175. Langhorst J, Elsenbruch S, Koelzer J, et al (2008) Noninvasive markers in the
      assessment of intestinal inflammation in inflammatory bowel diseases: Performance
      of fecal lactoferrin, calprotectin, and PMN-elastase, CRP, and clinical indices. Am J
      Gastroenterol 103:162–169. https://doi.org/10.1111/j.1572-0241.2007.01556.x

176. Dai J, Liu WZ, Zhao YP, et al (2007) Relationship between fecal lactoferrin and
      inflammatory bowel disease. Scand J Gastroenterol 42:1440–1444.
      https://doi.org/10.1080/00365520701427094

177. Delva E, Tucker DK, Kowalczyk AP (2009) The desmosome. Cold Spring Harb Perspect
      Biol 1:a002543–a002543. https://doi.org/10.1101/cshperspect.a002543

178. Cichon C, Sabharwal H, Rüter C, Schmidt MA (2014) MicroRNAs regulate tight junction
      proteins and modulate epithelial/endothelial barrier functions. Tissue Barriers
      2:e944446. https://doi.org/10.4161/21688362.2014.944446

179. Zhou G, Yang L, Gray A, et al (2017) The role of desmosomes in carcinogenesis. Onco
      Targets Ther 10:4059–4063. https://doi.org/10.2147/OTT.S136367

180. Tili E, Michaille JJ, Piurowski V, et al (2017) MicroRNAs in intestinal barrier function,
      inflammatory bowel disease and related cancers — their effects and therapeutic
      potentials. Curr Opin Pharmacol 37:142–150.
      https://doi.org/10.1016/j.coph.2017.10.010

181. Peck BCE, Weiser M, Lee SE, et al (2015) MicroRNAs classify different disease behavior
      phenotypes of Crohn's disease and may have prognostic utility. Inflamm Bowel Dis
      21:2178–2187. https://doi.org/10.1097/MIB.0000000000000478

182. Lewis A, Mehta S, Hanna LN, et al (2015) Low serum levels of microRNA-19 are
      associated with a stricturing Crohn's disease phenotype. Inflamm Bowel Dis 21:1926–
      1934. https://doi.org/10.1097/MIB.0000000000000443

183. Cheng X, Zhang X, Su J, et al (2015) MiR-19b downregulates intestinal SOCS3 to reduce
      intestinal inflammation in Crohn's disease. Sci Rep 5:10397.
      https://doi.org/10.1038/srep10397

184. Wu F, Zikusoka M, Trindade A, et al (2008) MicroRNAs are differentially expressed in
      ulcerative colitis and alter expression of macrophage inflammatory peptide-2 alpha.
      Gastroenterology 135:1624–1635.e24. https://doi.org/10.1053/j.gastro.2008.07.068

185. Feng X, Wang H, Ye S, et al (2012) Up-Regulation of microRNA-126 May Contribute to
      Pathogenesis of Ulcerative Colitis via Regulating NF-κB Inhibitor IκBα. PLoS One
      7:e52782. https://doi.org/10.1371/journal.pone.0052782

186. Yang Y, Ma Y, Shi C, et al (2013) Overexpression of miR-21 in patients with ulcerative
      colitis impairs intestinal epithelial barrier function through targeting the Rho GTPase
      RhoB. Biochem Biophys Res Commun 434:746–752.
      https://doi.org/10.1016/j.bbrc.2013.03.122

187. Polytarchou C, Hommes DW, Palumbo T, et al (2015) MicroRNA214 Is Associated With
      Progression of Ulcerative Colitis, and Inhibition Reduces Development of Colitis and
      Colitis-Associated Cancer in Mice. Gastroenterology 149:981–992.
      https://doi.org/10.1053/j.gastro.2015.05.057

188. Koukos G, Polytarchou C, Kaplan JL, et al (2015) A MicroRNA signature in pediatric
      ulcerative colitis: Deregulation of the miR-4284/CXCL5 pathway in the intestinal
      epithelium. Inflamm Bowel Dis 21:996–1005.
      https://doi.org/10.1097/MIB.0000000000000339

189. Bian Z, Li L, Cui J, et al (2011) Role of miR-150-targeting c-Myb in colonic epithelial
      disruption during dextran sulphate sodium-induced murine experimental colitis and
      human ulcerative colitis. J Pathol 225:544–553. https://doi.org/10.1002/path.2907

34 Hübenthal, Franke, Lipinski, Juzėnas

Acknowledgements

We wish to express our gratitude to Dirk Baumjohann for reviewing the draft. His additions and revisions improved the manuscript considerably. Furthermore, we want to thank Tine Pape for her assistance in designing the figures generated for this chapter.

RESEARCH ARTICLE

Sparse Modeling Reveals miRNA Signatures for Diagnostics of Inflammatory Bowel Disease

Matthias Hübenthal[1◉], Georg Hemmrich-Stanisak[1◉]*, Frauke Degenhardt[1],
Silke Szymczak[1¤], Zhipei Du[1], Abdou Elsharawy[1,2], Andreas Keller[3], Stefan Schreiber[1,4],
Andre Franke[1]

1 Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany, 2 Chemistry
Department, Division of Biochemistry, Faculty of Sciences, Damietta University, New Damietta City, Egypt,
3 Chair for Clinical Bioinformatics, Saarland University, Saarbruücken, Germany, 4 Department of Internal
Medicine I, University Hospital Schleswig-Holstein, Kiel, Germany

◉ These authors contributed equally to this work.
¤ Current address: Institute of Medical Informatics and Statistics, Christian-Albrechts-University of Kiel, Kiel,
Germany
* g.hemmrich-stanisak@ikmb.uni-kiel.de

Abstract

The diagnosis of inflammatory bowel disease (IBD) still remains a clinical challenge and the
most accurate diagnostic procedure is a combination of clinical tests including invasive endos-
copy. In this study we evaluated whether systematic miRNA expression profiling, in conjunc-
tion with machine learning techniques, is suitable as a non-invasive test for the major IBD
phenotypes (Crohn's disease (CD) and ulcerative colitis (UC)). Based on microarray technol-
ogy, expression levels of 863 miRNAs were determined for whole blood samples from 40 CD
and 36 UC patients and compared to data from 38 healthy controls (HC). To further discrimi-
nate between disease-specific and general inflammation we included miRNA expression data
from other inflammatory diseases (inflammation controls (IC): 24 chronic obstructive pulmo-
nary disease (COPD), 23 multiple sclerosis, 38 pancreatitis and 45 sarcoidosis cases) as well
as 70 healthy controls from previous studies. Classification problems considering 2, 3 or 4
groups were solved using different types of penalized support vector machines (SVMs). The
resulting models were assessed regarding sparsity and performance and a subset was
selected for further investigation. Measured by the area under the ROC curve (AUC) the corre-
sponding median holdout-validated accuracy was estimated as ranging from 0.75 to 1.00
(including IC) and 0.89 to 0.98 (excluding IC), respectively. In combination, the corresponding
models provide tools for the distinction of CD and UC as well as CD, UC and HC with expected
classification error rates of 3.1 and 3.3%, respectively. These results were obtained by incor-
porating not more than 16 distinct miRNAs. Validated target genes of these miRNAs have
been previously described as being related to IBD. For others we observed significant enrich-
ment for IBD susceptibility loci identified in earlier GWAS. These results suggest that the pro-
posed miRNA signature is of relevance for the etiology of IBD. Its diagnostic value, however,
should be further evaluated in large, independent, clinically well characterized cohorts.

Introduction

Inflammatory bowel disease (IBD) is a complex, polygenic, chronic intestinal disorder of unknown etiology, comprising two major types: Crohn's disease (CD) and ulcerative colitis (UC). IBD is believed to evolve through a dysregulated response of the immune system to the commensal microbiota associated with intestinal tissues in a genetically susceptible host. The diagnosis of IBD is often achieved only months or years after the first onset of symptoms and still requires a multitude of information from clinical, radiological, endoscopic and histological tests. Extensive heterogeneity is observed in terms of disease presentation, behavior, and response to treatment. However, a definite diagnosis of CD or UC cannot be established in approximately 10%–17% of colitis patients (known as "indeterminate colitis" (IC)) [1] and more than 10% of IBD patients change diagnosis (CD or UC) during the first year of the disease course [2]. Fecal and serological diagnostic tests, e.g. for calprotectin, lactoferrin or CRP (C-reactive protein) as well as serum antibodies like pANCAs (perinuclear antineutrophil cytoplasmic antibody) and ASCAs (anti-S.cerevisiae antibody), supplement invasive endoscopic/colonoscopic methods to verify IBD-diagnosis, to differentiate between the major subtypes or to evaluate disease progression [3,4]. In the last 10 years, several genome-wide association studies (GWAS) were carried out to identify common susceptibility variants for IBD. In a large meta-analysis of previous IBD GWAS, including more than 75,000 cases and controls, Jostins *et al.* identified 71 additional loci, increasing the total number of known IBD susceptibility loci with association of genome-wide significance to more than 163 [5]. While GWAS findings have added tremendously to the understanding of disease etiology and the genetic architecture, common genetic variants have low diagnostic value as shown for IBD [6] and other diseases [7]. Other studies, employing mRNA-based measurements of differential gene expression in tissue or peripheral blood of IBD patients of varying disease state, revealed distinct expression patterns [8–11]. Limitations of these studies were reported when comparing cases and healthy controls or trying to classify disease subphenotypes [12]. Non-coding, regulatory microRNAs (miRNAs) have been studied in the context of their function in IBD [13] but especially because of their ability to serve as diagnostic markers, as recently summarized by Chen *et al.* [14]. As miRNA expression levels are more stable in tissues and body fluids, such as peripheral blood, and as miRNAs act as master-regulators of mRNAs, differential signatures of miRNAs could serve as superior, non-invasive diagnostic markers to verify IBD diagnosis, discriminate between major IBD subphenotypes and to predict prognosis. A core set of deregulated miRNAs has been identified in a series of studies investigating differential miRNA expression in biopsies and peripheral blood of IBD patients [15–22]. Functional links gained from the analysis of IBD-associated miRNA target genes implicate an involvement of cellular pathways of the immune system (NF-κB, IL-23/IL-23R, IL-6/STAT3) [23–29], autophagy [13,30,31], epithelial barrier function [32,33], IBD-associated dysplasia and colorectal cancer [34–36] in IBD disease etiology. Besides these mechanistic insights into the disease, highly accurate predictive sets of miRNAs suitable for diagnostic purposes have not yet been reported. Interestingly, most of the afore-described studies, investigating deregulation of miRNAs, follow the classic approach of statistical hypothesis testing for significant differential expression of single candidate miRNAs. Some publications, however, point out an alternative way of employing large miRNA datasets and machine-learning techniques, such as support vector machines (SVMs) [37] or random forests (RFs) [38]. Keller and colleagues successfully applied SVM-based approaches to identify diagnostic miRNA-profiles for several different diseases [39], such as multiple sclerosis [40,41], lung cancer [42] or male infertility [43]. Others used similar analysis strategies to generate miRNA expression signatures for pharyngeal squamous cell carcinomas [44], thyroid lesions

[45], lung adenocarcinoma [46] or pulmonary tuberculosis [47]. Even ulcerative colitis has been investigated using SVMs, leading to a signature of platelet-derived miRNAs [48].

Here we investigate microarray-based miRNA expression profiles from peripheral blood of IBD patients, using penalized SVMs [49] and random forests for distinction of CD and UC from healthy controls and other complex inflammatory diseases (chronic obstructive pulmonary disease (COPD), multiple sclerosis, pancreatitis and sarcoidosis). The promising results of our pilot study show, that machine-learning techniques and miRNA signatures should be further investigated for IBD diagnostics. Moreover, the miRNA profiles identified yield further insight into the disease-relevant signaling pathways.

Material and Methods

Patient recruitment and sampling

Clinical data and sample material used in this study were obtained under written informed consent of patients as well as healthy donors, and under approvals of the local ethics committees (Biobank Popgen & Ethik-Komission der Medizinischen Fakultät, Universitätsklinikum Schleswig-Holstein, Kiel). We randomly selected blood samples of 40 CD, 36 UC patients and included 38 healthy controls (HC) from our biobank. Patients were collected at the UKSH tertiary referral center. Diagnoses were verified by a clinician after reviewing the respective medical health records. As shown in **Table 1**, patients of every group were matched regarding demographic parameters (mean age at diagnosis of 27.3 and 28.1 years for CD and UC cases, respectively; mean age at sampling of 46.0 and 43.8 years for CD and UC cases, respectively; fraction of males of 54.1% in CD and 53.1% in UC patients, respectively). The majority of the patients was treated with anti-TNF-α inhibitors, such as Infliximab or Mesalazine (67.6% of CD and 90.6% of UC cases) and is therefore assumed to be stable regarding the clinical presentation. The activity of immune cells is assumed to be altered partially since a fraction of the patients additionally was treated with immunosuppressive drugs, such as Azathioprine, Cyclosporine, 6-Mercaptopurine or Tacrolimus (48.6% of CD and 31.3% of UC cases). Furthermore a substantial fraction of the patients underwent the clinically common treatment with SAIDs (steroidal anti-inflammatory drugs; 29.7% of CD and 56.3% of the UC cases, respectively) and/or NSAIDs (non-steroidal anti-inflammatory drugs; 2.7% of CD and 6.3% of the UC cases, respectively). However, based on the available data exacerbation of IBD at sampling was ruled out for 51.4% of CD and 56.3% of UC cases.

Table 1. Characterization of the study subjects. Grouped by CD, UC and HC frequency information (in percent) on demographics (gender and smoking status), medication (anti-TNF-α, immunosuppressant, SAIDs and NSAIDs) as well as symptoms (disease attack at sampling, stenosis, fistula and surgery) is shown.

		demographics		medication				symptoms			
		male	smoker	anti-TNF-alpha	immunosuppressant	said	nsaid	disease attack at sampling	stenosis	fistula	surgery
CD	no	45.9	35.1	32.4	51.4	70.3	97.3	51.4	27.0	48.6	29.7
	yes	54.1	64.9	67.6	48.6	29.7	2.7	0.0	62.2	48.6	70.3
	NA	0.0	0.0	0.0	0.0	0.0	0.0	48.6	10.8	2.7	0.0
UC	no	46.9	68.8	9.4	68.8	43.8	93.8	56.3	78.1	84.4	90.6
	yes	53.1	31.3	90.6	31.3	56.3	6.3	0.0	3.1	3.1	3.1
	NA	0.0	0.0	0.0	0.0	0.0	0.0	43.8	18.8	12.5	6.3
HC	no	46.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	yes	53.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	NA	0.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

doi:10.1371/journal.pone.0140155.t001

miRNA extraction and microarray measurement

After sampling, peripheral blood was anticoagulated using ethylenediaminetetraacetic acid (EDTA) and immediately processed for RNA isolation. Total RNA, including miRNAs, was extracted using the miRNeasy Mini Kit (Qiagen GmbH, Hilden, Germany) and stored at -80°C. All samples were analyzed on the automated Geniom Real Time Analyzer (GRTA, febit biomed GmbH, Heidelberg, Germany) using the Geniom miRNA Biochip for Homo sapiens, covering 866 human miRNA species [50]. Since miRBase has been updated from version 12 to 14 during the time course of the study, we used 863 miRNAs that were consistently present in all three versions for the final data analysis. Biotin labeling was conducted by microfluidic enzymatic on-chip labeling of miRNAs as described previously [51]. Hybridization was carried out for 16 hours at 42°C followed by signal enhancement processing with GRTA. Detection images were analyzed using the Geniom Wizard Software.

Data preprocessing

Sample data for other inflammatory diseases, representing the inflammation control panel for the current investigation, was taken from a previously published study [39]. This dataset comprised 24 COPD, 23 multiple sclerosis, 38 pancreatitis and 45 sarcoidosis cases as well as another 70 healthy controls. Raw data of these samples was downloaded from Gene Expression Omnibus (GEO, Accession code: GSE31568) and analyzed jointly with raw data of samples generated for this study. Samples with median background-subtracted intensity exceeding $1.5 \cdot IQR$ where removed as outliers resulting in 273 samples, including 37 CD, 32 UC, 92 HC, 23 COPD, 23 multiple sclerosis, 35 pancreatitis and 32 sarcoidosis cases. To account for batch effects arising from differences in the source of data the background-subtracted intensity values were centered with regard to the medians of the healthy controls. Normalization then was performed using the R package vsn [52] for robust calibration and variance stabilization.

Classification with penalized support vector machines

To obtain mathematical models that allow diagnostic applications as well as the elucidation of the role of miRNAs in the development of IBD, different types of classification problems were investigated. Aiming for the distinction between CD, UC and HC initially a set of models considering 2 groups was examined (CD vs. HC, UC vs. HC, CD vs. UC). Classification problems additionally incorporating IC (CD vs. IC, UC vs. IC, IC vs. HC) were carried out to differentiate CD, UC and HC from general inflammation. Models aiming for the distinction of combinations of groups were examined by jointly considering 3 groups (CD vs. UC+HC, UC vs. CD+HC, HC vs. CD+UC as well as CD vs. UC+IC, UC vs. CD+IC, IC vs. CD+UC). Finally, also a set of models allowing for 4 groups was investigated (CD vs. UC+HC+IC, UC vs. CD+HC+IC, HC vs. CD+UC+IC, IC vs. CD+UC+HC). Each of the 16 classification problems was solved using different types of linear penalized support vector machines, namely LASSO SVM, elastic net SVM, SCAD SVM and elastic SCAD SVM. Additionally, the linear standard SVM not performing feature selection was used as a reference. It is worth noting that not every classification problem considered has a diagnostic meaning. However, for the subsequent construction of combined classifiers, none of these can be neglected.

Support vector machines (SVMs) are widely used for solving supervised classification problems. However, SVMs do not allow for the selection of important variables (feature selection). Applying the mathematical idea of regularization abolishes this limitation [49]. Accordingly, regularized incarnations of the standard SVM along with efficient algorithms for optimizing their objective functions have been proposed. All these methods share the use of penalties for

model complexity to provide sparse solutions, i.e. small sets of features that enable good classification.

For data $D \in \{(x_i, y_i)|x_i \in R^m, y_i \in \{-1, 1\}\}_{i=1}^n$ with input vectors x_i and class labels y_i for $i = 1, \ldots, n$ the SVM optimization problem corresponds to the minimization of $\|\bar{\beta}\|_2^2$ with respect to the decision rule $y_i(\bar{\beta} \cdot x_i - \beta_0) \geq 1$. As shown by Hastie [49] this also can be written as a regularization problem $\min_{\beta_0, \bar{\beta}} \sum_{i=1}^n L(y_i; f(x_i)) + p_\lambda(\bar{\beta})$ where $L(y_i; f(x_i)) = \max(0, 1 - y_i(\bar{\beta} \cdot x_i - \beta_0))$ is a loss (or cost) function and $p_\lambda(\bar{\beta})$ a penalty function with parameter λ. The classic choices of $p_\lambda(\bar{\beta})$ include the ridge penalty [49] (standard SVM, $p_\lambda(\bar{\beta}) = \lambda\|\bar{\beta}\|_2^2$) and the LASSO [53] (least absolute shrinkage and selection operator, $p_\lambda(\bar{\beta}) = \lambda\|\bar{\beta}\|_1$) as well as their combination known as the elastic net [54] ($p_\lambda(\bar{\beta}) = \lambda_1\|\bar{\beta}\|_1 + \lambda_2\|\bar{\beta}\|_2^2$, $\lambda_1, \lambda_1 \geq 0$). More recently a penalty function improving the properties of the LASSO was published. The SCAD (smoothly clipped absolute deviation) penalty [55] is given by the quadratic spline

$$p_\lambda^{\text{SCAD}}(\bar{\beta}) = \sum_{j=1}^m \left(\lambda|\beta_j|I(|\beta_j| \leq \lambda) + \frac{|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)} I(\lambda < |\beta_j| \leq a\lambda) + \frac{(a+1)\lambda^2}{2} I(|\beta_j| > a\lambda) \right)$$

for $a > 2$ and $\lambda > 0$ and indicator function $I(\cdot)$. Similar to the LASSO this function provides feature selection by shrinking small coefficients $|\beta_j| \leq \lambda$ to zero (resulting in a sparse model). However, in contrast to the LASSO it applies a constant penalty to large coefficients $|\beta_j| > a\lambda$ (resulting in an approximately unbiased model). Combining SCAD with the ridge penalty finally results in the elastic SCAD penalty [56] defined as $p(\bar{\beta}) = p_{\lambda_1}^{\text{SCAD}}(\bar{\beta}) + \lambda_2\|\bar{\beta}\|_2^2$ with tuning parameters $\lambda_1, \lambda_2 \geq 0$. Efficient implementations of SVMs regularized using the penalty functions mentioned before are available in the R package penalizedSVM (version 1.1) [57].

The normalized miRNA expression data were randomly split at a ratio of 5:3, preserving the proportion of samples per group. The first partition was used to construct the respective model, whereas the second was used for evaluation. To estimate the distribution of each model's predictive performance, the partitioning was conducted repeatedly applying 500-fold hold-out sampling (random choice of samples without replacement). The tuning parameters thereby were trained using 5-fold cross validation and fixed grid search based on the respective training datasets. The final SVMs then were obtained by selecting the sparsest median performing model for each investigated classification task. A classifier's performance thereby was measured by the area under the receiver operating characteristic (ROC) curve (AUC, sensitivity as a function of 1-specificity). For illustrative purpose, additional performance measures of varying informational content were determined, e.g. balanced accuracy (BAC), sensitivity (SN), specificity (SP). For each classification problem the sets of miRNAs (miRNA signatures) considered by the sparsest median performing model were selected for further investigation, including validation with random forests and target enrichment analysis.

According to the principal of majority voting [58], the selected models were used to construct combined classifiers for exemplary diagnostic problems (CD vs. UC, CD vs. UC vs. HC, CD vs. UC vs. IC and CD vs. UC vs. HC vs. IC). The diagnoses provided by these models were evaluated using the classification error rate estimated based on the complete dataset. Finally, the risk of observing small combined error rates by chance was assessed using the Z-statistic with parameters estimated based on 1,000-fold permutation of the class labels. Corresponding p-values were calculated using the normal cumulative distribution function and tested for significance using the standard significance level of 0.05.

Validation with random forests

A second machine-learning approach, random forest (RF), was used to analyze the reported miRNA dataset. RF is an ensemble tree method that was first introduced by Breiman *et al.* in 2001 [38], and has been shown to be accurate in both classification and regression problems. Randomization is introduced by constructing each decision tree with a randomly chosen bootstrap sample. Additionally, at each node the optimal splitting variable is selected among a random subset of variables (predictors). Variables selected in RF classification trees are assigned an importance score that is a measure of how much the particular predictor contributes to classifying the respective data. In this study relative recurrency variable importance metric (r2VIM), recently proposed as a measure of variable importance, was used. Based on the permutation importance scheme this measure reduces noisy signal selection [59]. For further details on the concept of RF refer to Strobl *et al.* [60]).

To evaluate the validity of the feature selection employed by the penalized SVMs, two random forests were built for each classification problem. While the first model incorporated variables per holdout selected by the RF, the second contained variables per holdout selected by the SVM (holdout signature). To further validate the meaningfulness of the proposed miRNA signature, another two random forests were built. This time the variable set was constant across the training datasets for each classification problem. For the third model all variables selected in at least one training dataset were ranked by the number of times they were selected and the 50% most frequently selected variables (top signature) were used for training of the RF. Finally, for the fourth model the variables incorporated by the sparsest median performing SVM (median signature) were used.

For comparability, model training, as well as evaluation, incorporated the randomly selected datasets (500-fold holdout partitioning) previously used to construct the SVM classifier. As a measure of model performance again the area under the ROC curve (AUC) was used. All RF analyses were performed in R (version 3.0.1) using the packages parallelRandomForest (version 4.6–7) and ROCR (version 1.0–5) [61]. For each forest, 500 trees (ntree) were built with a terminal node size (nodesize) of 10% of the sample size. The number of randomly selected variables at each node (mtry) was set to the square root of the total number of predictors. For each analysis a random seed was set to a randomly chosen number between 1 and 100,000.

miRNA target gene enrichment analysis

Experimentally validated miRNA target genes were extracted from miRTarBase [62] version 4.5 and tested for significant enrichment within the previously published IBD susceptibility loci [5]. In total more than 163 genetic risk loci have been previously identified as being associated with inflammatory bowel disease (CD: 30, UC: 23, IBD: 110) [5]. 49 out of 1332 experimentally validated miRNA target genes, as listed in miRTarBase, overlap with these loci. To test for overrepresentation of risk loci among the targets of the miRNAs selected for distinguishing CD, UC and HC, Fisher's exact test was applied (see also **S7 Table**). Enrichment was considered as being significant in case p-values were smaller than 0.05. Adjustment for multiple testing was conducted using Bonferroni correction.

Results

Differential expression analysis of peripheral blood miRNAs

To examine potential deregulation, we analyzed expression levels of 863 miRNAs in 40 Crohn's disease patients, 36 ulcerative colitis patients as well as 70 healthy control individuals. After RNA isolation from freshly drawn peripheral blood, miRNA expression data were generated

utilizing the Geniom Biochip miRNA (Homo sapiens). After batch-correction and normalization the background-subtracted microarray intensity values did not show considerable sample-based mean-variance dependencies or sample-based variability of dispersion estimates. As illustrated using multidimensional scaling based on Spearman's rank correlation distance (**S1 Fig**), the groups of interest are visually hardly distinguishable.

In the differential expression analysis (summarized in **S1 Table**) we were able to identify 292 and 353 miRNAs as being significantly deregulated in CD and UC, respectively, when compared to healthy controls (Student's t-test with significance threshold of 0.05 applied to p-values adjusted for multiple testing according to Holm's sequential Bonferroni method). In terms of miRNA expression level differences these results correspond well to previously published findings (see **S2 Fig** and **S2 Table**). The degree of consistency, thereby, increases with the sample size of the reference study. The correspondence to expression levels of the core set of altered miRNAs involved in IBD [14] was estimated to be 75.0%. Additional 20 miRNAs investigated by Wu *et al.* (14 cases of active CD, 10 cases of active UC, 13 HC) agree with our data in 45.0% of the cases [16]. Evaluation of another 7 miRNAs identified in a study employing 20 UC and 20 HC samples shows a correspondence of 71.4% [48]. Finally, the studies conducted by Zahm *et al.* [18] (11 deregulated miRNAs identified in 46 cases of active CD and 32 HC) and Paraskevi *et al.* [22] (17 miRNAs, 128 cases of active CD, 88 cases of active UC, 162 HC) completely overlap with our results (correspondence of 100.0%). Interestingly, a large proportion of miRNAs that have previously been reported as being differentially expressed only for a certain group (CD or UC) appear to be deregulated similarly in both subtypes in our data and thus may be general IBD miRNAs. This effect may be explained due to the smaller sample size and/or higher variability in previous studies.

Classification with penalized support vector machines

Since there are various ways to construct complex classifiers for the distinction between CD, UC and HC (and IC, respectively), we assessed different types of penalized SVMs as well as the corresponding sets of miRNAs based on model performance and sparsity. Considering models incorporating 2 groups, differences in holdout-based median classifier performance of the penalization methods were small. However, due to its theoretic properties, the elastic SCAD SVM (median AUC = 0.97) was chosen for further investigation. Plots and tables illustrating the performance of the LASSO SVM (median AUC = 0.96), elastic net SVM (median AUC = 0.94) and SCAD SVM (median AUC = 0.95) are shown in **S4**–**S6 Figs** and S3–S5 **Tables.**

Fig 1 summarizes the elastic SCAD SVM's performance in solving the 16 different diagnostic problems measured by the area under the curve (AUC). The models incorporating 2 groups show stable superiority (median AUC = 0.97; 0.98 including vs. 0.95 excluding IC) in comparison to the models considering 3 groups (median AUC = 0.92; 0.93 including vs. 0.92 excluding IC) or 4 groups (median AUC = 0.85). In addition, these models provide remarkable sparsity (median percentage of miRNAs removed = 99.3%, 99.4% including vs. 99.2% excluding IC) and only marginal loss of performance compared to the standard SVM. As shown in **Table 2**, in terms of median sensitivity and specificity, the performance of the selected models can be estimated as 1.00 and 0.90, respectively (1.00 and 0.91 including IC, 1.00 and 0.90 excluding IC). The median balanced accuracy (BAC) was 0.95 (0.96 including IC, 0.95 excluding IC). Additional performance measures (e.g. median Matthews correlation coefficient (MCC) and Youden's index (YOUDEN)) are listed in **S3 Table** for each particular classifier.

The final set of markers selected for diagnostic application is shown in **Fig 2**. It includes 16 distinct miRNAs originating from elastic SCAD SVMs incorporating 2 groups: hsa-miR-34b-
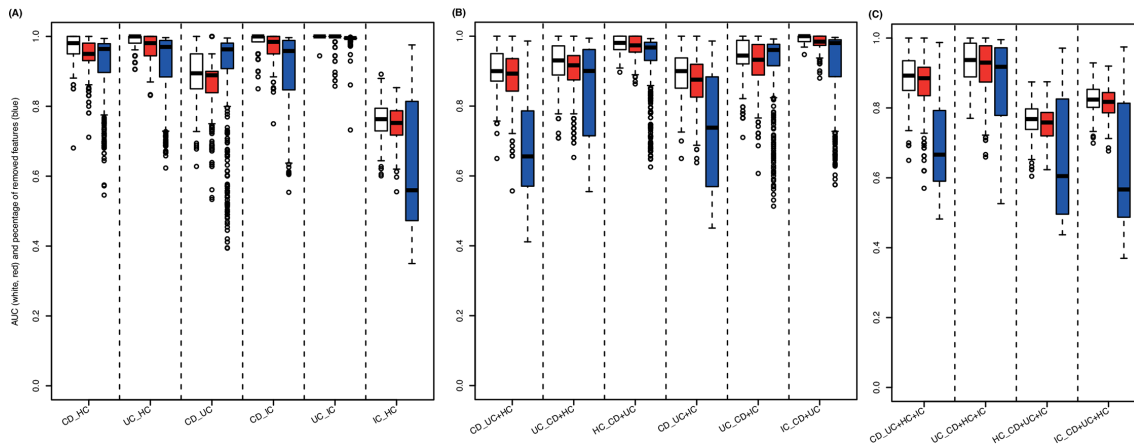
Fig 1. SVM classification results. Measured by the area under the ROC curve (AUC), classification performance is shown for models considering **(A)** 2 groups (CD vs. HC, UC vs. HC, CD vs. UC, CD vs. IC, UC vs. IC, IC vs. HC), **(B)** 3 groups (CD vs. UC+HC, UC vs. CD+HC, HC vs. CD+UC, CD vs. UC+IC, UC vs. CD+IC, IC vs. CD+UC) and **(C)** 4 groups (CD vs. UC+HC+IC, UC vs. CD+HC+IC, HC vs. CD+UC+IC, IC vs. CD+UC+HC). Performance of linear standard SVMs (considering every miRNA measured, white boxes) is compared to linear elastic SCAD SVMs (considering subsets of miRNAs measured, red boxes). In addition, as a measure of model complexity the percentage of miRNAs neglected for constructing the respective penalized SVMs are plotted (blue boxes).

doi:10.1371/journal.pone.0140155.g001

Table 2. Performance measures for the different classification models. Corresponding to the classification accuracy of the sparsest median performing penalized SVM (see **Fig 1**) for each classifier area under the ROC curve (AUC), sensitivity (SN = TPR, true positive rate), specificity (SP = TNR, true negative rate) and balanced accuracy (BAC = (SN+SP)/2) are shown.

#groups	classifier	AUC	SN	SP	BAC
	CD/HC	0.950	0.963	1.000	0.981
	UC/HC	0.981	1.000	0.900	0.950
	CD/UC	0.889	1.000	0.833	0.917
2	**median**	**0.950**	**1.000**	**0.900**	**0.950**
	CD/IC	0.984	1.000	0.909	0.955
	UC/IC	1.000	1.000	1.000	1.000
	IC/HC	0.752	0.750	0.765	0.757
	median	**0.984**	**1.000**	**0.909**	**0.955**
	CD/UC+HC	0.893	0.969	0.692	0.831
	UC/CD+HC	0.917	0.971	0.800	0.886
	HC/CD+UC	0.974	1.000	0.963	0.981
3	**median**	**0.917**	**0.971**	**0.800**	**0.886**
	CD/UC+IC	0.876	0.951	0.800	0.876
	UC/CD+IC	0.933	0.976	0.889	0.933
	IC/CD+UC	0.984	0.950	1.000	0.975
	median	**0.933**	**0.951**	**0.889**	**0.933**
	CD/UC+HC+IC	0.885	0.970	0.800	0.885
	UC/CD+HC+IC	0.930	0.985	0.800	0.893
4	HC/CD+UC+IC	0.758	0.830	0.708	0.769
	IC/CD+UC+HC	0.817	0.860	0.765	0.813
	median	**0.851**	**0.915**	**0.783**	**0.849**

doi:10.1371/journal.pone.0140155.t002

67

Fig 2. Expression profile of signature miRNAs. Median normalized expression levels for miRNAs considered by the final models (sparsest median performing elastic SCAD SVMs) used to distinguish CD, UC and HC are shown. The heat map was generated using a distance function based on Spearman's rank correlation coefficient and agglomerative hierarchical clustering using complete-linkage. Low and high expression levels are plotted using red and blue, respectively.

doi:10.1371/journal.pone.0140155.g002

3p, hsa-miR-142-5p, hsa-miR-205-5p, hsa-miR-424-5p, hsa-miR-570-3p, hsa-miR-885-5p, hsa-miR-1301-3p (CD vs. HC), hsa-miR-16-5p, hsa-miR-34b-3p, hsa-miR-99b-5p (UC vs.

68

Table 3. Comparison of classification approaches. The table shows the median classifier performance (AUC) of the classification problems considering different numbers of groups (2, 3 and 4) and models (SVM and RF). Performance of the standard SVM is compared to the elastic SCAD SVM (a). Performance of the RF per holdout sample using the miRNAs selected by the RF is compared to the RF per holdout sample using the miRNAs selected by the elastic SCAD SVM (b). Performance of the RF using the top 50% of the miRNAs most frequently selected by the RF across all runs is compared to the RF using the miRNAs selected by the median performing elastic SCAD SVM (c). For each comparison performance estimates based on miRNAs selected by the elastic SCAD SVM are enclosed in parentheses.

classification signature	2 groups	2 groups, no IC	3 groups	3 groups, no IC	4 groups
(a) SVM	0.990 (0.965)	0.981 (0.950)	0.938 (0.925)	0.931 (0.917)	0.858 (0.851)
(b) RF, holdout signature	0.992 (0.992)	0.985 (0.985)	0.978 (0.980)	0.969 (0.974)	0.919 (0.910)
(c) RF, top/median signature	0.996 (0.994)	0.992 (0.988)	0.982 (0.989)	0.977 (0.992)	0.941 (0.940)

doi:10.1371/journal.pone.0140155.t003

HC) and hsa-miR-34b-3p, hsa-miR-377-3p, hsa-miR-484, hsa-miR-574-5p, hsa-miR-656-3p, hsa-miR-744-5p, hsa-miR-1247-5p, hsa-miR-1908-5p (CD vs. UC, miRBase version 21 nomenclature). The corresponding models provide tools for the distinction of CD and UC as well as CD, UC and HC with remarkable small classification error rates of 3.1 and 3.3%, respectively (i.e. applying the proposed models will result in approximately 3 incorrect diagnoses per 100 tests). Notably, these estimates are not based on an independent dataset. Therefore, they are potentially optimistic but still provide a measure for the combined classifier's diagnostic value. This is confirmed by permutation tests showing significant deviation of the classification error from its random expectation. For further examples of classifier combinations and the corresponding size of miRNA signatures see S6 **Table**.

Validation using random forests

A second, independent machine learning approach, random forest analysis (RFs), was employed to validate our SVM-based miRNA signatures. As shown in **Table 3**, random forest analyses confirmed our SVM results. In the case of models considering 2 groups, the performance differences were small (AUC = 0.990 for the SVM using the entirety of the miRNAs, AUC = 0.992 for the RF per holdout sample using miRNAs selected by the RF and AUC = 0.996 for the RF using the top 50% of the miRNAs most frequently selected by the RF across all runs). When using the miRNAs selected by the elastic SCAD SVM for training the RF in the same way, highly accurate models were obtained: AUC = 0.965 for the elastic SCAD SVM, AUC = 0.992 for the RF per holdout sample using miRNAs selected by the elastic SCAD SVM and AUC = 0.994 for the RF using the miRNAs considered by the median performing elastic SCAD SVM. These results strongly support the validity of the miRNA combinations chosen as putative diagnostic markers by the SVM approach.

Target genes of the diagnostic miRNA signature correlate with susceptibility genes

To assess the potential biological significance of miRNAs within the signatures revealed by the machine-learning approaches we correlated previous knowledge about disease relevance of IBD-related genes to the experimentally validated target genes of the miRNA signatures (results summarized in **Table 4** and **S7 Table**). Irrespective of the miRNA signature tested (CD vs. HC, UC vs. HC or CD vs. UC), we observed an overlap of miRNA target genes and published IBD-related genes or, on the other hand, genes within known IBD susceptibility loci. Thus, target genes of signature miRNAs used for the distinction of CD and HC are significantly enriched for loci known to be associated with CD ($p = 3.22 \cdot 10^{-3}$), UC ($p = 1.09 \cdot 10^{-3}$) and suggestive for IBD ($p = 4.37 \cdot 10^{-2}$). Targets of signature miRNAs used for the distinction of UC and HC show suggestive enrichment for loci associated with CD ($p = 3.34 \cdot 10^{-2}$). Considering

Table 4. Signature miRNAs regulate target genes previously identified as IBD-risk genes. Both CD and UC diagnostic signatures contain several miR-NAs that regulate experimentally validated target genes known to be involved in IBD-related phenotypes in humans and/or mice. Genes marked with * have even been reported as candidate genes in susceptibility loci identified in recent IBD GWAS.

signature	miRNA	target gene	function/disease implication	reference
CD/HC	hsa-miR-205	LRRK2 *	susceptibility gene for CD	[63]
		SHIP2/INPPL1	regulator of PI3K, therapeutical target in inflammation	[64]
		ZEB1	regulates intestinal cell growth	[65]
		E2F1	activation promoted by chronic inflammation	[66]
		ERBB3	inhibits treatment of IBD	[67]
	hsa-miR-142-5p	NFE2L2/NRF2	susceptibility for DSS-induced colitis	[68]
	hsa-miR-424	MYB	colonic epithelial disruption by mir-150	[69]
		CUL2 *	susceptibility gene for CD	[70]
		PU.1	role in T-cell mediated colitis	[71]
	hsa-miR-34b	HNF4A *	susceptibility gene for early onset CD	[72]
		CREB1	diverse implications in CD	[73]
UC/HC	hsa-miR-34b	HNF4A *	susceptibility gene for UC	[74]
		NOTCH1	regulator of intestinal epithelial barrier	[75]
		c-MET/HGFR	upregulated in UC	[76]
		CAV1	upregulated in UC inflamed tissue	[77]
	hsa-miR-99b	RAVER2 *	susceptibility gene for UC	[78]
		mTOR	inhibition depletes mouse colitis	[79]
	hsa-miR-16	HMGA1/2	P-ANCA autoantigens	[80]
		ACVR2a	associated with IBD-related CRC	[81]

doi:10.1371/journal.pone.0140155.t004

the targets of the complete set of miRNAs used for the distinction of CD, UC and HC suggestive enrichment is observed for previously published susceptibility loci of CD ($p = 4.80 \cdot 10^{-3}$) and UC ($p = 4.80 \cdot 10^{-3}$).

In a next step we investigated whether previously identified genetic variation in the IBD susceptibility genes could directly play a role in miRNA-target gene interaction. We used dbSNP annotations of the human genome provided by the UCSC genome browser to identify SNPs that could interfere with miRNA binding sites. As a result we found that most 3'-UTRs of the analyzed IBD-risk genes indeed exhibit genetic variation (SNPs and small InDels) but mostly not in the respective signature-miRNA binding site regions. Only for hsa-miR-99b, which is part of the UC signature, we were able to identify potentially interesting SNPs located in the essential miRNA binding site seed regions of RAVER2 (rs183861354, chr1:64831085, G>A) and mTOR (rs375505566, chr1:11107188, G>A). Strikingly, both SNPs change the same nucleotide position within the seed region of the miRNA binding site. Whether this single nucleotide variant affects the binding behavior and as a consequence the gene functions in IBD cases compared to healthy controls, remains to be shown.

Discussion

In this study we compared miRNA expression profiles of whole peripheral blood samples from patients with inflammatory bowel disease (Crohn's disease and ulcerative colitis) to healthy controls and "disease controls". We were able to confirm significantly deregulated miRNAs in blood that were previously reported by others and could further add new candidates to the catalogue of IBD-associated miRNAs. To our knowledge this study represents the largest (both in terms of samples and measured miRNAs) blood-based miRNA-expression study for IBD published to date. Our analysis, however, was focused on the identification of disease specific,

diagnostic classification signatures derived from the overall miRNA expression profiles irrespective of single miRNA deregulation.

miRNAs are often referred to as "blood-based biomarkers" for diagnosing disease or monitoring disease progression. As it has been shown for several types of cancer this holds true as long as a relatively stable condition, such as a recurrent aberrant gene expression in certain tissues or exosomal miRNA content can be measured repeatedly. Concerning blood-based miRNA expression in inflammatory or auto-immune diseases, however, the assumption of stable conditions is often violated. Numerous known comorbidities as well as environmental and life-style factors, treatment and disease activity may influence miRNA levels in the blood stream and lead to intra- and interindividual miRNA-expression variability. Also general factors like blood cell composition, depending on the type of disease may vary significantly, and hence impact miRNA levels in peripheral blood. Thus, instead of aiming to identify single miRNA "biomarkers", to enhance predictive power it appears more promising to investigate complex predictors that are based on larger numbers of miRNAs. In this way, besides simple deregulation, also certain combinations of regulatory effects are taken into account for diagnostic or predictive models.

In this work we demonstrated the use of machine-learning techniques to construct IBD-specific miRNA signatures and we were able to reveal highly accurate classification models that distinguish healthy and diseased individuals as well as the two main IBD subtypes and other inflammatory conditions from each other. Furthermore, a minimal set of not more than 16 miRNAs, being sufficient for sensitive and specific classification, holds great promises and should be further evaluated in independent sample panels.

The here-investigated models represent solutions to construct classifiers for miRNA expression data but they also exhibit some limitations, most notably the limited generalizability of the models to other technologies. All models are trained based on the same type of data that originate from a certain technology (here the Geniom Array). Application of these models to independent samples in a clinical or diagnostic setting would always require to remove technology biases. In addition to that, the here-presented classifiers remain restricted to the set of miRNAs that are present on the microarray used to detect differential expression. Future studies utilizing next generation sequencing (NGS) will presumably overcome this limitation as all present miRNAs in a sample are theoretically detectable by this technology. Furthermore, implementing approaches that include more levels of available information e.g. genetic variants, microbiome data or clinical data from electronic health records (that include information on differential diagnoses, medication, disease activity, etc.) will potentially add to the predictive power needed for highly sensitive and specific classification.

Regularized instances of support vector machines incorporate penalties for model complexity to prevent overfitting and to provide sparse solutions. In the here-presented study this property is used to obtain small sets of miRNAs suitable for diagnostic application. It is expected that miRNAs essential for solving a particular classification problem likewise are selected by random forests using the recurrent relative variable importance. However, this approach does not aim at selecting a minimal set of features so that one does not expect miRNA signatures to be fully overlapping. To obtain more comparable results, future studies might consider regularized random forests as introduced by Deng and Runger [82]. In this work the miRNA signature selected using the elastic SCAD SVM was confirmed by comparably high classification performance of random forests as an independent classification approach. For this purpose aforementioned limitations can be neglected.

To obtain a model applicable with high accuracy to independent data we chose the sparsest median performing elastic SCAD SVM along with the corresponding miRNA signature. Both, the regularization approach and the comprehensive holdout sampling decrease the model's

probability of being overfitted to the dataset generated for this study. However, due to correlating expression profiles it is expected that models with matching accuracy potentially incorporate differing miRNAs. For the same reason more complex signatures may exist which merely incoporate additional highly correlated miRNAs.

Classifiers for complex diagnostic problems were constructed by majority voting of simpler models. As shown in this study, this approach results in remarkable low classification error rates. However, follow-up studies could potentially incorporate the estimation of class probabilities to enhance the interpretability of the classification results.

To get insights into functional implications of the miRNAs contained in the revealed IBD signature, we screened current databases for experimentally validated miRNA-target gene interactions. Notably, a considerable fraction of the target genes within the IBD miRNA signatures has been implicated in intestinal diseases (see **Table 4**). Many of those targets were identified in recent IBD GWAS but most of the genetic variation detected does not correlate (and thus not interfere) with miRNA regulatory binding sites. Only the hsa-mir-99b binding sites in the 3'-UTRs of the IBD susceptibility gene RAVER2, a ribonucleoprotein (hnRNP) involved in regulation of splicing and mTOR, a serine/threonine proteine kinase, shown to be involved in activation of autophagy, represent good candidates for further experimental investigation. In the future, more complete data on genetic varation in 3'-UTRs of IBD related genes will supposedly come from whole genome sequencing approaches and will thus enable for more complete analyses of miRNA target genes. In a recent review on genetic studies in IBD Liu and Anderson [83] conclude that most of the identified GWAS loci actually reside in noncoding regions of the genome and that a vast number of these noncoding variants will likely play a role in gene regulation. miRNAs are certainly an important part of the regulatory machinery of the genome, but besides their utility in diagnostics, miRNA signatures might also give valuable insights into disease development and progression.

Supporting Information

S1 Fig. MDS (multidimensional scaling) plots for visualization of background-subtracted intensity values. Background-subtracted intensity values normalized using variance stabilization (**A**) before and (**B**) after median centering based on the batches observed for healthy controls. The corresponding medians are indicated by black circles. MDS was performed using a distance function based on Spearman's rank correlation coefficient. Data points of each group are represented by their α-shape (generalized convex hull). The second plot visualizes the batch-corrected normalized data used for diagnostic classification.
(TIFF)

S2 Fig. Median expression levels of miRNAs previously published as being deregulated in CD, UC and HC. The horizontal side bar indicates the correspondence between the literature and the dataset used for this study. Measurements with directions of effect deviating from the literature are marked using black bars. The heat map was generated using a distance function based on Spearman's rank correlation coefficient and agglomerative hierarchical clustering using complete-linkage. Low and high expression levels are plotted using red and blue, respectively.
(TIFF)

S3 Fig. Median expression profiles of significantly deregulated miRNAs in CD, UC and HC. For each pair of groups two-sample t-tests were applied. Deregulation was considered as being significant for Holm-corrected p-values <0.05. Not significantly differentially expressed miRNAs were neglected. 667 out of 863 miRNAs were differentially deregulated in any of the

comparisons.
(TIFF)

S4 Fig. Classification results for LASSO SVM. Measured by the area under the ROC curve (AUC) classification performance is shown for models considering **(A)** 2 groups (CD vs. HC, UC vs. HC, CD vs. UC, CD vs. IC, UC vs. IC, IC vs. HC), **(B)** 3 groups (CD vs. UC+HC, UC vs. CD+HC, HC vs. CD+UC, CD vs. UC+IC, UC vs. CD+IC, IC vs. CD+UC) and **(C)** 4 groups (CD vs. UC+HC+IC, UC vs. CD+HC+IC, HC vs. CD+UC+IC, IC vs. CD+UC+HC). Performance of linear standard SVMs (considering every miRNA measured, white boxes) is compared to linear LASSO SVMs (considering subsets of miRNAs measured, red boxes). In addition, as a measure of model complexity the percentage of miRNAs neglected for constructing the respective penalized SVMs are plotted (blue boxes).
(TIFF)

S5 Fig. Classification results for elastic net SVM. Measured by the area under the ROC curve (AUC) classification performance is shown for models considering **(A)** 2 groups (CD vs. HC, UC vs. HC, CD vs. UC, CD vs. IC, UC vs. IC, IC vs. HC), **(B)** 3 groups (CD vs. UC+HC, UC vs. CD+HC, HC vs. CD+UC, CD vs. UC+IC, UC vs. CD+IC, IC vs. CD+UC) and **(C)** 4 groups (CD vs. UC+HC+IC, UC vs. CD+HC+IC, HC vs. CD+UC+IC, IC vs. CD+UC+HC). Performance of linear standard SVMs (considering every miRNA measured, white boxes) is compared to linear elastic net SVMs (considering subsets of miRNAs measured, red boxes). In addition, as a measure of model complexity the percentage of miRNAs neglected for constructing the respective penalized SVMs are plotted (blue boxes).
(TIFF)

S6 Fig. Classification results for SCAD SVM. Measured by the area under the ROC curve (AUC) classification performance is shown for models considering **(A)** 2 groups (CD vs. HC, UC vs. HC, CD vs. UC, CD vs. IC, UC vs. IC, IC vs. HC), **(B)** 3 groups (CD vs. UC+HC, UC vs. CD+HC, HC vs. CD+UC, CD vs. UC+IC, UC vs. CD+IC, IC vs. CD+UC) and **(C)** 4 groups (CD vs. UC+HC+IC, UC vs. CD+HC+IC, HC vs. CD+UC+IC, IC vs. CD+UC+HC). Performance of linear standard SVMs (considering every miRNA measured, white boxes) is compared to linear SCAD SVMs (considering subsets of miRNAs measured, red boxes). In addition, as a measure of model complexity the percentage of miRNAs neglected for constructing the respective penalized SVMs are plotted (blue boxes).
(TIFF)

S7 Fig. Comparison of SVM and random forest. Measured by the area under the ROC curve (AUC) classification performance is shown for models considering **(A)** 2 groups (CD vs. HC, UC vs. HC, CD vs. UC, CD vs. IC, UC vs. IC, IC vs. HC), **(B)** 3 groups (CD vs. UC+HC, UC vs. CD+HC, HC vs. CD+UC, CD vs. UC+IC, UC vs. CD+IC, IC vs. CD+UC) and **(C)** 4 groups (CD vs. UC+HC+IC, UC vs. CD+HC+IC, HC vs. CD+UC+IC, IC vs. CD+UC+HC). Classification performance of the linear elastic SCAD SVM (white box) is compared to a **Random forests per holdout sample considering variables selected using the SVM** (red box) and the Random forest itself (blue box), respectively.
(TIFF)

S8 Fig. Comparison of SVM and random forest. Measured by the area under the ROC curve (AUC) classification performance is shown for models considering **(A)** 2 groups (CD vs. HC, UC vs. HC, CD vs. UC, CD vs. IC, UC vs. IC, IC vs. HC), **(B)** 3 groups (CD vs. UC+HC, UC vs. CD+HC, HC vs. CD+UC, CD vs. UC+IC, UC vs. CD+IC, IC vs. CD+UC) and **(C)** 4 groups (CD vs. UC+HC+IC, UC vs. CD+HC+IC, HC vs. CD+UC+IC, IC vs. CD+UC+HC).

Classification performance of the linear elastic SCAD SVM (white box) is compared to a **Random forests considering variables selected using the median performing SVM** (red box). Additionally, Random forests were trained with the top 50% of the variables ranked by their frequency of selection (blue box).

(TIFF)

S1 Table. Differential expression analysis. For each binary combination of groups t-tests for differential miRNA expression were conducted. The test results are summarized by the fold change (fc), the t-statistic (t), the p-value (p) and the p-values adjusted for multiple testing using Holm-correction (padj).

(XLSX)

S2 Table. miRNAs previously described to be deregulated in IBD. Tables were adapted from Chen *et al.* **(A)** and Coscun et al. **(B)**, respectively. For binary comparisons (CD vs. HC, UC vs. HC, IBD vs. HC and CD vs. UC). Directions of effect known from the literature as well as measured by the microarray used for this study are summarized.

(XLSX)

S3 Table. Performance measures for LASSO SVM. Corresponding to the classification accuracy of the sparsest median performing penalized SVM (see S4 Fig) for each classification task area under the ROC curve (AUC), Matthews correlation coefficient (MCC), balanced accuracy (BAC), Youden's index (YOUDEN), sensitivity (SN = TPR), specificity (SP = TNR), positive predictive value (PPV), false discovery rate (FDR), negative predictive value (NPV) and false omission rate (FOR) are shown.

(XLSX)

S4 Table. Performance measures for elastic net SVM. Corresponding to the classification accuracy of the sparsest median performing penalized SVM (see **S5 Fig**) for each classification task area under the ROC curve (AUC), Matthews correlation coefficient (MCC), balanced accuracy (BAC), Youden's index (YOUDEN), sensitivity (SN = TPR), specificity (SP = TNR), positive predictive value (PPV), false discovery rate (FDR), negative predictive value (NPV) and false omission rate (FOR) are shown.

(XLSX)

S5 Table. Performance measures for SCAD SVM. Corresponding to the classification accuracy of the sparsest median performing penalized SVM (see **S6 Fig**) for each classification task area under the ROC curve (AUC), Matthews correlation coefficient (MCC), balanced accuracy (BAC), Youden's index (YOUDEN), sensitivity (SN = TPR), specificity (SP = TNR), positive predictive value (PPV), false discovery rate (FDR), negative predictive value (NPV) and false omission rate (FOR) are shown.

(XLSX)

S6 Table. Exemplary diagnostic application. The final median performing models were used to predict the disease status based on each individual's miRNA expression data. For each combined classifier, constructed using majority voting, the number of groups, n(groups), considered by the atomic models as well as the respective number of miRNAs, n(mirnas), and unique miRNAs, n(unique), are shown. In addition the classification errors per individual's group (ε(CD), ε(UC), ε(HC) and ε(IC)) were estimated. Furthermore, for each classifier the mean classification error (mean) as well as the corresponding permutation based Z-score (Z(mean)) were calculated. Z-scores corresponding to p-values lower than the significance threshold of 0.05 are marked using [*].

(XLSX)

PLOS | ONE

S7 Table. Enrichment analysis for validated signature miRNA targets. The proportion of validated targets of the miRNAs selected for diagnostic prediction (CD vs. HC, UC vs. HC, CD vs. UC and CD vs. UC vs. HC) known to be coded at risk loci (CD, UC and IBD) is compared to the proportion of general miRNA targets known to be coded at risk loci (CD, UC and IBD). Targets of signature miRNA coded at risk and non-risk loci are denoted as C_r and C, repectively. miRNA targets excluding signature targets coded at risk and non-risk loci are denoted as R_r and R, repectively. The total number of validated miRNA targets is denoted as N. Enrichment analysis was performed using Fisher's exact test, resulting p-values are marked as being nominal significant (*, $p < 0.05$) and significant after Bonferroni correction for multiple testing (**; $p_{adj} < 0.05$).
(XLSX)

S8 Table. Raw miRNA expression data. Background subtracted microarray intensities from Geniom Wizard Software that were used to infer miRNA expression levels.
(XLSX)

Acknowledgments

Author Contributions

Conceived and designed the experiments: MH GH AF. Performed the experiments: MH GH. Analyzed the data: MH GH FD S. Szymczak. Contributed reagents/materials/analysis tools: ZGD AE AK S. Schreiber. Wrote the paper: MH GH FD S. Szymczak AF.

References

1. Dotan I, Fishman S, Dgani Y, Schwartz M, Karban A, Lerner A, et al. Antibodies against laminaribioside and chitobioside are novel serologic markers in Crohn's disease. Gastroenterology. 2006; 131: 366–78. doi: 10.1053/j.gastro.2006.04.030 PMID: 16890590

2. Moum B, Ekbom A, Vatn MH, Aadland E, Sauar J, Lygren I, et al. Inflammatory bowel disease: re-evaluation of the diagnosis in a prospective population based study in south eastern Norway. Gut. 1997; 40: 328–32. PMID: 9135520

3. Lewis JD. The utility of biomarkers in the diagnosis and therapy of inflammatory bowel disease. Gastroenterology. 2011; 140: 1817–1826.e2. doi: 10.1053/j.gastro.2010.11.058 PMID: 21530748

4. Iskandar HN, Ciorba M. Biomarkers in inflammatory bowel disease: current practices and recent advances. Transl Res. Mosby, Inc.; 2012; 159: 313–25. doi: 10.1016/j.trsl.2012.01.001 PMID: 22424434

5. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. Nature Publishing Group; 2012; 491: 119–24. doi: 10.1038/nature11582 PMID: 23128233

6. Jostins L, Barrett JC. Genetic Risk Prediction in Complex Disease. Hum Mol Genet. 2011; 1–7. doi: 10.1093/hmg/ddr378

PLOS ONE

7. Jakobsdottir J, Gorin MB, Conley YP, Ferrell RE, Weeks DE. Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. PLoS Genet. 2009; 5: e1000337. doi: 10.1371/journal.pgen.1000337 PMID: 19197355

8. Clark PM, Dawany N, Dampier W, Byers SW, Pestell RG, Tozeren A. Bioinformatics analysis reveals transcriptome and microRNA signatures and drug repositioning targets for IBD and other autoimmune diseases. Inflamm Bowel Dis. 2012; 18: 2315–33. doi: 10.1002/ibd.22958 PMID: 22488912

9. Gologan S, Iacob R, Iancu D, Iacob S, Cotruta B, Vadan R, et al. Inflammatory gene expression profiles in Crohn's disease and ulcerative colitis: a comparative analysis using a reverse transcriptase multiplex ligation-dependent probe amplification protocol. J Crohns Colitis. European Crohn's and Colitis Organisation; 2013; 7: 622–30. doi: 10.1016/j.crohns.2012.08.015 PMID: 23014361

10. van Lierop PPE, Swagemakers SM, de Bie CI, Middendorp S, van Baarlen P, Samsom JN, et al. Gene expression analysis of peripheral cells for subclassification of pediatric inflammatory bowel disease in remission. PLoS One. 2013; 8: e79549. doi: 10.1371/journal.pone.0079549 PMID: 24260248

11. Montero-Meléndez T, Llor X, García-Planella E, Perretti M, Suárez A. Identification of novel predictor classifiers for inflammatory bowel disease by gene expression profiling. PLoS One. 2013; 8: e76235. doi: 10.1371/journal.pone.0076235 PMID: 24155895

12. Granlund AVB, Flatberg A, Østvik AE, Drozdov I, Gustafsson BI, Kidd M, et al. Whole genome gene expression meta-analysis of inflammatory bowel disease colon mucosa demonstrates lack of major differences between Crohn's disease and ulcerative colitis. PLoS One. 2013; 8: e56818. doi: 10.1371/journal.pone.0056818 PMID: 23468882

13. Brest P, Lapaquette P, Souidi M, Lebrigand K, Cesaro A, Vouret-Craviari V, et al. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. Nat Genet. Nature Publishing Group; 2011; 43: 242–5. doi: 10.1038/ng.762 PMID: 21278745

14. Chen W-X, Ren L-H, Shi R-H. Implication of miRNAs for inflammatory bowel disease treatment: Systematic review. World J Gastrointest Pathophysiol. 2014; 5: 63–70. doi: 10.4291/wjgp.v5.i2.63 PMID: 24891977

15. Iborra M, Bernuzzi F, Correale C, Vetrano S, Fiorino G, Beltrán B, et al. Identification of serum and tissue micro-RNA expression profiles in different stages of inflammatory bowel disease. Clin Exp Immunol. 2013; 173: 250–8. doi: 10.1111/cei.12104 PMID: 23607522

16. Wu F, Guo NJ, Tian H, Marohn M, Gearhart S, Bayless TM, et al. Peripheral blood microRNAs distinguish active ulcerative colitis and Crohn's disease. Inflamm Bowel Dis. 2011; 17: 241–50. doi: 10.1002/ibd.21450 PMID: 20812331

17. Wu F, Zhang S, Dassopoulos T, Harris ML, Bayless TM, Meltzer SJ, et al. Identification of microRNAs associated with ileal and colonic Crohn's disease. Inflamm Bowel Dis. 2010; 16: 1729–38. doi: 10.1002/ibd.21267 PMID: 20848482

18. Zahm AM, Thayu M, Hand NJ, Horner A, Leonard MB, Friedman JR. Circulating MicroRNA Is a Biomarker of Pediatric Crohn Disease. J Pediatr Gastroenterol Nutr. 2011; 53: 26–33. doi: 10.1097/MPG.0b013e31822200cc PMID: 21546856

19. Fasseu M, Tréton X, Guichard C, Pedruzzi E, Cazals-Hatem D, Richard C, et al. Identification of restricted subsets of mature microRNA abnormally expressed in inactive colonic mucosa of patients with inflammatory bowel disease. PLoS One. 2010; 5. doi: 10.1371/journal.pone.0013160

20. Zahm AM, Hand NJ, Tsoucas DM, Le Guen CL, Baldassano RN, Friedman JR. Rectal microRNAs are perturbed in pediatric inflammatory bowel disease of the colon. J Crohns Colitis. European Crohn's and Colitis Organisation; 2014; doi: 10.1016/j.crohns.2014.02.012

21. Wu F, Zikusoka M, Trindade A, Dassopoulos T, Harris ML, Bayless TM, et al. MicroRNAs are differentially expressed in ulcerative colitis and alter expression of macrophage inflammatory peptide-2 alpha. Gastroenterology. 2008; 135: 1624–1635.e24. doi: 10.1053/j.gastro.2008.07.068 PMID: 18835392

22. Paraskevi A, Theodoropoulos G, Papaconstantinou I, Mantzaris G, Nikiteas N, Gazouli M. Circulating MicroRNA in inflammatory bowel disease. J Crohns Colitis. European Crohn's and Colitis Organisation; 2012; 6: 900–4. doi: 10.1016/j.crohns.2012.02.006 PMID: 22386737

23. Ghorpade DS, Sinha AY, Holla S, Singh V, Balaji KN. NOD2-nitric oxide-responsive microRNA-146a activates sonic hedgehog signaling to orchestrate inflammatory responses in murine model of inflammatory bowel disease. J Biol Chem. 2013; doi: 10.1074/jbc.M113.492496

24. Kanaan Z, Rai SN, Eichenberger MR, Barnes C, Dworkin AM, Weller C, et al. Differential microRNA expression tracks neoplastic progression in inflammatory bowel disease-associated colorectal cancer. Hum Mutat. 2012; 33: 551–60. doi: 10.1002/humu.22021 PMID: 22241525

25. Li Z, Wu F, Brant SR, Kwon JH. IL-23 receptor regulation by Let-7f in human CD4+ memory T cells. J Immunol. 2011; 186: 6182–90. doi: 10.4049/jimmunol.1000917 PMID: 21508257

76

Sparse Modeling for IBD Diagnostics

26. Koukos G, Polytarchou C, Kaplan JL, Morley-Fletcher A, Gras-Miralles B, Kokkotou E, et al. MicroRNA-124 regulates STAT3 expression and is down-regulated in colon tissues of pediatric patients with ulcerative colitis. Gastroenterology. Elsevier, Inc; 2013; 145: 842–52.e2. doi: 10.1053/j.gastro.2013.07.001 PMID: 23856509

27. Chuang AY, Chuang JC, Zhai Z, Wu F, Kwon JH. NOD2 expression is regulated by microRNAs in colonic epithelial HCT116 cells. Inflamm Bowel Dis. 2014; 20: 126–35. doi: 10.1097/01.MIB.0000436954.70596.9b PMID: 24297055

28. Chen Y, Wang C, Liu Y, Tang L, Zheng M, Xu C, et al. miR-122 targets NOD2 to decrease intestinal epithelial cell injury in Crohn's disease. Biochem Biophys Res Commun. Elsevier Inc.; 2013; 438: 133–9. doi: 10.1016/j.bbrc.2013.07.040 PMID: 23872065

29. Feng X, Wang H, Ye S, Guan J, Tan W, Cheng S, et al. Up-regulation of microRNA-126 may contribute to pathogenesis of ulcerative colitis via regulating NF-kappaB inhibitor IκBα. PLoS One. 2012; 7: e52782. doi: 10.1371/journal.pone.0052782 PMID: 23285182

30. Thu Nguyen HT, Dalmasso G, Müller S, Carrière J, Seibold F, Darfeuille-Michaud A. Crohn's Disease-associated Adherent Invasive Escherichia coli Affect Levels of microRNAs in Intestinal Epithelial Cells to Reduce Autophagy. Gastroenterology. Elsevier Ltd; 2013; doi: 10.1053/j.gastro.2013.10.021

31. Zhai Z, Wu F, Chuang AY, Kwon JH. miR-106b Fine Tunes ATG16L1 Expression and Autophagic Activity in Intestinal Epithelial HCT116 Cells. 2013; 19: 17–19. doi: 10.1097/MIB.0b013e31829e71cf

32. Shi C, Liang Y, Yang J, Xia Y, Chen H, Han H, et al. MicroRNA-21 knockout improve the survival rate in DSS induced fatal colitis through protecting against inflammation and tissue injury. PLoS One. 2013; 8: e66814. doi: 10.1371/journal.pone.0066814 PMID: 23826144

33. Yang Y, Ma Y, Shi C, Chen H, Zhang H, Chen N, et al. Overexpression of miR-21 in patients with ulcerative colitis impairs intestinal epithelial barrier function through targeting the Rho GTPase RhoB. Biochem Biophys Res Commun. Elsevier Inc.; 2013; 434: 746–52. doi: 10.1016/j.bbrc.2013.03.122 PMID: 23583411

34. Olaru AV, Selaru FM, Mori Y, Vazquez C, David S, Paun B, et al. Dynamic changes in the expression of MicroRNA-31 during inflammatory bowel disease-associated neoplastic transformation. Inflamm Bowel Dis. 2011; 17: 221–31. doi: 10.1002/ibd.21359 PMID: 20848542

35. Olaru AV, Yamanaka S, Vazquez C, Mori Y, Cheng Y, Abraham JM, et al. MicroRNA-224 negatively regulates p21 expression during late neoplastic progression in inflammatory bowel disease. Inflamm Bowel Dis. 2013; 19: 471–80. doi: 10.1097/MIB.0b013e31827e78eb PMID: 23399735

36. Ludwig K, Fassan M, Mescoli C, Pizzi M, Balistreri M, Albertoni L, et al. PDCD4/miR-21 dysregulation in inflammatory bowel disease-associated carcinogenesis. Virchows Arch. 2013; 462: 57–63. doi: 10.1007/s00428-012-1345-5 PMID: 23224068

37. Cortes C, Vapnik V. Support-Vector Networks. Mach Learn. 1995; 20: 273–297.

38. Breiman L. Random Forests. Mach Learn. 2001; 45: 5–32.

39. Keller A, Leidinger P, Bauer A, ElSharawy A, Haas J, Backes C, et al. Toward the blood-borne miR-Nome of human diseases. Nat Methods. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2011; 8: 841–843. doi: 10.1038/nmeth.1682 PMID: 21892151

40. Keller A, Leidinger P, Steinmeyer F, Stähler C, Franke A, Hemmrich-Stanisak G, et al. Comprehensive analysis of microRNA profiles in multiple sclerosis including next-generation sequencing. Mult Scler. 2014; 20: 295–303. doi: 10.1177/1352458513496343 PMID: 23836875

41. Keller A, Leidinger P, Lange J, Borries A, Schroers H, Scheffler M, et al. Multiple sclerosis: microRNA expression profiles accurately differentiate patients with relapsing-remitting disease from healthy controls. PLoS One. 2009; 4: e7440. doi: 10.1371/journal.pone.0007440 PMID: 19823682

42. Keller A, Backes C, Leidinger P, Kefer N, Boisguerin V, Barbacioru C, et al. Next-generation sequencing identifies novel microRNAs in peripheral blood of lung cancer patients. Mol Biosyst. 2011; 7: 3187–99. doi: 10.1039/c1mb05353a PMID: 22027949

43. Abu-Halima M, Hammadeh M, Backes C, Fischer U, Leidinger P, Lubbad AM, et al. A panel of five microRNAs as potential biomarkers for the diagnosis and assessment of male infertility. Fertil Steril. 2014; doi: 10.1016/j.fertnstert.2014.07.001

44. Lajer CB, Nielsen FC, Friis-Hansen L, Norrild B, Borup R, Garnæs E, et al. Different miRNA signatures of oral and pharyngeal squamous cell carcinomas: a prospective translational study. Br J Cancer. 2011; 104: 830–40. doi: 10.1038/bjc.2011.29 PMID: 21326242

45. Keutgen XM, Filicori F, Crowley MJ, Wang Y, Scognamiglio T, Hoda R, et al. A panel of four miRNAs accurately differentiates malignant from benign indeterminate thyroid lesions on fine needle aspiration. Clin Cancer Res. 2012; 18: 2032–8. doi: 10.1158/1078-0432.CCR-11-2487 PMID: 22351693

77

46. Patnaik SK, Yendamuri S, Kannisto E, Kucharczuk JC, Singhal S, Vachani A. MicroRNA expression profiles of whole blood in lung adenocarcinoma. PLoS One. 2012; 7: e46045. doi: 10.1371/journal.pone.0046045 PMID: 23029380

47. Miotto P, Mwangoka G, Valente IC, Norbis L, Sotgiu G, Bosu R, et al. miRNA signatures in Sera of patients with active pulmonary tuberculosis. PLoS One. 2013; 8: e80149. doi: 10.1371/journal.pone.0080149 PMID: 24278252

48. Duttagupta R, DiRienzo S, Jiang R, Bowers J, Gollub J, Kao J, et al. Genome-wide maps of circulating miRNA biomarkers for ulcerative colitis. PLoS One. 2012; 7: e31241. doi: 10.1371/journal.pone.0031241 PMID: 22359580

49. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Springer Series in Statistics). 2nd ed. 2009.

50. Güimil R, Beier M, Scheffler M, Rebscher H, Funk J, Wixmerten A, et al. Geniom technology—the benchtop array facility. Nucleosides Nucleotides Nucleic Acids. 22: 1721–3. doi: 10.1081/NCN-120023122 PMID: 14565504

51. Vorwerk S, Ganter K, Cheng Y, Hoheisel J, Stähler PF, Beier M. Microfluidic-based enzymatic on-chip labeling of miRNAs. N Biotechnol. 25: 142–9. doi: 10.1016/j.nbt.2008.08.005 PMID: 18786664

52. Huber W, von Heydebreck A, Sueltmann H, Poustka A, Vingron M. Parameter estimation for the calibration and variance stabilization of microarray data. Stat Appl Genet Mol Biol. 2003; 2. doi: 10.2202/1544-6115.1008

53. Bradley PS, Mangasarian OL. Feature Selection via Concave Minimization and Support Vector Machines. Proceedings of the Fifteenth International Conference on Machine Learning ICML '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1998. pp. 82–90.

54. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Statistical Methodol. 2005; 67: 301–320.

55. Zhang HH, Ahn J, Lin X, Park C. Gene selection using support vector machines with non-convex penalty. Bioinformatics. 2006; 22: 88–95. doi: 10.1093/bioinformatics/bti736 PMID: 16249260

56. Becker N, Toedt G, Lichter P, Benner A. Elastic SCAD as a novel penalization method for SVM classification tasks in high-dimensional data. BMC Bioinformatics. BioMed Central Ltd; 2011; 12: 138. doi: 10.1186/1471-2105-12-138 PMID: 21554689

57. Becker N, Werft W, Toedt G, Lichter P, Benner A. penalizedSVM: a R-package for feature selection SVM classification. Bioinformatics. 2009; 25: 1711–1712. doi: 10.1093/bioinformatics/btp286 PMID: 19398451

58. Friedman JH. Another approach to polychotomous classification. 1996.

59. Szymczak S, Holzinger E, Dasgupta A, Malley J, Molloy A, Mills J, et al. r2VIM: A new variable selection method for random forests in genome-wide association studies. submitted.

60. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods. 2009; 14: 323–48. doi: 10.1037/a0016973 PMID: 19968396

61. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005; 21: 3940–1. doi: 10.1093/bioinformatics/bti623 PMID: 16096348

62. Hsu S-D, Tseng Y-T, Shrestha S, Lin Y-L, Khaleel A, Chou C-H, et al. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. Nucleic Acids Res. 2014; 42: D78–85. doi: 10.1093/nar/gkt1266 PMID: 24304892

63. Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet. 2010/11/26 ed. 2010; 42: 1118–1125. doi: 10.1038/ng.717 PMID: 21102463

64. Blunt MD, Ward SG. Targeting PI3K isoforms and SHIP in the immune system: new therapeutics for inflammation and leukemia. Curr Opin Pharmacol. 2012; 12: 444–51. doi: 10.1016/j.coph.2012.02.015 PMID: 22483603

65. Chen Y, Xiao Y, Ge W, Zhou K, Wen J, Yan W, et al. miR-200b inhibits TGF-β1-induced epithelial-mesenchymal transition and promotes growth of intestinal epithelial cells. Cell Death Dis. 2013; 4: e541. doi: 10.1038/cddis.2013.22 PMID: 23492772

66. Ying L, Marino J, Hussain SP, Khan MA, You S, Hofseth AB, et al. Chronic inflammation promotes retinoblastoma protein hyperphosphorylation and E2F1 activation. Cancer Res. 2005; 65: 9132–6. doi: 10.1158/0008-5472.CAN-05-1358 PMID: 16230367

67. Frey MR, Brent Polk D. ErbB receptors and their growth factor ligands in pediatric intestinal inflammation. Pediatr Res. 2014; 75: 127–32. doi: 10.1038/pr.2013.210 PMID: 24402051

68. Khor TO, Huang M-T, Kwon KH, Chan JY, Reddy BS, Kong A-N. Nrf2-deficient mice have an increased susceptibility to dextran sulfate sodium-induced colitis. Cancer Res. 2006; 66: 11580–4. doi: 10.1158/0008-5472.CAN-06-3562 PMID: 17178849

69. Bian Z, Li L, Cui J, Zhang H, Liu Y, Zhang C-Y, et al. Role of miR-150-targeting c-Myb in colonic epithelial disruption during dextran sulphate sodium-induced murine experimental colitis and human ulcerative colitis. J Pathol. 2011; 225: 544–53. doi: 10.1002/path.2907 PMID: 21590770

70. Balzola F, Bernstein C, Ho GT, Russell RK. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease: Commentary. 2011/10/11 ed. Inflammatory Bowel Disease Monitor. 2012. pp. 126–127.

71. Gerlach K, Hwang Y, Nikolaev A, Atreya R, Dornhoff H, Steiner S, et al. TH9 cells that express the transcription factor PU.1 drive T cell-mediated colitis via IL-9 receptor signaling in intestinal epithelial cells. Nat Immunol. 2014; 15: 676–86. doi: 10.1038/ni.2920 PMID: 24908389

72. Marcil V, Sinnett D, Seidman E, Boudreau F, Gendron F-P, Beaulieu J-F, et al. Association between genetic variants in the HNF4A gene and childhood-onset Crohn's disease. Genes Immun. Nature Publishing Group; 2012; 13: 556–65. doi: 10.1038/gene.2012.37 PMID: 22914433

73. Diegelmann J, Czamara D, Le Bras E, Zimmermann E, Olszak T, Bedynek A, et al. Intestinal DMBT1 expression is modulated by Crohn's disease-associated IL23R variants and by a DMBT1 variant which influences binding of the transcription factors CREB1 and ATF-2. PLoS One. 2013; 8: e77773. doi: 10.1371/journal.pone.0077773 PMID: 24223725

74. Barrett JC, Lee JC, Lees CW, Prescott NJ, Anderson CA, Phillips A, et al. Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. Nat Genet. 2009; 41: 1330–4. doi: 10.1038/ng.483 PMID: 19915572

75. Dahan S, Rabinowitz KM, Martin AP, Berin MC, Unkeless JC, Mayer L. Notch-1 signaling regulates intestinal epithelial barrier function, through interaction with CD4+ T cells, in mice and humans. Gastroenterology. 2011; 140: 550–9. doi: 10.1053/j.gastro.2010.10.057 PMID: 21056041

76. Sipos F, Galamb O, Herszényi L, Molnár B, Solymosi N, Zágoni T, et al. Elevated insulin-like growth factor 1 receptor, hepatocyte growth factor receptor and telomerase protein expression in mild ulcerative colitis. Scand J Gastroenterol. 2008; 43: 289–98. Available: http://www.ncbi.nlm.nih.gov/pubmed/18938767 PMID: 18938767

77. Andoh A, Saotome T, Sato H, Tsujikawa T, Araki Y, Fujiyama Y, et al. Epithelial expression of caveolin-2, but not caveolin-1, is enhanced in the inflamed mucosa of patients with ulcerative colitis. Inflamm Bowel Dis. 2001; 7: 210–4. Available: http://www.ncbi.nlm.nih.gov/pubmed/11515846 PMID: 11515846

78. Bouzid D, Fourati H, Amouri A, Marques I, Abida O, Haddouk S, et al. Association of the RAVER2 gene with increased susceptibility for ulcerative colitis. Hum Immunol. 2012; 73: 732–5. doi: 10.1016/j.humimm.2012.04.018 PMID: 22561236

79. Bhonde MR, Gupte RD, Dadarkar SD, Jadhav MG, Tannu AA, Bhatt P, et al. A novel mTOR inhibitor is efficacious in a murine model of colitis. Am J Physiol Gastrointest Liver Physiol. 2008; 295: G1237–45. doi: 10.1152/ajpgi.90537.2008 PMID: 18927209

80. Sobajima J, Ozaki S, Osakada F, Uesugi H, Shirakawa H, Yoshida M, et al. Novel autoantigens of perinuclear anti-neutrophil cytoplasmic antibodies (P-ANCA) in ulcerative colitis: non-histone chromosomal proteins, HMG1 and HMG2. Clin Exp Immunol. 1997; 107: 135–40. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1904558&tool=pmcentrez&rendertype=abstract PMID: 9010268

81. Schulmann K, Mori Y, Croog V, Yin J, Olaru A, Sterian A, et al. Molecular phenotype of inflammatory bowel disease-associated neoplasms with microsatellite instability. Gastroenterology. 2005; 129: 74–85. Available: http://www.ncbi.nlm.nih.gov/pubmed/16012936 PMID: 16012936

82. Deng H, Runger G. Gene selection with guided regularized random forest. 2012; Available: http://arxiv.org/abs/1209.6425

83. Liu JZ, Anderson C. Genetic studies of Crohn's disease: past, present and future. Best Pract Res Clin Gastroenterol. Elsevier Ltd; 2014; 28: 373–86. doi: 10.1016/j.bpg.2014.04.009 PMID: 24913378
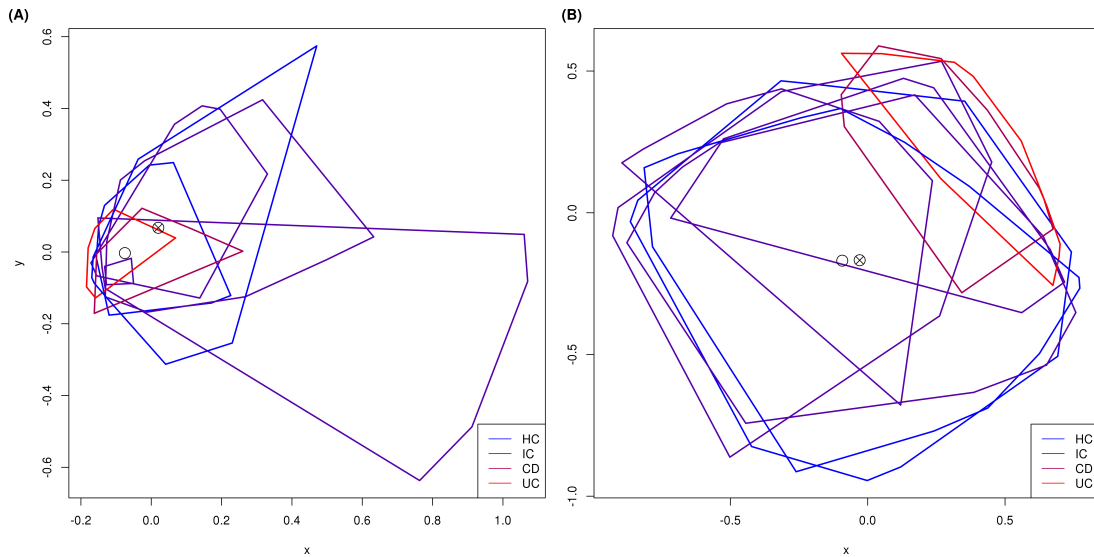
Supplementary Figure S2.1: MDS (multidimensional scaling) plots for visualization of background-subtracted intensity values. Background-subtracted intensity values normalized using variance stabilization (A) before and (B) after median centering based on the batches observed for healthy controls. The corresponding medians are indicated by black circles. MDS was performed using a distance function based on Spearman's rank correlation coefficient. Data points of each group are represented by their α-shape (generalized convex hull). The second plot visualizes the batch-corrected normalized data used for diagnostic classification.
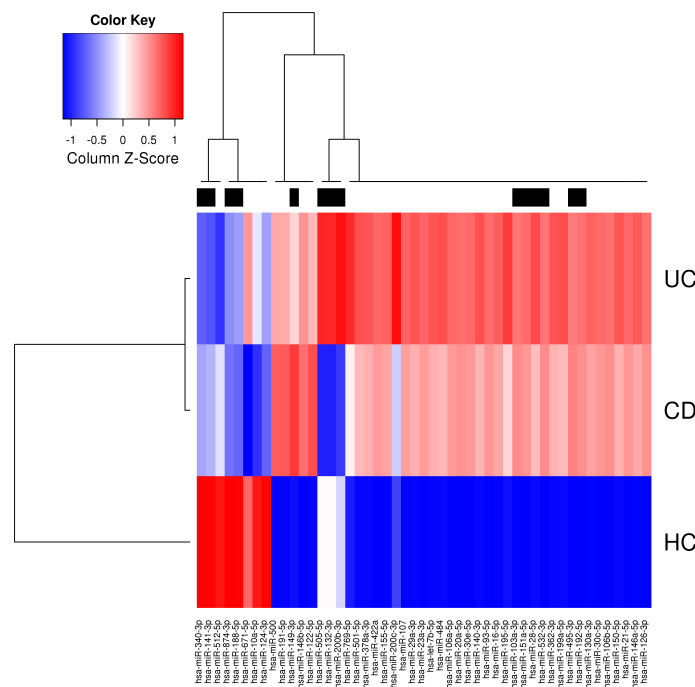
Supplementary Figure S2.2: Median expression levels of miRNAs previously published as being deregulated in CD, UC and HC. The horizontal side bar indicates the correspondence between the literature and the dataset used for this study. Measurements with directions of effect deviating from the literature are marked using black bars. The heat map was generated using a distance function based on Spearman's rank correlation coefficient and agglomerative hierarchical clustering using complete-linkage. Low and high expression levels are plotted using red and blue, respectively.
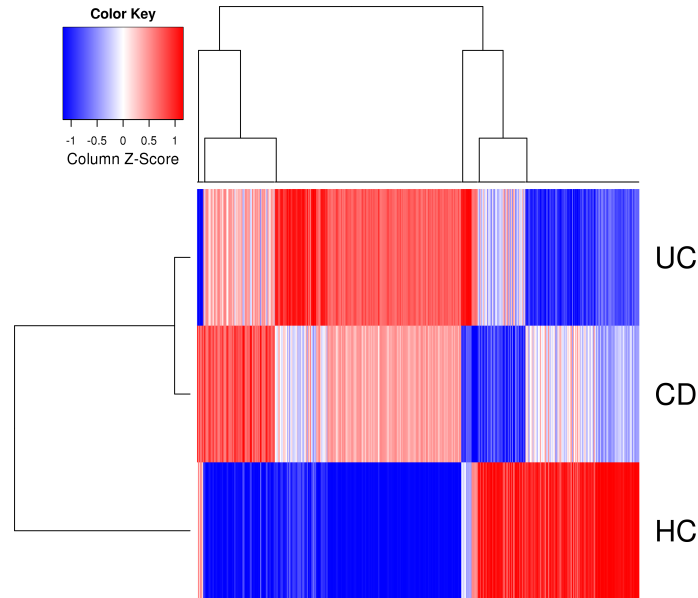
Supplementary Figure S2.3: Median expression profiles of significantly deregulated miRNAs in CD, UC and HC. For each pair of groups two-sample t-tests were applied. Deregulation was considered as being significant for Holm-corrected p-values < 0.05. Not significantly differentially expressed miRNAs were neglected. 667 out of 863 miRNAs were differentially deregulated in any of the comparisons.
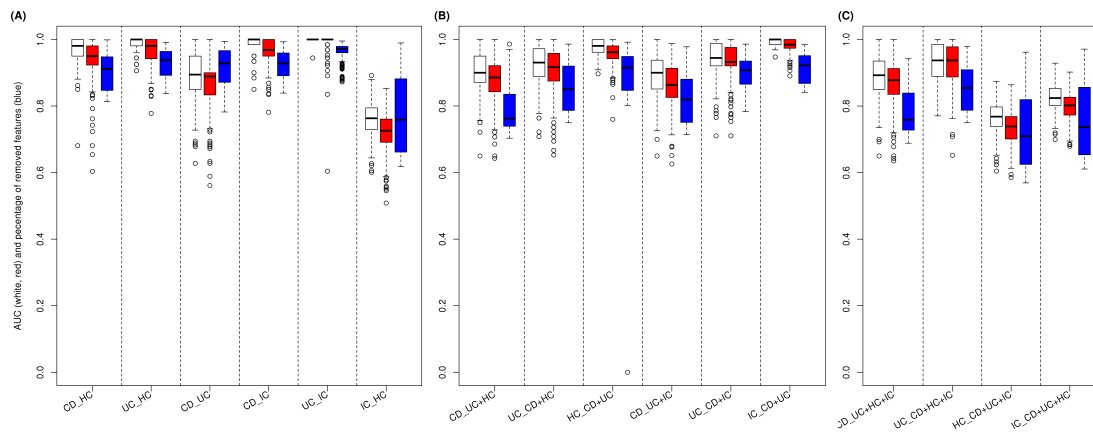
Supplementary Figure S2.4: Classification results for LASSO SVM. Measured by the area under the ROC curve (AUC) classification performance is shown for models considering (A) 2 groups (CD vs. HC, UC vs. HC, CD vs. UC, CD vs. IC, UC vs. IC, IC vs. HC), (B) 3 groups (CD vs. UC+HC, UC vs. CD+HC, HC vs. CD+UC, CD vs. UC+IC, UC vs. CD+IC, IC vs. CD+UC) and (C) 4 groups (CD vs. UC+HC+IC, UC vs. CD+HC+IC, HC vs. CD+UC+IC, IC vs. CD+UC+HC). Performance of linear standard SVMs (considering every miRNA measured, white boxes) is compared to linear LASSO SVMs (considering subsets of miRNAs measured, red boxes). In addition, as a measure of model complexity the percentage of miRNAs neglected for constructing the respective penalized SVMs are plotted (blue boxes).

Supplementary Figure S2.5: Classification results for elastic net SVM. Measured by the area under the ROC curve (AUC) classification performance is shown for models considering (A) 2 groups (CD vs. HC, UC vs. HC, CD vs. UC, CD vs. IC, UC vs. IC, IC vs. HC), (B) 3 groups (CD vs. UC+HC, UC vs. CD+HC, HC vs. CD+UC, CD vs. UC+IC, UC vs. CD+IC, IC vs. CD+UC) and (C) 4 groups (CD vs. UC+HC+IC, UC vs. CD+HC+IC, HC vs. CD+UC+IC, IC vs. CD+UC+HC). Performance of linear standard SVMs (considering every miRNA measured, white boxes) is compared to linear elastic net SVMs (considering subsets of miRNAs measured, red boxes). In addition, as a measure of model complexity the percentage of miRNAs neglected for constructing the respective penalized SVMs are plotted (blue boxes).
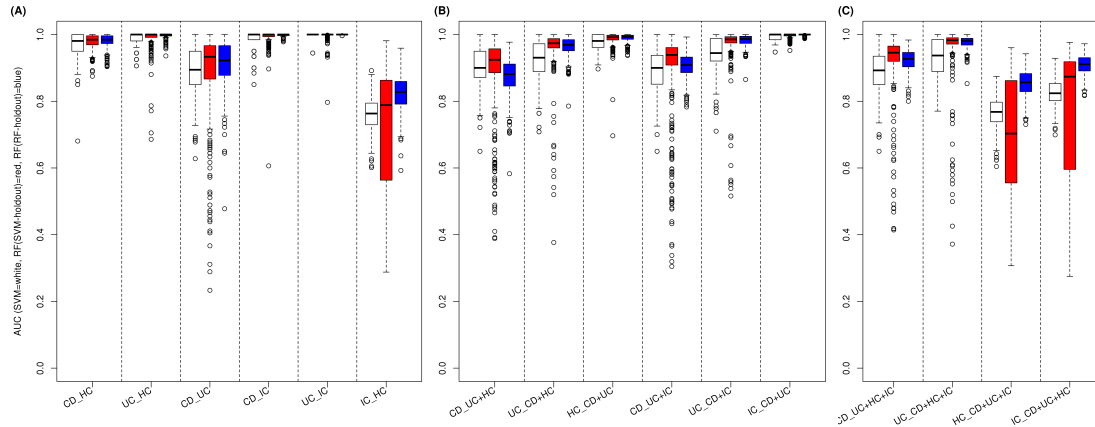

Supplementary Figure S2.6: Classification results for SCAD SVM. Measured by the area under the ROC curve (AUC) classification performance is shown for models considering (A) 2 groups (CD vs. HC, UC vs. HC, CD vs. UC, CD vs. IC, UC vs. IC, IC vs. HC), (B) 3 groups (CD vs. UC+HC, UC vs. CD+HC, HC vs. CD+UC, CD vs. UC+IC, UC vs. CD+IC, IC vs. CD+UC) and (C) 4 groups (CD vs. UC+HC+IC, UC vs. CD+HC+IC, HC vs. CD+UC+IC, IC vs. CD+UC+HC). Performance of linear standard SVMs (considering every miRNA measured, white boxes) is compared to linear SCAD SVMs (considering subsets of miRNAs measured, red boxes). In addition, as a measure of model complexity the percentage of miRNAs neglected for constructing the respective penalized SVMs are plotted (blue boxes).
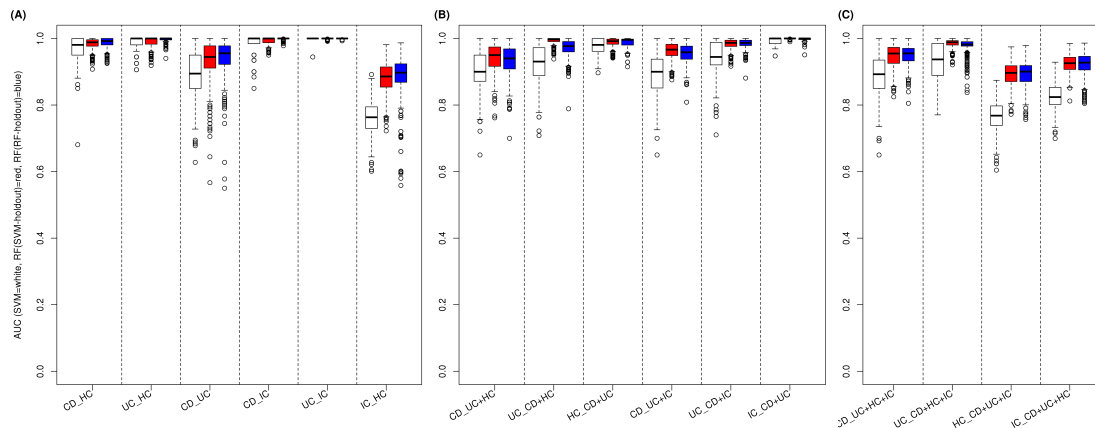
Supplementary Figure S2.7: Comparison of SVM and random forest. Measured by the area under the ROC curve (AUC) classification performance is shown for models considering (A) 2 groups (CD vs. HC, UC vs. HC, CD vs. UC, CD vs. IC, UC vs. IC, IC vs. HC), (B) 3 groups (CD vs. UC+HC, UC vs. CD+HC, HC vs. CD+UC, CD vs. UC+IC, UC vs. CD+IC, IC vs. CD+UC) and (C) 4 groups (CD vs. UC+HC+IC, UC vs. CD+HC+IC, HC vs. CD+UC+IC, IC vs. CD+UC+HC). Classification performance of the linear elastic SCAD SVM (white box) is compared to a Random forests per holdout sample considering variables selected using the SVM (red box) and the Random forest itself (blue box), respectively.

Supplementary Figure S2.8: Comparison of SVM and random forest. Measured by the area under the ROC curve (AUC) classification performance is shown for models considering (A) 2 groups (CD vs. HC, UC vs. HC, CD vs. UC, CD vs. IC, UC vs. IC, IC vs. HC), (B) 3 groups (CD vs. UC+HC, UC vs. CD+HC, HC vs. CD+UC, CD vs. UC+IC, UC vs. CD+IC, IC vs. CD+UC) and (C) 4 groups (CD vs. UC+HC+IC, UC vs. CD+HC+IC, HC vs. CD+UC+IC, IC vs. CD+UC+HC). Classification performance of the linear elastic SCAD SVM (white box) is compared to a Random forests considering variables selected using the median performing SVM (red box). Additionally, Random forests were trained with the top 50% of the variables ranked by their frequency of selection (blue box).

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11)   **EP 3 212 803 B1**

(12)   **EUROPEAN PATENT SPECIFICATION**

(45) Date of publication and mention
of the grant of the patent:
26.12.2018 Bulletin 2018/52

(51) Int Cl.:
C12Q 1/68 (2018.01)

(21) Application number: **15741191.9**

(86) International application number:
PCT/EP2015/066832

(22) Date of filing: **23.07.2015**

(87) International publication number:
WO 2016/066288 (06.05.2016 Gazette 2016/18)

(54) **MIRNAS AS NON-INVASIVE BIOMARKERS FOR INFLAMMATORY BOWEL DISEASE**

MIRNAS ALS NICHTINVASIVE BIOMARKER FÜR ENTZÜNDLICHE DARMERKRANKUNG

MIARN UTILISÉS COMME BIOMARQUEURS NON INVASIFS DE LA MALADIE INFLAMMATOIRE DE L'INTESTIN

(84) Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR

(30) Priority: **28.10.2014 EP 14190596**
02.03.2015 EP 15157101

(43) Date of publication of application:
06.09.2017 Bulletin 2017/36

(73) Proprietor: **Hummingbird Diagnostics GmbH**
69120 Heidelberg (DE)

(72) Inventors:
• **BREFORT, Thomas**
85560 Ebersberg (DE)
• **FRANKE, Andre**
24119 Kronshagen (DE)
• **HEMMRICH-STANISAK, Georg**
24783 Osterrönfeld (DE)
• **HÜBENTHAL, Matthias**
24116 Kiel (DE)

• **SCHEFFLER, Matthias**
69493 Hirschberg (DE)

(74) Representative: **Geling, Andrea**
ZSP Patentanwälte PartG mbB
Hansastraße 32
80686 München (DE)

(56) References cited:
WO-A1-2013/043482   WO-A2-2009/120877

• **FENG WU ET AL: "Peripheral blood MicroRNAs distinguish active ulcerative colitis and Crohn's disease", INFLAMMATORY BOWEL DISEASES, vol. 17, no. 1, 1 January 2011 (2011-01-01), pages 241-250, XP55046385, ISSN: 1078-0998, DOI: 10.1002/ibd.21450**
• **PARASKEVI ARCHANIOTI ET AL: "Micro-RNAs as regulators and possible diagnostic bio-markers in inflammatory bowel disease", JOURNAL OF CROHN'S AND COLITIS, ELSEVIER BV, NL, vol. 5, no. 6, 21 May 2011 (2011-05-21), pages 520-524, XP028116598, ISSN: 1873-9946, DOI: 10.1016/J.CROHNS.2011.05.007 [retrieved on 2011-06-01]**

EP 3 212 803 B1

The full version of the patent application is available online at:
`https://register.epo.org/application?number=EP15741191`

US 20170306407A1

(19) **United States**

(12) **Patent Application Publication**   (10) Pub. No.: **US 2017/0306407 A1**
Brefort et al.   (43) **Pub. Date: Oct. 26, 2017**

(54) **MIRNAS AS NON-INVASIVE BIOMARKERS FOR INFLAMMATORY BOWEL DISEASE**

(71) Applicant: **Hummingbid Diagnostics GmbH,** Heidelberg (DE)

(72) Inventors: **Thomas Brefort**, Walldorf (DE); **Andre Franke**, Kronshagen (DE); **Georg Hemmrich-Stanisak**, Osterronfeld (DE); **Matthias Hubenthal**, Kiel (DE); **Matthias Scheffler**, Hirschberg (DE)

(21) Appl. No.:   **15/520,751**

(22) PCT Filed: **Jul. 23, 2015**

(86) PCT No.:   **PCT/EP2015/066832**

§ 371 (c)(1),
(2) Date: **Apr. 20, 2017**

(30)   **Foreign Application Priority Data**

Oct. 28, 2014   (EP) 14190596.8
Mar. 2, 2015   (EP) 15157101.5

Publication Classification

(51) **Int. Cl.**
C12Q 1/68   (2006.01)
(52) **U.S. Cl.**
CPC *C12Q 1/6883* (2013.01); *C12Q 2600/158* (2013.01); *C12Q 2600/178* (2013.01)

(57)   **ABSTRACT**

The present invention relates to a method for diagnosis of Inflammatory Bowel Disease (IBD) based on the determination of expression profiles of miRNAs representative for diagnosis of IBD compared to a reference. In addition, the present invention relates to a kit for diagnosis of IBD comprising means for determining expression profiles of miRNAs representative for IBD.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
6 May 2016 (06.05.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/066288 A1

(51) **International Patent Classification:**
C12Q 1/68 (2006.01)

(21) **International Application Number:**
PCT/EP2015/066832

(22) **International Filing Date:**
23 July 2015 (23.07.2015)

(25) **Filing Language:** English

(26) **Publication Language:** English

(30) **Priority Data:**
14190596.8   28 October 2014 (28.10.2014)   EP
15157101.5   2 March 2015 (02.03.2015)   EP

(71) **Applicant:** COMPREHENSIVE BIOMARKER CEN-TER GMBH [DE/DE]; Im Neuenheimer Feld 583, 69120 Heidelberg (DE).

(72) **Inventors:** BREFORT, Thomas; Dannheckerstr. 19, 69190 Walldorf (DE). FRANKE, Andre; Hindenburgstr. 2, 24119 Kronshagen (DE). HEMMRICH-STANISAK, Georg; Königsbergerstr. 14, 24783 Osterrönfeld (DE). HÜBENTHAL, Matthias; Kronshagener Weg 38, 24116 Kiel (DE). SCHEFFLER, Matthias; Steig 16B, 69493 Hirschberg (DE).

(74) **Agent:** GELING, Andrea; Zwicker Schnappauf & Partner Patentanwälte PartG mbB, Radlkoferstr. 2, 81373 München (DE).

(81) **Designated States** *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report (Art. 21(3))*
— *with sequence listing part of description (Rule 5.2(a))*

WO 2016/066288 A1

(54) **Title:** MIRNAS AS NON-INVASIVE BIOMARKERS FOR INFLAMMATORY BOWEL DISEASE

(57) **Abstract:** The present invention relates to a method for diagnosis of Inflammatory Bowel Disease (IBD) based on the determination of expression profiles of miRNAs representative for diagnosis of IBD compared to a reference. In addition, the present invention relates to a kit for diagnosis of IBD comprising means for determining expression profiles of miRNAs representative for IBD.

The full version of the patent application is available online at:
https://worldwide.espacenet.com/publicationDetails/biblio?CC=WO&NR=2016066288A1&KC=A1

Comprehensive discriminative analysis of sequencing-based isomiR expression profiles reveals highly accurate tools for diagnostics of inflammatory bowel disease

Matthias Hübenthal[1], Simonas Juzėnas[1], Susanna Nikolaus[1], Dominik Schulte[2], Sebastian Zeißig[2], Nina Strüning[2], Jonas Halfvarson[3], Mauro D'Amato[4,5,6,7], Andreas Keller[5], Limas Kupčinskas[9,10], Stefan Schreiber[1,2], Georg Hemmrich-Stanisak[1], Andre Franke[1]

[1]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany, [2]Department of Internal Medicine I, University Hospital Schleswig-Holstein, Kiel, Germany, [3]Department of Gastroenterology, Faculty of Medicine and Health, Örebro University, Örebro, Sweden, [4]Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden, [5]Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden, [6]Department of Gastrointestinal and Liver Diseases, Biodonostia Health Research Institute, San Sebastián, Spain, [7]IKERBASQUE, Basque Science Foundation, Bilbao, Spain, [8]Chair for Clinical Bioinformatics, Saarland University, Saarbrücken, Germany, [9]Institute for Digestive Research, Academy of Medicine, Lithuanian University of Health Sciences, Kaunas, Lithuania, [10]Department of Gastroenterology, Academy of Medicine, Lithuanian University of Health Sciences, Kaunas, Lithuania

Abstract

Inflammatory bowel disease (IBD) is a chronic relapsing disorder of the alimentary tract, encompassing two clinical entities, namely Crohn's disease (CD) and ulcerative colitis (UC). In the course of the last decades IBD has been recognized as an increasingly common diagnosis, rendering a significant burden to societies and healthcare systems around the world. The diagnostic procedure for both IBD subtypes involves a combination of clinical, endoscopic, histological and radiological parameters. Numerous biomarkers have been proposed to complement IBD diagnosis. However, due to insufficient accuracy none of them can be recommended for clinical diagnostic practice. With this study we present the first-time evaluation of the clinical utility of miRNA variants (isomiRs) in the context of IBD.

Global profiling of isomiRs expression has been conducted by next generation sequencing based on whole-blood samples drawn from a cohort of 515 individuals (346 cases of untreated/treated IBD as well as 124 healthy and 45 symptomatic controls). The resulting profiles have been used to identify trait-specific signatures of isomiR expression, providing the foundation to derive binomial as well as multinomial classification models for IBD-related diagnostic problems. For pairs of untreated phenotypes (CD−, UC−, including HC) we estimated these models to perform with a median BAC of 0.9286. Models distinguishing pairs of treated phenotypes (CD+, UC+, including HC), in turn, we estimated to perform with a median BAC of 0.9655. According to the principal of majority voting, binomial models have been used to construct combined classifiers, solving 3-nomial (CD+ vs. UC+ vs. HC, CD− vs. UC− vs. HC and CD− vs. UC− vs. SC) as well as 4-nomial diagnostic problems (CD− vs. UC− vs. SC vs. HC). With mean BACs of up to 0.9000 and 0.8667, respectively, the performance of the models likewise has been estimated to be remarkably high.

Overall, we do not only provide evidence for sequencing-based isomiR expression profiles being a valuable tools for diagnosing IBD but also for their superiority over established biomarkers. However, the translatability of these findings into routine clinical practice remains to be proven.

1

1   Introduction

Inflammatory bowel disease (IBD) is a chronic relapsing disorder of the alimentary tract. It encompasses two major subtypes, namely Crohn's disease (CD) and ulcerative colitis (UC), each showing extensive heterogeneity in terms of clinical and histological presentation as well as response to treatment. A growing body of evidence suggests the disease being a results of dysregulated inflammatory response to commensal microbes in a genetically susceptible host. In a series of genome-wide association studies (GWAS, including [11, 25]) 241 genetic loci have been identified to be associated with IBD. While these GWAS findings added tremendously to the understanding of the genetic architecture of the disease, the exact mechanisms of its pathogenesis is barely understood and linking SNP associations to functional mechanisms is a major challenge to the field. Consequently, the different aspects of gene regulation also have been investigated in IBD.

microRNAs (miRNAs) represent a class of short endogenous, non-coding ribonucleic acids (RNAs), acting as master post-transcriptional regulators of a wide range of physiological and pathological processes. Thus, the molecules modulate the activity of specific messenger RNAs (mRNAs) by predominantly targeting their 3' UTR region [2]. As evident from a growing number of studies each single miRNA hairpin arm (3' and 5', respectively) can give rise to numerous variants, referred to as isomiRs, which differ in length and sequence composition [16, 36, 49]. These variants can be classified into (templated or non-templated) 3' and 5' isomiRs, respectively, as well as polymorphic isomiRs. Their abundance is shown to vary depending on gender, ethnicity [29], tissue, cell type [21] and state of health [49]. It has been shown that 5' isomiRs extend target repertoire of their corresponding reference miRNAs [7], whereas 3' isomiRs seem to be involved in modulation of miRNA targeting effectiveness [5] and even in subcellular and extracellular miRNA trafficking [23, 53]. However, their biological significance still needs to be investigated.

Due to their short length miRNAs are considered to be more stable than mRNAs [22], even in tissues and body fluids, such as peripheral blood [35]. Furthermore, they are shown to be deregulated in a variety of conditions, including cancer [6, 31] and autoimmune disorders [1, 13, 46], and signatures of differentially expressed miRNAs have been intensively studied as biomarkers for the prediction of diseases. Noteworthy, in the context of IBD the number of blood-based studies of differential miRNA expression is limited. Thus, for instance, Paraskevi et al. [39] as well as Schaefer et al. [45] reported distinct expression patterns for a small number of candidate miRNAs comparing CD, UC with the healthy state (HC) based on real-time quantitative PCR (RT-qPCR). To-date, the only miRNome-wide analyses of IBD have been performed by Wu et al. [52] and Hübenthal et al. [17]. Employing microarray technology these studies revealed putative biomarkers (along with mathematical models) for the distinction of CD, UC and HC.

Relying on predefined annealing probes, microarray technology solely allows the detection of annotated miRNAs and isomiRs, respectively. Next generation sequencing (NGS) technology, instead, enables the quantification of the entire set of isomiRs being expressed in a given sample. Recent studies of colonic/ileal miRNAs/isomiRs employed NGS for the distinction between CD, UC and healthy controls [4, 27, 28] or between CD subtypes among each other [40]. However, global NGS-based profiles of miRNAs/isomiRs circulating in whole blood so far have not been investigated in the context of IBD.

This study now examines isomiR expression profiles of 515 individuals (346 cases of IBD as well as 124 healthy and 45 symptomatic controls) and state-of-the-art machine learning methods to identify trait-specific signatures. Therewith, we provide the to-date most comprehensive discriminative analysis of sequencing-based isomiR expression profiles from whole blood, contributing to a more comprehensive understanding of inflammatory bowel disease. Ultimately, we present classification models for highly accurate sequencing-based isomiR-guided diagnostics of IBD and a multitude of other diagnostically relevant problems. These models serve as exemplary applications of signatures of differentially expressed isomiRs and represent the first-time evaluation of the discriminative power of these molecules in the context of IBD. Our data clearly shows that isomiR expressions profiles are

2

a powerful tool for the distinction of CD and UC in both, treated and untreated patients. However, the signature enabling the distinction varies depending on the treatment status.

2 Materials and methods

2.1 Patients and samples

For study participants recruited in Germany and Sweden biological samples as well as clinical meta-data was collected. The study has been approved by the respective local ethics committees (PopGen 2.0 Netzwerk (P2N) and ethics committee of the medical faculty, University Hospital Schleswig-Holstein, Kiel, Germany; ethics committee of the Karolinska Institutet, Stockholm, Sweden). All participants provided written informed consent.

The German cohort comprises 94 healthy individuals as well as 271 patients diagnosed with inflammatory conditions. The diagnoses include Crohn's disease (treated CD denoted as CD+, n=146) and ulcerative colitis (treated UC denoted as UC+, n=125). The Swedish cohort (Swedish Inception Cohort in IBD, SIC IBD) in turn, includes 30 healthy individuals as well as 120 patients diagnosed with inflammatory conditions. The diagnoses comprise Crohn's disease (untreated CD denoted as CD−, n=27) and ulcerative colitis (untreated UC denoted as UC−, n=48). Patients with gastro-intestinal symptoms not being diagnosed with IBD have been considered as symptomatic controls (SC, n=45). For a summary reference is made to Table S1.

A major difference between the German and Swedish cohort is their treatment status. Swedish samples were included in the SIC IBD cohort at first diagnosis of IBD. This implies their naïvety with regard to any IBD medication. More precisely, criteria for excluding patients were systemic treatment with immunosuppressants, such as Infliximab (Remicade, Remsima, Inflectra), Adalimumab (Humira, Exemptia), Methotrexat (Trexall, Rheumatrex), Azathioprine (Azasan, Imuran), Tacrolimus (Prograf, Advagraf) and Ciclosporin (Neoral, Sandimmune). Furthermore, topical treatment (by suppository or enema) with Mesalazine (Azacol, Lialda), systemic treatment with corticosteroids, such as budesonide (Entocort) as well as with sulfasalazine (Azulfidine, Salazopyrin), 6-Mercaptopurine (Purinethol), Vedolizumab (Entyvio) or Certolizumab (Cimzia), respectively, lead to exclusion of participants. In contrast German samples are assumed to be treated with one or more of the drugs mentioned before. For further details see Table 1.

CD and UC cases of both cohorts have been evaluated with regard to disease activity the Harvey-Bradshaw index (HBI), the clinical activity index according to Rachmilewitz (CAI) as well as the partial Mayo score, respectively [8, 43, 44]. In addition to medication and disease activity, patients have been evaluated with regard to common clinical parameters, such as gender and smoking status. Furthermore, information on disease location, behavior, extent (Montreal classification) as well as serological (C-reactive protein, CRP; albumin) and fecal markers (calprotectin) has been collected. For details including parameter's distributional information reference is made to Table 1.

2.2 RNA isolation and sequencing

Stabilization and purification of total RNA (including miRNA) from whole blood samples was conducted employing the PAXgene Blood miRNA System (Qiagen) in accordance to the manufacturers' instructions. The products were subjected to Illumina's TruSeq Small RNA Sample Preparation protocol to generate small RNA libraries for each sample. Subsequently, sequencing was conducted on Illumina HiSeq 2500 machines (24 samples per lane, single-end reads of 50bp length, TruSeq sequencing chemistry v2.5, 3.0 or 4.0).

Raw sequencing reads as well as quantified read-count data will be deposited at NCBI Gene Expression Omnibus (GEO) [10].

3

2.3 Processing of small RNA-seq data

Raw small RNA sequencing reads were processed using cutadapt v1.3 [33] to remove adapters, low-quality bases (<Q20) and reads shorter than 18bp. Processed reads which mapped to viral genomes, viral miRNA precursors and human non-miRNA short RNAs were filtered out. The remaining reads were mapped to miRNA sequences from miRBase release 22 [24] using mirAligner [37] with default parameters. The R package isomiRs [38] then was used to generate read count data. Thereby for a given reference miRNA mismatches observed in $< 20\%$ of the reads were considered as sequencing errors. The corresponding reads have been removed accordingly. Furthermore requiring the sum of counts to be $\geq 10^6$ for each sample and the number of samples to be ≥ 10 for each technical batch under consideration we obtained count matrices of 14874 isomiRs for the German (comprising 365 samples) as well well as the Swedish cohort (comprising 150 samples). By considering only reads with counts exceeding 5 in a fraction of samples of $>10\%$ per batch, sex and trait the number of isomiRs reduced to 2895 for the German and 1954 for the Swedish cohort. We only incorporate isomiRs featuring canonical nucleotide substitutions (cytidine (C) to uridine (U) and (A) to inosine (I) deamination) as well as canonical nucleotide additions (poly(A) and poly(T) tailing), further reducing dimensionality to 2426/1565 for the German/Swedish cohort. Besides 282/213 reference miRNAs, the final datasets comprise 1739/1100 3' trimming modifications, 723/405 5' trimming modifications, 737/460 3' end additions as well as 8/6 nucleotide substitutions (including 3/2 seed alterations).

This dataset was normalized by performing variance stabilizing transformation (VST) based on parametric dispersion estimates as implemented in the R package DESeq2 [30]. Samples for which normalized counts x met $x \gtreqless |\text{median}(x) \pm 3 \cdot \text{ICQ}(x)|$ (with respect to batch and trait) were considered as being outliers and removed. Subsequently, the same criterion was applied to exclude outliers with regard to relative log expression (RLE). After renormalization adjustment for biological (gender) as well as experimental batches (sequencing technician, chemistry and run) has been performed, employing the empirical Bayesian framework implemented by the R function sva::ComBat [20, 26]. Explorative plots then were generated based on a multidimensional scaling (MDS) of the data employing Spearman dissimilarities.

2.4 Differential expression analysis

Subsequent to the preprocessing of the isomiR read count data differential expression analysis has been performed to facilitate the identification of trait-specific isomiR expression patterns. The analysis is based on generalized linear models (GLMs) allowing for pairwise comparisons of the traits of interest while adjusting for unwanted biological/experimental variability. Significance of isomiR deregulation thereby corresponds to non-negligibility of log fold changes ($\log_2 \text{FC}$) and is assessed by applying Wald tests to the coefficients of the GLM. For the Swedish dataset we conducted 6 pairwise comparisons, incorporating untreated IBD (CD− and UC−) as well as controls (SC and HC). For the German dataset, in turn, we conducted 4 pairwise comparisons, incorporating treated IBD (CD+ and UC+) as well as controls (HC). Direct comparison between subsets of the Swedish and German cohorts have not been performed. However, by employing HC as a mutual reference, exclusive features of isomiR expression profiles (for CD−, UC−, CD+, UC+, SC) can be identified post hoc and regardless of the country of origin. To illustrate the resulting expression signatures, heatmaps and quasi-proportional Venn diagrams [41] have been generated.

2.5 Literature-based validation

There is an growing body of evidence for various miRNAs being deregulated in IBD (CD vs. HC, UC vs. HC as well as CD vs. UC). However, the vast majority of studies used samples different from whole blood. We employed rentrez::entrez_search [50] to gather information available to date (for the corresponding PubMed query please see Figure S2). Subsequent manual inspection of the retrieved papers resulted in a list of 25 recent articles reporting

4

differentially expressed miRNAs in whole blood. The isomiRs being differentially expressed (Wald $p < .05$ and $\log_2 |\mathrm{FC}| > 0$) in the study presented here have been grouped with regard to the respective miRNA arms in order to allow a comparison to these earlier studies not investigating miRNA isoforms. A previously published set of deregulated miRNAs was assumed to support our findings in case it was overlapping with them. We performed enrichment analysis employing Fisher's exact test to evaluate these overlaps and by this the significance of our findings. Note that, given the precise definition of the traits as well as composition and size of the respective study populations varied, this approach might have a limited power, so the sizes of the overlaps are likely to be underestimated.

2.6 Mathematical modeling

With this study we aim for the establishment of state-of-the-art machine learning methods for the mathematical modeling of differential expression of isomiRs between the diagnoses under consideration, and to apply the resulting models for in-silico predictions of these traits. In total, mathematical models for 9 distinct clinical classification problems have been considered. This corresponds to all $(4 \cdot 3/2) = 6$ binary comparisons incorporating the 4 Swedish (CD−, UC−, SC and HC) as well as all $(3 \cdot 2/2) = 3$ binary comparisons incorporating the 3 German phenotypes (CD+, UC+ and HC). It is worth noting that the considered classification problems vary with respect to the importance for clinical practice. However, to answer more complex diagnostic questions by subsequent construction of combined classification models, none of these can be neglected.

Each of the classification problems mentioned before has been solved by employing classical types of support vector machines (SVMs), namely linear and radial SVMs as well as polynomial SVMs (R package e1071 [34]). In addition, regularized incarnations of the SVM, including elastic net and SCAD (smoothly clipped absolute deviation) SVM have been employed to generate distinguishing models of higher sparsity (R package penalizedSVM [3]). For technical details on these models reference is made to Hübenthal et al. [17]. In order to increase stability and sparsity of the models, only isomiRs exhibiting differential expression ($p < 0.05$ and $|\log_2 \mathrm{FC}| > 0$) with regard to the respective classification problem have been included in the modeling process. Each single model has been evaluated using common metrics for the assessment of binary classifiers. Thus, for instance true positives and negatives (TP and TN) as well as false positives and negatives (FP and FN), along with positive and negative predictive value (PPV and NPV) as well as sensitivity (SN), specificity (SP), balanced accuracy (BAC) and Matthews correlation coefficient (MCC) have been calculated. Estimation of model parameters and model performance has been performed based on randomly sampled subpopulations. Repetitions of the sampling have been performed to determine distributional properties of each model's performance. Optimization of the model's hyperparameters for each of these samples has been conducted based on nested cross validation (CV) and grid search as implemented in R functions e1071::tune.svm and penalizedSVM::svm.fs. Models performing with highest classification accuracy and stability (as estimated by median and interquartile range (IQR) of BAC) then have been chosen for further investigation. According to the principal of majority voting [14], these models have been used to construct combined classifiers that solve more complex diagnostic problems, such as the distinction between 3 (CD− vs. UC− vs. HC, CD− vs. UC− vs. SC, CD+ vs. UC+ vs. HC) or even 4 traits (CD− vs. UC− vs. SC vs. HC). Combined classifiers' performances have been estimated employing leave-on-out cross validation (LOO-CV).

Throughout the analysis the available data has been used based on the premise of good generalizability. Construction of the atomic classification models has been conducted on a dataset of 500 samples, remaining after random removal of 35 samples. The process of model construction thereby involved training (n=310) and evaluation datasets (n=155) resulting from random splits into 1000 folds at an ratio of 2:1. Overall 45000 computational experiments (model estimation and evaluation for 9 problems, 5 models and 1000 holdout samples) have been performed. They all have been conducted employing the on-site high-performance computing (HPC) resources as one job per holdout and phenotype group. After determination of the final models the 35 samples drawn initially

<div align="center">5</div>

have been used to evaluate the model's prediction resulting from majority voting. Thereby the evaluation scheme of LOO-CV has been followed. For a flowchart of the full setup of the classification experiment reference is made to Figure 4.

3   Results

3.1   Explorative analysis reveals disease- and treatment-specific differences

We used plots based on multidimensional scaling to explore general features of the given isomiR expression data. As evident from Figure 1, data points form homogeneous clusters for each trait under consideration. Judged by visual appearance neither location nor dispersion differ considerably. This observation is further supported by Kruskal-Wallis tests revealing no significant distributional differences for both, the Swedish and the German cohort (data not shown). Likewise, this confirms that the adjustment for unwanted variation has been successful (see Figure S1). Being hardly distinguishable on a global scale, samples of each trait show distinct expression patterns for a subset of isomiRs. Thus, pairwise Wald tests (nominal $p < 0.05$ and $|\log_2 FC| > \log_2 1.5$) of all substrata of the Swedish cohort reveal differential expression for a median number of 30 isomiRs (24 miRNA arms, see Table S3A). The substrata of the German cohort, in turn, reveal differential expression for a median number of 155 isomiRs (85 miRNA arms, see Table S3B). To identify isomiRs being exclusively deregulated in a given trait we examined the relationships between sets of differentially expressed isomiRs using Venn diagrams. However, to allow for a comparison between the cohorts, we only considered pairwise comparisons with respect to HC. In this way we identified 60 isomiRs (34 miRNA arms) as being exclusively deregulated in the untreated samples (CD− vs. HC: 12 (9), UC− vs. HC: 23 (12) and SC vs. HC: 25 isomiRs (13 miRNA arms), see Figure 2A) as well as 86 isomiRs (32 miRNA arms) as being exclusively deregulated in the treated samples (CD+ vs. HC: 38 (16), UC+ vs. HC: 48 isomiRs (16 miRNA arms), see Figure 2B). The signatures of untreated and treated CD comprise 30 and 151 exclusively deregulated isomiRs (21 and 78 miRNA arms, see Figure 2C). The signatures of untreated and treated UC, in turn, comprise 53 and 159 exclusively deregulated isomiRs (29 and 71 miRNA arms, see Figure 2D). Expression patterns (in terms of logarithmized fold changes with respect to HC) of isomiRs being exclusively differentially expressed in all the traits under consideration are depicted in Figure 3. For CD−, UC−, CD+, UC+ and SC the heatmap comprises 11, 21, 34, 45 and 17 trait-specific isomiRs from 7, 8, 11, 11 and 7 trait-specific miRNA arms. Only 11 of these 128 miRNAs correspond to reference miRNAs being annotated in miRBase release 22. Each of the remaining miRNAs exhibits at least one sequence modification (102 3' trimming and 32 5' trimming modifications, along with 1 nucleotide substitution). Notably, not a single 3' addition has been detected as being trait-specific.

3.2   Study adds to the current knowledge gathered from literature

We consider the present study as a pilot assessing the capability of circulating isomiRs predicting inflammatory bowel disease. Furthermore, it is the first study investigating IBD-associated blood-born miRNAs employing next generation sequencing (NGS) technology. Nonetheless, previous publications reported alternative approaches addressing the same question. Enrichment analysis (comparing our study to earlier studies) was performed to assess the significance of our findings in terms of overlapping sets of differentially expressed isomiRs and miRNA arms, respectively.

Several studies based their investigation on whole blood samples. Thus, Paraskevi et al. [39] and Schaefer et al. [45], for instance, employed real-time quantitative PCR (RT-qPCR) to assay deregulation of miRNAs in CD and UC with respect to the healthy state (HC). Enrichment analysis reveals non-random overlaps between each of these studies and ours, for both, the untreated (p=3.66E-02 and 1.57E-03) as well as the treated (p=4.44E-11 and 2.65E-03) IBD cohort. Microarray studies by Wu et al. [52] and Hübenthal et al. [17] focusing on CD vs. HC,

6

UC vs. HC and CD vs. UC confirm our results as well by likewise exhibiting a significant enrichment. Notably, this only holds true for the treated cohort (p=4.17E-05 and 3.06E-03). Further studies by Iborra et al. [19] and Zahm et al. [54, 55] based their investigation on serum-born miRNAs assayed by RT-qPCR. All three studies focus on the comparison of CD and UC with the healthy state and support our results as well. Specifically, for the untreated IBD cohort randomness as a source for overlapping sets of miRNAs has been ruled out with p=1.12E-04, 3.49E-02 and 2.32E-02. The same applies to treated the IBD cohort with p=7.18E-10, 2.63E-08 and 1.17E-02. Table 2 briefly summarizes the enrichment analyses. A more comprehensive overview including identifiers of the overlapping miRNA arms are provided in Table S5. For the underlying raw data resulting from the PubMed query (Figure S2) reference is made to Table S4.

3.3   Mathematical modeling allows highly accurate trait prediction

Specific properties of isomiR expression profiles being detected in the course of explorative analysis have been further used to build mathematical models for in-silico distinction of the traits under investigation. Thus, (penalized) support vector machines have been generated to enable binomial classification of the clinical conditions. Distributional features of each classifiers' performance thereby has been evaluated based on 500-fold holdout sampling. With median balanced accuracy of 0.9182 the overall performance (considering 5 model types and 9 classification problems) has been estimated to be remarkably high. However, with an inter quartile range (IQR) of 0.2049, it varied considerably. This variability illustrates the substantial differences in the difficulty of the classification problems under consideration. Thus, for instance, the distinction between pairs of untreated phenotypes (Swedish samples including HC but excluding SC) resulted in a median BAC of 0.9286 (IQR=0.2143). Distinguishing pairs of treated phenotypes (German samples including HC), in turn, resulted in a comparably higher median BAC of 0.9655 (IQR=0.1812). This indicates treatment-dependent changes of the isomiR expression profiles, leading to an improvement of diagnostic utility in treated patients. Irrespective of the treatment the distinction between disease traits (CD, UC and SC) is estimated to be a bigger challenge than the distinction between disease and health (HC). Figure 5 summarizes distributional features of the performance of all binomial classifiers built in the course of this study.

The qualitative behavior of multinomial models solving more complex classification problems naturally depends on the choice of the binomial models considered in the process of majority voting. Accordingly, each classifier has been selected to optimize accuracy (as measured by median BAC) and stability of its votes (as measured by IQR of BAC). Applying these criteria, classical SVMs (including linear SVM: BAC=0.9375, IQR=0.1989; radial SVM: BAC=0.9375, IQR=0.2004; polynomial SVM: BAC=0.9375, IQR=0.2023) clearly outperform the penalized SVMs (including elastic net SVM: BAC=0.8750, IQR=0.2261; SCAD SVM: BAC=0.8462, IQR=0.2160). Remarkably, the least complex model (linear SVM) has been estimated to optimize both measures. However, to further increase the predictive performance we allow the combined models to incorporate different types of binomial models (see Figure 5).

Performances of the multinomial classifiers we report in Table 3. To increase reliability of the performance measures their estimation has been based on independent samples neither used for training nor evaluation of the underlying binomial models. To derive respective distributional properties LOO-CV has been employed. Increasing complexity of a classification problem compromises the performance of the model solving it. Accordingly, it is expected that multinomial classifiers are outperformed by individual binomial classifiers. However, with BACs ranging from 0.7333 to 0.9000 the estimated performance of the combined classifiers under consideration is remarkably high. Specifically, the model distinguishing CD−, UC− and SC has been estimated to perform with a mean BAC of 0.7333. The model classifying CD−, UC− and HC incorporates the comparatively less demanding distinction between IBD and HC. Accordingly, the classifier has been estimated to perform even better (mean BAC=0.8667). This equally applies to the distinction between CD−, UC−, SC and HC (mean BAC=0.8667).

<center>7</center>

However, these performances even have been exceeded while classifying CD+, UC+ and HC (BAC=0.9000), again illustrating the dependency of the models' utility on IBD treatment.

4 Discussion

With the here-presented study we, for the first time, report isomiR-based mathematical models for the distinction of CD and UC among each other as well as from healthy (HC) and symptomatic controls (SC). We report standard SVMs to be able to predict disease phenotypes with a remarkably high median balanced accuracy of 0.9375. Penalized SVMs showed comparably lower performance and/or stability and were therefore investigated in less detail. However, they might provide a foundation for functional and clinical research due to their sparsity. In contrast to the non-penalized SVMs they base phenotype predictions on isomiR signatures of minimal size. For each binomial classification problem we provide candidate isomiRs that exhibit differential expression but also allow accurate predictions in random subsamples of our data. This implies generalizability of both, the binomial as well as derived multinomial models. Notably, assessing combined models based on a single random dataset might result in optimistic or pessimistic estimates of classifier performance. However, employing LOO-CV reduces the bias.

We show validity of the models by conducting enrichment analysis of signatures of miRNA arms (derived from sets of differentially expressed isomiRs) with regard to previous work. We report non-random concordance of signatures allowing the distinction between CD, UC and HC. Notably, highest consistency we observe for classification problems incorporating treated IBD samples. This suggests that earlier studies neglected treatment as a factor influencing disease-related deregulation of isomiR expression [19, 39, 54]. While such a sampling strategy certainly corresponds to daily clinical practice, it impedes the identification of disease-specific signatures. However, not all of the studies report the information needed to prove this hypothesis [9, 42, 48]. Beyond that, signature validation by enrichment analysis assumes compatibility of underlying study cohorts. Since early biomarker studies frequently employ small study cohorts of varying homogeneity, this assumption is likely to be violated. However, by expectation this should lead to underestimation of enrichment. Accordingly, the here-reported estimates of concordance should be rather conservative.

All comparisons between the treatment groups have been conducted with respect to healthy controls. This approach enables the identification of treatment-induced differences specific for a given disease. In contrast, differences being treatment-induced but unrelated to the diseases, have not been considered to ensure the neglection of unwanted biological and technical variation. Accordingly, reliability of the results has been increased at the cost of overall comprehensiveness of the study. Notably, the excluded comparisons are assumed to be of comparably low clinical relevance. However, future investigations might consider them to elucidate the role of isomiRs with respect to treatment response.

Serological markers, such as AMCA (anti-mannobioside carbohydrate antibody), ASCA (anti-saccharomyces cerevisiae antibody) and ANCA (anti-neutrophil cytoplasmic antibody) represent common tools complementing IBD diagnostics, particularly in case of diagnostic uncertainty. Nonetheless, their performance varies considerably. In accordance to this, sole reliance on these biomarkers can not be recommended [15, 32]. AMCA and ASCA are estimated to allow a distinction between IBD and HC with BACs of 0.5050–0.6150 and 0.6050–0.7250, respectively (see [12, 47]). With median BACs of 1.0000 (for untreated CD or UC vs. HC) and 0.9828 (for treated CD or UC vs. HC) the corresponding isomiR-based classifiers are clearly superior. ASCA+/ANCA− and ASCA+ allow the more demanding distinction between CD and UC to be made with BACs of 0.7200–0.7900 and 0.5950–0.8600, respectively (see [12, 47] among others). With a BAC of 0.7857 the here-presented models are competitive. Notably, the most valid comparison of these performances certainly would have been based on biomarker levels measured for the same samples. Unfortunately, the corresponding data was not available for both, the Swedish and the German cohort.

8

The multinomial models we propose represent a distinct feature of this investigation. Whereas previous studies (with the exception of earlier work of our group [18]) focus in pairwise distinction of CD, UC, SC and HC, we provide tools for a joint discrimination of arbitrary combinations of these traits. In the first place we present models to solve 3-nomial (CD+ vs. UC+ vs. HC, CD− vs. UC− vs. HC and CD− vs. UC− vs. SC) as well as 4-nomial classification problems (CD− vs. UC− vs. SC vs. HC). With mean BACs of up to 0.9000 and 0.8667, respectively, their performance likewise has been estimated to be remarkably high. In favor of generalizability, we restricted the models to the given main phenotypes. However, extensions to multinomial models further stratifying with respect to clinical parameters, such as disease location and disease activity, do not represent a technical challenge.

Previously being dismissed as experimental artifacts, isomiRs are now considered to result from alterations in miRNA biogenesis. Recent studies show these alterations to vary with respect to gender and ethnicity as well as tissue and cell type [21, 29]. Here we ultimately provide evidence for the existence of blood-born isomiRs being specific for IBD-related traits. This suggests sequence variation of miRNAs having an impact on the pathogenesis of inflammatory bowel disease. On the other hand it questions the cell type specificity of our findings and therefore motivates further investigation.

5 Acknowledgements

6 Author contributions

Conception and design of the study: MH, GHS and AF. Generation, collection, assembly of data: MH, SJ, SN, DS, SZ, NS, JH, MD, AK, LK, SS, GHS and AF. Analysis of data: MH and SJ. Drafting or revision of the manuscript: MH and SJ. Approval of the final version of the manuscript: MH, SJ, GHS and AF.

9

Figures and tables

	German cohort			Swedish cohort
variable	CD+	UC+	CD-	UC-
general				
female	92/50 [64.8/35.2]	65/57 [53.3/46.7]	13/17 [43.3/56.7]	21/23 [47.7/52.3]
age	39.0 (28.5-48.0) [66.9]	40.0 (28.0-48.0) [66.4]	34.0 (22.0-48.0) [100.0]	33.0 (26.0-44.0) [100.0]
weight	68.0 (60.0-83.5) [66.9]	79.0 (68.5-92.0) [65.6]		
size	170.0 (166.0-178.0) [66.9]	175.0 (170.0-182.0) [66.4]		
BMI	23.3 (21.2-26.4) [66.9]	25.5 (23.1-28.4) [66.4]		
smoking	21/73 [14.8/51.4]	10/71 [8.2/58.2]		
allergy	8/6 [5.6/4.2]	6/10 [4.9/8.2]		
laboratory				
leucocytes	80.5 (43.0-120.0) [66.2]	82.0 (41.0-118.0) [64.8]		
erythrocytes	51.5 (27.0-75.0) [66.2]	61.0 (41.5-80.0) [64.8]		
platelets	64.5 (35.0-102.0) [66.2]	63.0 (39.5-95.5) [64.8]		
HB	33.0 (20.0-41.0) [66.2]	34.0 (22.0-43.0) [64.8]		
hematocrit	40.0 (36.0-42.0) [66.2]	40.0 (37.0-42.0) [64.8]		
CRP	2.5 (1.1-6.8) [66.2]	1.9 (1.0-7.2) [62.3]	32.0 (6.8-54.0) [43.3]	7.4 (2.1-9.1) [38.6]
calprotectin			1668.0 (337.0-4589.0) [20.0]	939.5 (459.5-1763.5) [9.1]
albumin			35.0 (31.5-38.0) [40.0]	37.0 (34.0-41.0) [38.6]
severity				
CDAI	57.0 (23.0-78.5) [56.3]			
Mayo (partial)				5.0 (3.0-6.0) [97.7]
Mayo (endoscopic)				2.0 (2.0-2.5) [100.0]
Mayo				7.0 (5.0-8.0) [97.7]
medication				
azathioprin	21/73 [14.8/51.4]	19/61 [15.6/50.0]		
adalimumab	17/77 [12.0/54.2]	0/80 [0.0/65.6]		
infliximab	38/56 [26.8/39.4]	10/70 [8.2/57.4]		
methotrexat	0/94 [0.0/66.2]	1/79 [0.8/64.8]		
ciclosporin	0/94 [0.0/66.2]	0/80 [0.0/65.6]		
mesalazin (systemic)	14/80 [9.9/56.3]	53/26 [43.4/21.3]		
mesalazin (topical)	5/89 [3.5/62.7]	17/64 [13.9/52.5]		
corticosteroids (topical)	2/92 [1.4/64.8]	7/72 [5.7/59.0]		
corticosteroids (systemic)	25/69 [17.6/48.6]	28/52 [23.0/42.6]		

Table 1: **Clinical characteristics.** Stratified by country-of-origin and trait the study population is summarized with regard to clinical parameters, including measures of disease severity. Categorical variables (e.g. gender) are reported in terms of absolute numbers as well as percentages of non-missing datapoints per category. Quantitative variables (e.g. age) are reported in terms of median (M), lower quartile (LQ) and upper quartile (UQ) as well as the percentage of non-missing datapoints.

10
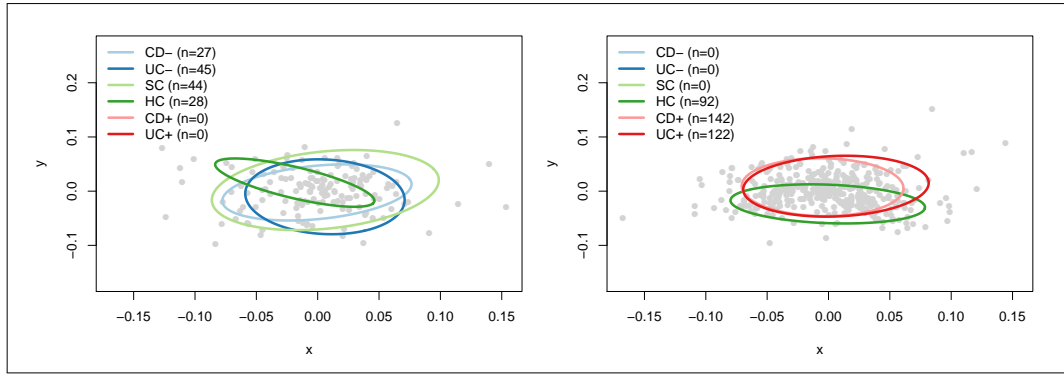
Figure 1: **Visual inspection by multidimensional scaling.** For both cohorts logarithmically transformed data with correction for technical batches as well as gender is shown. MDS plots have been generated using distance based on Spearman correlation. For each trait an ellipse tracing the bivariate normal density contour (with 25% of the samples being excluded) has been added. Ellipses are colored according to the sample's trait.
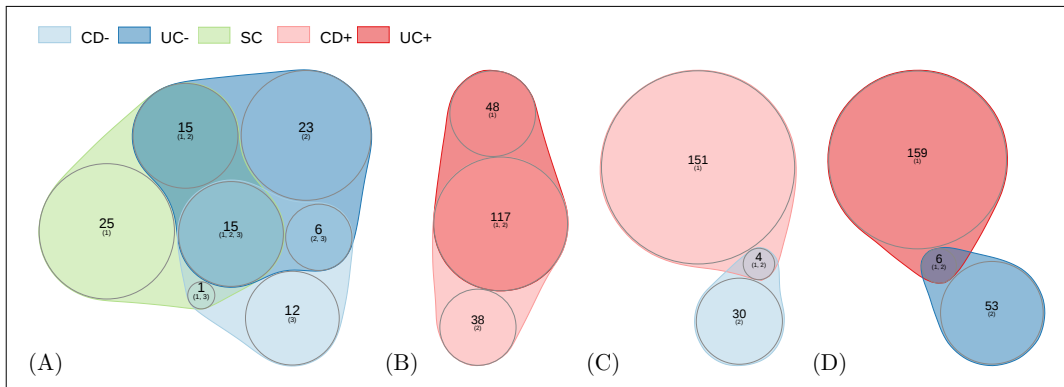
Figure 2: **Trait-specific DE miRNA arms.** Quasi-proportional Venn diagrams [41] have been generated to illustrate ovarlaps of signatures of DE miRNA arms for (A) CD−, UC− and SC, (B) CD+ and UC+, (C) CD− and CD+ as well as (D) UC− and UC+, each with respect to HC.

11
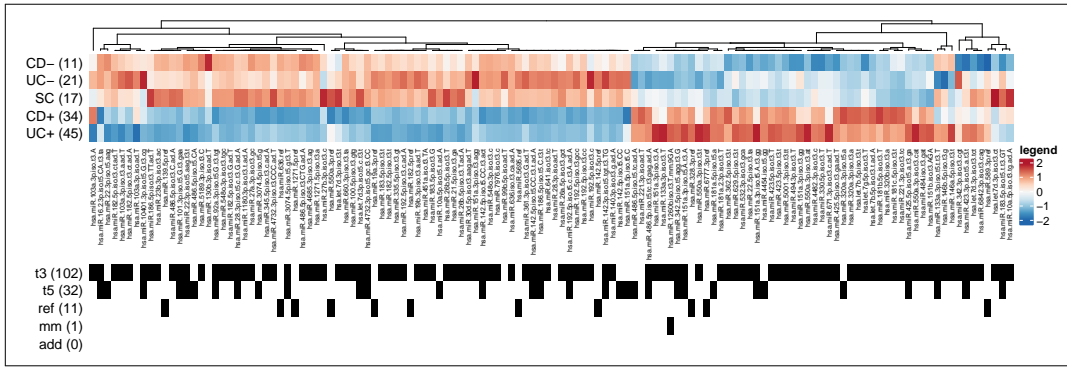
Figure 3: **Expression profile of trait-specific miRNAs.** For CD−, UC− and SC as well as CD+ and UC+ expression levels of isomiRs being significantly deregulated ($p < 0.05$ and $|\log_2 FC| > \log_2 1.5$) with respect to HC are shown as a heatmap of logarithmized fold changes. The sidebars at the bottom of the panel indicate the modification type of a given isomiR. Numbers of deregulated isomiRs per trait as well per modification type are added in parentheses.

study	technology	comparison	sample size	mirnas	overlap	Swedish cohort p-value	overlap	German cohort p-value
Paraskevi et al. [39]	RT-qPCR	CD vs. HC, UC vs. HC	378	32	5	3,66E-02*	23	4,44E-11*
Schaefer et al. [45]	RT-qPCR	CD vs. HC, UC vs. HC	90	7	4	1,57E-03*	5	2,65E-03*
Iborra et al. [19]	RT-qPCR	CD vs. HC, UC vs. HC	69	31	9	1,12E-04*	21	7,18E-10*
Zahm et al. [54]	RT-qPCR	CD vs. HC, UC vs. HC	78	21	4	3,49E-02*	16	2,63E-08*
Zahm et al. [55]	RT-qPCR	CD vs. HC, UC vs. HC	47	3	2	2,32E-02*	3	1,17E-02*
Wu et al. [52]	microarray/RT-qPCR	CD vs. HC, UC vs. HC, CD vs. UC	55	20	1	6,69E-01	11	4,17E-05*
Hübenthal et al. [17]	microarray	CD vs. UC, CD vs. HC, UC vs. HC	114	16	2	2,33E-01	7	3,06E-03*
Wu et al. [51]	microarray/RT-qPCR	CD vs. HC	24	3	0	1,00E+00	3	1,17E-02*
Duttagupta et al. [9]	microarray/RT-qPCR	UC vs. HC	40	39	4	1,72E-01	21	1,54E-08*
Takagi et al. [48]	microarray/RT-qPCR	UC vs. HC	24	2	0	1,00E+00	2	4,29E-02*
Polytarchou et al. [42]	microarray	UC vs. HC	67	4	2	3,36E-02*	2	9,50E-02

Table 2: **DE miRNAs overlapping with the literature stratified by study.** Two sets of miRNAs reported as being DE (set *A* from the literature and set *B* from our study) are compared to each other. Comparisons CD vs. HC, UC vs. HC and CD vs. UC have been investigated. In case study *B* was focusing on a particular subphenotype (inactive or active CD or UC, respectively) the corresponding main phenotype (CD or UC, respectively) has been considered. To assess whether or not the overlap with the literature was occurring by chance Fisher's exact test was applied (significance criterion p < 0.05, non-significant studies neglected). Overlaps are presented for both, the untreated Swedish as well as the treated German cohort. For a more comprehensive table, including the list of DE isomiRs per study, reference is made to Table S5.

12
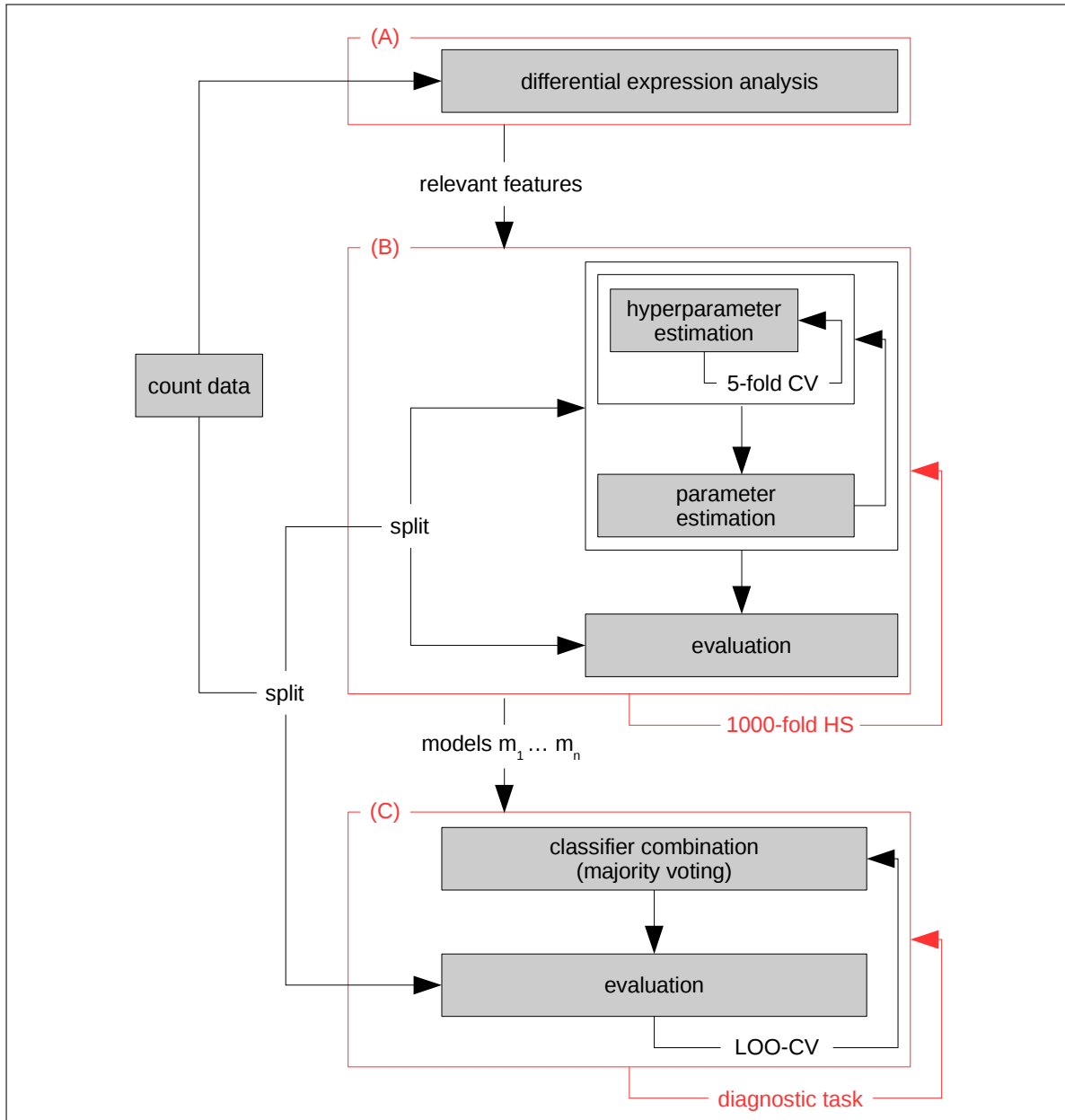
Figure 4: **Setup of the classification experiment.** The flowchart illustrates how isomiR count data has been used throughout the individual analysis steps. (A) Differential expression analysis employs the complete dataset to identify trait-specific isomiR signatures and to reduce dimensionality of the classification problems to be tackled. Random splitting into subsets is then performed for construction of atomic classifiers (B) and their subsequent combination aiming for solutions to more complex classification problems (C). Parameter estimation and subsequent evaluation is based distinct datasets repeatedly drawn at random without replacement (1000-fold holdout sampling). Selection of hyperparameters for each model being optimized thereby employs internal 5-fold cross validation. To apply binary models to solve a given diagnostic task the classification results are combined according to the principle of majority voting. Voting as well as subsequent evaluation of the final classification result is based on leave-on-out cross validation (LOO-CV).

13

Figure 5: **Classifier assessment and selection.** Performance of different types of SVMs in terms of balanced accuracy (BAC) as resulting from the application to pairwise classification problems incorporating CD−, UC−,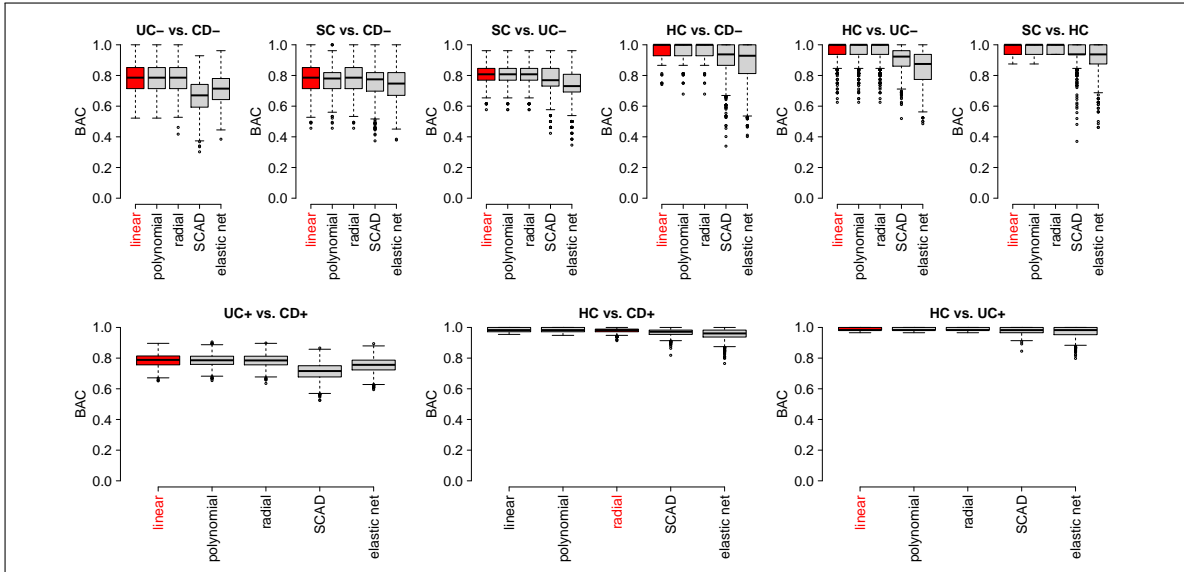 SC and HC (upper panel) as well as CD+, UC+ and HC (lower panel). For each panel pairwise comparisons have been ordered by median performance with respect to all models under consideration. Furthermore, for each pairwise comparison the model type optimizing median performance has been marked in red.

classifier	#tests	TP	FN	FP	TN	MCC	BAC	SN	SP	PPV	NPV
CD− vs. UC− vs. SC	3	11	1	6	6	0,5091	0,7333	0,9333	0,5333	0,6678	0,7785
CD− vs. UC− vs. HC	3	12	0	3	9	0,7611	0,8667	1,0000	0,7333	0,7900	1,0000
CD− vs. UC− vs. SC vs. HC	6	23	1	6	18	0,7487	0,8667	0,9667	0,7667	0,8060	0,9391
CD+ vs. UC+ vs. HC	3	10	2	1	11	0,8034	0,9000	0,8667	0,9333	0,9303	0,8921

Table 3: **Details on selected combined classifiers.** Performance in terms of Matthews correlation coefficient (MCC), balanced accuracy (BAC), sensitivity (SN), specificity (SP), positive predictive value (PPV) and negative predictive value (NPV) binary classifiers applied to combined classification problems incorporating main phenotypes (CD-, UC-, CD+, UC+, HC, SC). For each pairwise comparison the model type has been chosen to maximize the median classification performance as measured by BAC. Accuracy estimates represent the mean based on LOO-CV given 5 independent samples per trait.

14

Supplementary figures and tables

batch	CD-	UC-	SC	HC
B0001	14 (14)	11 (11)	15 (14)	
B0002	13 (13)	12 (11)	15 (15)	
B0003		11 (10)	15 (15)	
B0004		14 (13)		15 (15)
B0005				15 (13)

batch	CD+	UC+	HC
B0006	15 (15)		78 (76)
B0004	24 (20)	25 (24)	
B0007	80 (80)	58 (57)	
B0008	27 (27)	30 (29)	
B0009		12 (12)	
B0010			16 (16)

Table S1: **Experimental batches.** This table summarizes the grouping of samples being investigated in regard to sequencing technician, chemistry and run for both, the Swedish (top) and the German cohort (bottom).
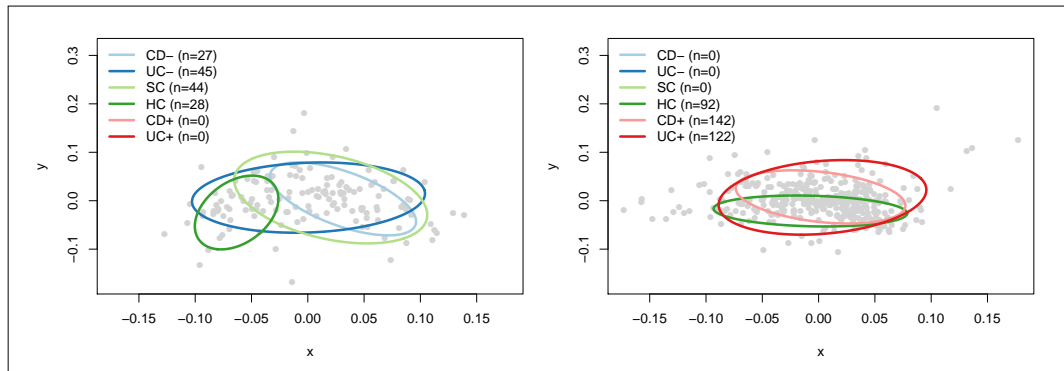
Figure S1: **Visual inspection by multidimensional scaling.** For both cohorts logarithmically transformed data is shown. MDS plots have been generated using distance based on Spearman correlation. For each trait an ellipse tracing the bivariate normal density contour (with 25% of the samples being excluded) has been added. Ellipses are colored according to the sample's trait.

Table S2: **Sequencing and mapping characteristics.** For each sample sequenced in the course of this study sequencing parameters, such as the number of raw, preprocessed and filtered reads (raw_reads, pre_processed_reads and filtered_reads), as well as mapping statistics, including the mapping efficiency are listed.

Table S3: **Differential expression analysis.** For each pairwise classification problem incorporating (A) Swedish and (B) German phenotypes, summary statistics of DE analysis, including effect size (log2FC) and effect significance (p-value), are reported.

15

```
(
  microrna[Title/Abstract] OR mirna[Title/Abstract]
) AND (
  (inflammatory bowel disease[Title/Abstract]) OR (IBD[Title/Abstract]) OR
  (ulcerative colitis[Title/Abstract]) OR (UC[Title/Abstract]) OR
  (crohn's disease[Title/Abstract]) OR (CD[Title/Abstract])
) AND (
  min_date[Date - Publication]:max_date[Date - Publication]
)
```

Figure S2: **PubMed query.** Employing rentrez::entrez_search this query has been used to gather current information on miRNAs/isomiRs being differentially expressed in regard to the traits under investigation.

Table S4: **Differentially expressed miRNAs being previously reported in the literature.** Publications resulting from the PubMed query (Figure S2) have been screened to obtain information on miRNAs being reported as DE and complemented with meta-information, such as samples type and disease/control cohort.

Table S5: **DE miRNAs overlapping with the literature.** Two sets of miRNAs reported as being DE (set A from the literature and set B from our study) are compared to each other. Comparisons CD vs. HC, UC vs. HC and CD vs. UC have been investigated. In case study B was focusing on a particular subphenotype (inactive or active CD or UC, respectively) the corresponding main phenotype (CD or UC, respectively) has been considered. To assess whether or not the overlap with the literature was occurring by chance Fisher's exact test was applied (significance criterion $p < 0.05$, non-significant studies neglected). This summary extends Table 2 by naming the DE miRNA arms being replicated from each study.

16

References

[1] G. Amarilyo et al. "miRNA in systemic lupus erythematosus." In: *Clinical immunology* 144.1 (July 2012), pp. 26–31.

[2] D. P. Bartel. "MicroRNAs: genomics, biogenesis, mechanism, and function." In: *Cell* 116.2 (Jan. 2004), pp. 281–97.

[3] N. Becker et al. *penalizedSVM: Feature Selection SVM using penalty functions.* R package version 1.1. 2012.

[4] S. Ben-Shachar et al. "MicroRNAs Expression in the Ileal Pouch of Patients with Ulcerative Colitis Is Robustly Up-Regulated and Correlates with Disease Phenotypes." In: *PloS one* 11.8 (2016), e0159956.

[5] A. M. Burroughs et al. "A comprehensive survey of 3' animal miRNA modification events and a possible role for 3' adenylation in modulating miRNA targeting effectiveness." In: *Genome research* 20.10 (Oct. 2010), pp. 1398–410.

[6] G. A. Calin et al. "MicroRNA signatures in human cancers." In: *Nature reviews. Cancer* 6.11 (Nov. 2006), pp. 857–66.

[7] N. Cloonan et al. "MicroRNAs and their isomiRs function cooperatively to target common biological pathways". In: *Genome Biology* 12.12 (2011), R126.

[8] A. D. Dhanda et al. "Can endoscopy be avoided in the assessment of ulcerative colitis in clinical trials?" In: *Inflammatory bowel diseases* 18.11 (Nov. 2012), pp. 2056–62.

[9] R. Duttagupta et al. "Genome-wide maps of circulating miRNA biomarkers for ulcerative colitis." In: *PloS one* 7.2 (Jan. 2012), e31241. PMID: 22359580.

[10] R. Edgar et al. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." In: *Nucleic acids research* 30.1 (Jan. 2002), pp. 207–10.

[11] D. Ellinghaus et al. "Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci". In: *Nature Genetics* 48.5 (May 2016), pp. 510–518.

[12] M. Ferrante et al. "New serological markers in inflammatory bowel disease are associated with complicated disease behaviour." In: *Gut* 56.10 (2007), pp. 1394–403.

[13] M. Filková et al. "MicroRNAs in rheumatoid arthritis: potential role in diagnosis and therapy." In: *BioDrugs : clinical immunotherapeutics, biopharmaceuticals and gene therapy* 26.3 (June 2012), pp. 131–41.

[14] J. H. Friedman. *Another approach to polychotomous classification.* Tech. rep. Department of Statistics, Stanford University, 1996.

[15] F. Gomollón et al. "3rd European Evidence-based Consensus on the Diagnosis and Management of Crohn's Disease 2016: Part 1: Diagnosis and Medical Management". In: *Journal of Crohn's and Colitis* 11.1 (2017), pp. 3–25.

[16] L. Guo et al. "A challenge for miRNA: multiple isomiRs in miRNAomics." In: *Gene* 544.1 (July 2014), pp. 1–7.

[17] M. Hübenthal et al. "Sparse modeling reveals miRNA signatures for diagnostics of inflammatory bowel disease". In: *PLoS ONE* 10.10 (2015), pp. 1–20. eprint: 26466382.

[18] M. Hübenthal et al. "Sparse modeling reveals miRNA signatures for diagnostics of inflammatory bowel disease". In: *PLoS ONE* 10.10 (2015), pp. 1–20. PMID: 26466382.

[19] M. Iborra et al. "Identification of serum and tissue micro-RNA expression profiles in different stages of inflammatory bowel disease." In: *Clinical and experimental immunology* 173.2 (Aug. 2013), pp. 250–8.

17

[20] W. E. Johnson et al. "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1 (2007), pp. 118–127.

[21] S. Juzenas et al. "A comprehensive, cell specific microRNA catalogue of human peripheral blood." In: *Nucleic acids research* 45.16 (Sept. 2017), pp. 9290–9301.

[22] A. Keller et al. "miRNAs in Ancient Tissue Specimens of the Tyrolean Iceman." In: *Molecular biology and evolution* 34.4 (Apr. 2017), pp. 793–801.

[23] D. Koppers-Lalic et al. "Nontemplated Nucleotide Additions Distinguish the Small RNA Composition in Cells from Exosomes". In: *Cell Reports* 8.6 (Sept. 2014), pp. 1649–1658.

[24] A. Kozomara et al. "miRBase: annotating high confidence microRNAs using deep sequencing data." In: *Nucleic acids research* 42.Database issue (Jan. 2014), pp. D68–73.

[25] K. M. de Lange et al. "Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease". In: *Nature Genetics* 49.2 (Jan. 2017), pp. 256–261.

[26] J. T. Leek et al. "The SVA package for removing batch effects and other unwanted variation in high-throughput experiments". In: *Bioinformatics* 28.6 (2012), pp. 882–883.

[27] J. Lin et al. "Novel MicroRNA Signature to Differentiate Ulcerative Colitis from Crohn Disease: A Genome-Wide Study Using Next Generation Sequencing". In: *MicroRNA* 5.3 (Jan. 2017), pp. 222–229.

[28] J. Lin et al. "Novel specific microRNA biomarkers in idiopathic inflammatory bowel disease unrelated to disease activity". In: *Modern Pathology* 27.4 (Apr. 2014), pp. 602–608.

[29] P. Loher et al. "IsomiR expression profiles in human lymphoblastoid cell lines exhibit population and gender dependencies." In: *Oncotarget* 5.18 (Sept. 2014), pp. 8790–802.

[30] M. I. Love et al. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15 (12 2014), p. 550.

[31] J. Lu et al. "MicroRNA expression profiles classify human cancers." In: *Nature* 435.7043 (June 2005), pp. 834–8.

[32] F. Magro et al. "Third European Evidence-based Consensus on Diagnosis and Management of Ulcerative Colitis. Part 1: Definitions, Diagnosis, Extra-intestinal Manifestations, Pregnancy, Cancer Surveillance, Surgery, and Ileo-anal Pouch Disorders". In: *Journal of Crohn's and Colitis* 11.6 (2017), pp. 649–670.

[33] M. Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet.journal* 17.1 (2011), pp. 10–12.

[34] D. Meyer et al. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.* R package version 1.6-8. 2017.

[35] P. S. Mitchell et al. "Circulating microRNAs as stable blood-based markers for cancer detection." In: *Proceedings of the National Academy of Sciences of the United States of America* 105.30 (July 2008), pp. 10513–8.

[36] C. T. Neilsen et al. "IsomiRs — the overlooked repertoire in the dynamic microRNAome". In: *Trends in Genetics* 28.11 (Nov. 2012), pp. 544–9.

[37] L. Pantano et al. "SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells". In: *Nucleic Acids Research* 38.5 (Mar. 2010), e34–e34.

[38] L. Pantano et al. *isomiRs: Analyze isomiRs and miRNAs from small RNA-seq.* 2017.

18

[39] A. Paraskevi et al. "Circulating MicroRNA in inflammatory bowel disease." In: *Journal of Crohn's & colitis* 6.9 (Oct. 2012), pp. 900–4. PMID: 22386737.

[40] B. C. E. Peck et al. "MicroRNAs classify different disease behavior phenotypes of Crohn's disease and may have prognostic utility." In: *Inflammatory Bowel Diseases* 21.9 (2015), pp. 2178–2187.

[41] J. G. Pérez-Silva et al. "nVenn: generalized, quasi-proportional Venn and Euler diagrams". In: *Bioinformatics* 34.13 (July 2018). Ed. by J. Wren, pp. 2322–2324.

[42] C. Polytarchou et al. "Assessment of Circulating MicroRNAs for the Diagnosis and Disease Activity Evaluation in Patients with Ulcerative Colitis by Using the Nanostring Technology." In: *Inflammatory bowel diseases* 21.11 (2015), pp. 2533–9. PMID: 26313695.

[43] D. Rachmilewitz. "Coated mesalazine (5-aminosalicylic acid) versus sulphasalazine in the treatment of active ulcerative colitis: a randomised trial". In: *Bmj* 298.6666 (Jan. 1989), pp. 82–6.

[44] W. J. Sandborn et al. "A review of activity indices and efficacy endpoints for clinical trials of medical therapy in adults with Crohn's disease". In: *Gastroenterology* 122.2 (Feb. 2002), pp. 512–530.

[45] J. S. Schaefer et al. "MicroRNA signatures differentiate Crohn's disease from ulcerative colitis". In: *BMC Immunology* 16.1 (2015), pp. 1–13. PMID: 25886994.

[46] M. R. Schneider. "MicroRNAs as novel players in skin development, homeostasis and disease." In: *The British journal of dermatology* 166.1 (Jan. 2012), pp. 22–8.

[47] C. H. Seow et al. "Novel anti-glycan antibodies related to inflammatory bowel disease diagnosis and phenotype." In: *The American journal of gastroenterology* 104.6 (2009), pp. 1426–34.

[48] T. Takagi et al. "Increased expression of microRNA in the inflamed colonic mucosa of patients with active ulcerative colitis." In: *Journal of gastroenterology and hepatology* 25 Suppl 1 (May 2010), S129–33. PMID: 20586854.

[49] A. G. Telonis et al. "Beyond the one-locus-one-miRNA paradigm: microRNA isoforms enable deeper insights into breast cancer heterogeneity." In: *Nucleic acids research* 43.19 (Oct. 2015), pp. 9158–75.

[50] D. Winter. *rentrez: Entrez in R*. R package version 1.1.0. 2017.

[51] F. Wu et al. "Identification of microRNAs associated with ileal and colonic Crohn's disease." In: *Inflammatory bowel diseases* 16.10 (Oct. 2010), pp. 1729–38. PMID: 20848482.

[52] F. Wu et al. "Peripheral blood microRNAs distinguish active ulcerative colitis and Crohn's disease." In: *Inflammatory bowel diseases* 17.1 (Jan. 2011), pp. 241–50. PMID: 20812331.

[53] F. Yu et al. "Naturally existing isoforms of miR-222 have distinct functions". In: *Nucleic Acids Research* 45.19 (Nov. 2017), pp. 11371–11385.

[54] A. M. Zahm et al. "Circulating microRNA is a biomarker of pediatric Crohn disease." In: *Journal of pediatric gastroenterology and nutrition* 53.1 (July 2011), pp. 26–33. PMID: 21546856.

[55] A. M. Zahm et al. "Rectal microRNAs are perturbed in pediatric inflammatory bowel disease of the colon." In: *Journal of Crohn's & colitis* 8.9 (Sept. 2014), pp. 1108–17. PMID: 24613022.

19

A comprehensive, cell specific microRNA catalogue of human peripheral blood

Simonas Juzenas[1,2,†]**, Geetha Venkatesh**[1,†]**, Matthias Hübenthal**[1,†]**, Marc P. Hoeppner**[1]**, Zhipei Gracie Du**[1]**, Maren Paulsen**[1]**, Philip Rosenstiel**[1]**, Philipp Senger**[3]**, Martin Hofmann-Apitius**[3]**, Andreas Keller**[4]**, Limas Kupcinskas**[2,5]**, Andre Franke**[1,*]** and Georg Hemmrich-Stanisak**[1,*]

[1]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, 24105 Kiel, Germany, [2]Institute for Digestive Research, Academy of Medicine, Lithuanian University of Health Sciences, Kaunas LT 44307, Lithuania, [3]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), 53754 Sankt Augustin, Germany, [4]Clinical Bioinformatics, Saarland University, 66125 Saarbrücken, Germany and [5]Department of Gastroenterology, Academy of Medicine, Lithuanian University of Health Sciences, Kaunas LT 50161, Lithuania

ABSTRACT

With this study, we provide a comprehensive reference dataset of detailed miRNA expression profiles from seven types of human peripheral blood cells (NK cells, B lymphocytes, cytotoxic T lymphocytes, T helper cells, monocytes, neutrophils and erythrocytes), serum, exosomes and whole blood. The peripheral blood cells from buffy coats were typed and sorted using FACS/MACS. The overall dataset was generated from 450 small RNA libraries using high-throughput sequencing. By employing a comprehensive bioinformatics and statistical analysis, we show that 3′ trimming modifications as well as composition of 3′ added non-templated nucleotides are distributed in a lineage-specific manner—the closer the hematopoietic progenitors are, the higher their similarities in sequence variation of the 3′ end. Furthermore, we define the blood cell-specific miRNA and isomiR expression patterns and identify novel cell type specific miRNA candidates. The study provides the most comprehensive contribution to date towards a complete miRNA catalogue of human peripheral blood, which can be used as a reference for future studies. The dataset has been deposited in GEO and also can be explored interactively following this link: http://134.245.63.235/ikmb-tools/bloodmiRs.

INTRODUCTION

MicroRNAs (miRNAs) are fundamental regulators in many cell biological processes. They represent a class of short (∼22 nucleotides long), non-coding RNAs that regulate gene expression at the post-transcriptional level by predominantly targeting the 3′ UTR region of target messenger RNAs (1). To date, >2500 miRNA sequences are known in humans (miRBase v21 (2)) and it has been predicted that 30–80% of human genes are influenced by at least one miRNA (3,4). Numerous mature miRNAs have been implicated in a wide range of physiologic and pathologic processes. Since the deregulation of miRNAs has been demonstrated in a range of diseases, including several types of cancer (5–8) and heart disease (9), there is a significant focus on miRNAs in biomarker research to utilize them in the prediction and early detection of diseases.

In recent years, the long-standing theory that each miRNA precursor (more precisely each arm of the hairpin molecule) produces one constant mature miRNA sequence, was disproved by results from high-throughput sequencing and subsequent bioinformatics analyses. It has been shown that a single miRNA hairpin arm can give rise to multiple distinct isoforms (isomiRs) that are now referred to as the mature miRNA transcripts and that can differ in their length and sequence composition (10). The term 'miRNA-arm' is used to define the set of all mature transcripts deriving from one arm (5p or 3p) of the miRNA hairpin molecule (11). IsomiRs are categorized into three main classes: 5′ isomiRs, 3′ isomiRs, and polymorphic isomiRs. Additionally, 5′ and 3′ isomiRs are subclassified into templated or non-templated modifications (12). Several studies have shown that isomiR expression

*To whom correspondence should be addressed. Tel: +49 0 431 500 15110; Fax: +49 0 431 500 15168; Email: a.franke@mucosa.de
Correspondence may also be addressed to Georg Hemmrich-Stanisak. Email: g.hemmrich-stanisak@ikmb.uni-kiel.de
†These authors contributed equally to this work as first authors.

profiles depend on tissue type, gender, population, ethnicity, disease type and subtype (13–16), and most importantly, that isomiRs exhibit functional differences in comparison to their archetype miRNAs (the miRNA sequences listed in public database) (17,18). These findings provide solid evidence that isomiRs have functional relevance and cannot be dismissed as experimental artefacts.

In 2008, miRNAs were first characterized in blood and described as circulating miRNAs (19–21). The detection of miRNAs in circulation provided an opportunity to use miRNAs as non-invasive biomarkers for the distinction of biological/clinical conditions. Since then numerous studies have reported circulating miRNAs as biomarkers for a variety of cancers and other diseases (22–24). However, the origin and especially the sequence variation of these miRNAs has been poorly investigated. In a recent study, Pritchard *et al.* (25) examined 79 solid tumor circulating miRNA biomarkers reported in the literature and showed that 46 of 79 are highly expressed in different blood cells. This provided first insights that blood cells substantially contribute to circulating miRNA levels, be it because of contamination by hemolysis or regular physiological processes. Since then, only a few research projects focused on the miRNA transcriptomes of either several or single separated blood cell types (26–29) and even fewer focused on the miRNA sequence variations in distinct cell types (30,31). Yet, there is still only scarce knowledge about the blood cell origin of circulating miRNAs, their expression pattern with respect to different cell types and especially about their sequence variations in different blood compounds. Therefore, it is important to further study miRNA expression of different blood cells in order to identify cell type specificity.

Because of its high availability in clinical practice, peripheral blood is also the most commonly collected body fluid. The majority of non-invasive miRNA biomarker studies, thus, currently focus on the measurement of blood miRNAs, although numerous other fluid compounds are also under investigation. Methodologically, besides quantitative PCR (qPCR) and array-based technologies, small RNA sequencing (sRNA-Seq) has become a popular technology to establish miRNA-based transcriptional profiles. Interpretation of results gained from those blood-based sRNA-Seq studies, however, remains a challenging task (32–34). Most sRNA-Seq protocols are PCR-based and depend on adapter-ligation steps and often suffer from the limitation that the measurement of a particular miRNA is not independent from other miRNAs. The resulting introduction of biases towards certain miRNAs may lead to discrepancies in the overall abundance of sequenced miRNAs. In addition to the above-described technological problems, blood as liquid tissue has only a relatively low abundance of miRNAs compared to solid tissues, which decreases the signal-to-noise ratio and renders the measurement of rare miRNAs difficult. Moreover, due to the differential miRNA expression in different blood cell types and the uneven distribution of cell types based on different conditions, e.g. diseases, it is difficult to distinguish between true positive, false positive but also potential false negative signals in sRNA-Seq datasets. Finally, depending on the treatment of the sample material, hemolysis can occur and introduce another source for altered measurements (34).

Given the above-described difficulties, when studying sRNA-Seq-based miRNA profiles, it is crucial to first understand how miRNAs and their isoforms are distributed in human blood before interpreting a deregulated state, e.g. in the context of a disease. For this purpose, a comprehensive reference dataset for miRNA expression in the different components of healthy human blood (cell types, vesicles and fluid) is urgently needed but currently only fragmentary in the public domain. The results presented here thus summarize our (still non-exhaustive, but substantial) contribution towards such a reference dataset. We report the transcriptional profiles of miRNAs and isomiRs from seven distinct peripheral blood cell populations of 43 healthy donors using sRNA-Seq. In addition, we also examined whole blood profiles ($n = 77$) as well as circulating miRNAs from serum ($n = 38$) and blood-borne exosomes ($n = 38$). The results can be interactively browsed at http://134.245.63.235/ikmb-tools/bloodmiRs.

MATERIALS AND METHODS

Study samples

A total of 162 blood samples from healthy volunteers (routine blood donors from transfusion medicine) were included in this study. Of the 162 samples, 43 buffy coats remaining from plasma donation were collected. Forty two serum samples were collected into serum collection tubes (BD Vacutainer 366643) and 77 whole blood samples were collected into PAXgene RNA blood tubes (Qiagen). All donors signed a written informed consent form. Approval for the study was received from the Ethics Committee of the Medical Faculty, University, Kiel.

Isolation and purification of cell types

Human peripheral blood cells were isolated from leukocyte concentrates (buffy coats) of healthy donors by Ficoll density gradient centrifugation. Briefly, buffy coats (~50 ml) were dissolved in 100 ml phosphate-buffered saline (PBS) and 15 ml lymphocyte separation medium/Ficoll ($d = 1.077$ g/ml) were carefully added and centrifuged at 2000 rpm for 20 min (divided in four separate 50 ml tubes, without using the brake function). After centrifugation, peripheral blood mononuclear cells (PBMCs) enriched in the interface between plasma and Ficoll were carefully gathered. After washing three times with PBS (1×5 min, 1600 rpm; 2×10 min, 1000 rpm), cells were resuspended and an aliquot counted by fluorescence-activated cell sorting (FACS). Granulocytes and RBCs, which are not within the PBMC fraction, were extracted separately by pipetting 1000 μl from the interface between Ficoll and red blood cells (RBCs) or 100 μl from the bottom of the tube, respectively.

Magnetic activated cell sorting (MACS, Miltenyi Biotec, Bergisch-Gladbach, Germany) was used to purify CD56+ (NK cell), CD19+ (B cell), CD8+ (cytotoxic T cell), CD4+ (T helper cell), CD14+ (monocyte), CD15+ (neutrophil) and CD235a+ (erythrocyte) cell populations, following the manufacturer's instructions. CD14+, CD15+, CD19+ and CD235a+ cells were isolated by positive selection, whereas CD4+, CD8+ and CD56+ cells were isolated by negative

selection. Purity of the individual cell populations was assessed using FACS.

Total RNA isolation and sRNA-Seq

Total RNA including small RNA from sorted cell types was extracted using the mirVana RNA Isolation Kit (Ambion). For whole blood, we used the PAXgene Blood miRNA Kit (Qiagen). For serum samples miRNeasy Serum and Plasma kit (Qiagen) was used and exosomes were processed using the Total Exosome Isolation Reagent and the Total Exosome RNA and Protein Isolation Kit (Life Technologies). All isolation protocols were conducted according to the manufacturers' instructions, without further modifications.

Extracted total RNAs were combined with a spike-in cocktail (except whole blood samples) as previously described by Hafner *et al.* (35). The products were then subjected to Illumina TruSeq Small RNA Sample Preparation protocol to generate small RNA libraries for each sample. Subsequently, the libraries were randomized and pooled with six samples per lane for serum and exosomes, and 24 samples per lane for cell types and whole blood. Sequencing was conducted on an Illumina HiSeq 2500 (1 × 50 bp SR, v3). Raw sequencing reads and quantified read-count data have been deposited at NCBI Gene Expression Omnibus (GEO) (36) under the accession number GSE100467.

Data analysis

Obtained raw reads were subjected to exhaustive adapter trimming using cutadapt v1.3 (37). Low quality 3′ ends of reads falling below Q20 (Phred score) were discarded. Only the quality trimmed reads that were longer than 18 bp were retained. Subsequently, the reads mapping to viral genomes, viral miRNA precursors and human non-miRNA short RNAs were filtered out. The remaining reads were mapped to miRNA sequences from miRBase v21 (2) using mirAligner (38) with default parameters (1 mismatch, 3 nt in the 3′ or 5′ trimming variants and the 3 nt in 3′-addition variants) and to spike-in sequences. As the latter were not available for all samples analysed, the respective read counts were neglected for further analysis. The R package isomiRs (38) was used to generate counts of miRNA-arms (read count sums of all isomiRs including archetype miRNA which align to the same miRBase entry) and isomiRs. Samples with fewer than one million of initial reads and those for which the number of expressed miRNA reads on a logarithmic scale was falling below Q1 − 1.5 IQR were excluded from further analysis. miRNA-arm and isomiR sequences that were expressed (expression value ≥ 5) in at least 85% of the samples in at least one of the cell types were normalized using DESeq2 (39). Normalized counts were checked for potential confounding effects such as age, gender, cell counts, cell purity, levels of hematocrit, number of leukocytes and erythrocytes. The counts were then subjected to pairwise comparisons between different cell types or combined cell type groups. The *P*-values resulting from Wald tests were corrected for multiple testing according to Benjamini and Hochberg (40). The miRNAs with a corrected *P*-value ≤ 0.001 and |log₂FC| > 1 were considered to be significantly differentially expressed. The miRNAs which were uniquely significantly upregulated in one

of the cell types when compared to every other cell type were called as being cell type specific miRNAs. In order to determine the genomic context of cell type specific miRNAs, the Ensembl API was used to retrieve annotated features (introns, exons and UTRs of coding and non-coding transcripts) that overlapped miRNA precursors. The miRNAs that did not intersect any transcript features were considered as intergenic. Furthermore, miRNA profiles from cell types were qualitatively compared with those from whole blood samples, serum and exosomes.

Statistical analyses and data processing were done in R version 3.3.3 (41). Visualization of graphs and heatmaps was performed with ggplot2 (42) and ComplexHeatmap (43) packages, respectively.

Detection of novel miRNAs

In order to detect novel miRNAs, raw sequencing reads from all samples were pooled into a single file to increase the power of prediction (44). Subsequently, pooled data were trimmed for adapters using cutadapt v1.3 and collapsed to obtain unique sequences while saving the information about read counts. The resulting data were mapped and filtered against spike-ins, viral genomes (RefSeq (45)), viral miRNA precursors (miRBase v21 (2)) and other human small RNAs (rRNAs, tRNAs, snRNAs and sRNAs (Rfam (46)) using BLASTN v2.2.30 (47). The remaining reads were mapped to the human genome (version hg19) using Bowtie v1.1.1 (48) and only those reads that mapped to the reference genome were passed down for novel miRNA prediction employing miRDeep2 v2.0.07 (49) with default parameters and as input for related species using known miRNAs from *Hominidae* family (miRBase v21). In order to reduce the number of false positively predicted miRNAs, several filtering steps were applied to remove: (a) novel miRNA candidates with miRDeep2 score < 1; (b) predicted precursors, which overlapped with human coding sequences (CDS) and non-coding RNAs (ncRNAs) (reference genome version hg38, Ensembl (50)); (c) novel miRNA precursors which were enriched with repetitive DNA element sequences from Dfam (51); and d) predicted miRNAs, which had outlying values based on quantiles (lower quantile = 0.01, upper quantile = 0.99) of GC content of known miRNA mature sequences belonging to *Hominidae* family (miRBase 21). Filtered novel miRNA candidates were ranked according to their distance to the features in the early miRBase (v1–v7) versions using novo-miRank (52).

Sample-wise quantification of candidate novel miRNAs was performed using quantifier.pl module from mirDeep2 v2.0.07 with default parameters. Normalization and statistical analysis of novel miRNA candidate count dataset was performed as described in the previous section.

Web-page implementation

To make our results easily accessible to researchers in the field, we implemented a web-tool for interactive browsing of miRNA expression results of peripheral blood cell populations. It contains miRNA-arm and isomiR expression data which can be explored by cell type or by miRNA of interest. The differential expression analysis can be performed

4 Nucleic Acids Research, 2017

in real time by selecting different cell types or combined cell type groups for pairwise comparisons. The web-tool was implemented using shiny web application framework for R (53), as well as R packages such as DESeq2 (39), dplyr (54), plotly (54) and iheatmapr and is available at http://134.245.63.235/ikmb-tools/bloodmiRs.

RESULTS

Overview of the small RNA transcriptome data

Illumina deep sequencing was conducted to profile miRNA expression in major peripheral blood cell populations, as well as serum, exosomes and whole blood of healthy individuals. We used magnetic activated cell sorting (MACS) to purify blood cells (see Figure 1) from the lymphoid lineage (CD56+ (NK cells), CD19+ (B cells), CD8+ (cytotoxic T cells), CD4+ (T helper cells)) and the myeloid lineage (CD14+ (monocytes), CD15+ (neutrophils) and CD235a+ (erythrocytes)). Purity of the sorted cells was tested using FACS analysis. Most cell populations were rather homogeneous with purity over 90%. Neutrophils, however, displayed considerable heterogeneity with a mean sorting purity of 79% (see Supplementary Figure S1A). Overall, 298 samples were sorted; however, we failed to sort erythrocytes for three samples, and T helper cells for one sample.

Subsequent total RNA isolation and small RNA sequencing of the above-mentioned cell fractions and blood compounds resulted in approximately 2.56 billion (B) raw sequencing reads, ranging from 212.7 thousand (K) to 32.7 million (M) reads per sample (Supplementary Figure S1B). This wide range in the raw sequences was expected and could occur due to individual variability (health, blood counts etc. of the donor) and due to variability in experimental factors such as different RNA isolation methods (i.e. mirVana, PAXgene, miRNeasy etc.), and/or different amounts of small RNAs in distinct blood compounds, which lead to unequal adapter dimer contamination and unequal spike-in distribution when preparing libraries (55).

Figure 1. Simplified human hematopoietic tree of different cell compounds (modified from Häggström (69)). The miRNA expression profiles were generated for natural killer (NK) cell (CD56+), B lymphocyte (CD19+), cytotoxic T cell (CD8+), T helper cell (CD4+), monocyte (CD14+), neutrophil (CD15+) and erythrocyte (CD235a+) populations (highlighted in red).

After the adapter and quality trimming step, we retained 91.3% (2.34B) of the initial sequencing reads, which shows the high quality of the dataset. The bulk of those reads were 19–24 nt length which corresponds to the length range of mature miRNA sequences (Supplementary Figure S1C). The filtering steps (see Methods section for details) retained 81.4% (2.09B) of the initial sequencing reads. The majority of filtered reads belonged to human non-miRNA short RNAs from Rfam database (Supplementary Figure S1D). The remaining high quality reads were mapped to known miRNA (miRBase v21) and spike-in sequences (see Materials and Methods section for details). The overall composition of the initial reads per sample in distinct blood compounds is presented in Supplementary Figure S1E. Quantification of filtered reads yielded 719.4M sequences to be mapped to 2106 unique known miRNA sequences from miRBase v21. As expected, the lowest numbers of miRNA counts were observed in transcriptionally non-active blood compounds such as RBCs, exosomes and serum (Supplementary Figure S1F). However, low miRNA read counts were also observed in neutrophils (CD15+) which was probably due to the lower cell sorting purity. After applying our QC threshold, 19 low read and/or outlying samples were removed, leaving 431 samples for downstream analysis. On average, we mapped 540 unique miRNA species per peripheral blood cell type, 162, 267 and 645 unique miRNA species per exosomes, serum and whole blood, respectively (Supplementary Figure S1F).

Sequencing and mapping statistics for all libraries used in this study can be found in Supplementary Table S1.

Anucleate blood harbors a substantial miRNA repertoire

Triggered by earlier reports about miRNAs being also present in RBCs (25, 27), we first qualitatively characterized the RBC miRNA content. Only miRNAs having at least five reads in at least 85% of the samples were considered in our analysis. Surprisingly, as anucleate cells should not be transcriptionally active, we identified 271 miRNA species within RBCs—several of them with considerable read counts (>1000/sample, see also Figure 2). In addition, we generated miRNA data from serum samples ($n = 38$) and exosomal vesicle preparations for 38 samples (34 of which are paired samples) of healthy donors, which resulted in another 90 and 51 miRNA species, respectively. When comparing the miRNA content of RBCs to those of serum and exosomes, we found a considerable overlap between all three fractions pointing towards a common miRNA repertoire of those blood compounds. We found 38 miRNA species present commonly in serum, exosomes and RBCs, whereas 19 miRNA species were found to be uniquely present in serum, 2 and 21 miRNA species in exosomes and RBCs, respectively (Figure 2). In total, we were thus able to detect 93 unique miRNA species outside the transcriptionally active cellular components of healthy human blood.

Sequence variation analysis reveals lineage specific 3' end modification patterns of isomiRs

The growing evidence of constitutive variability in miRNA biogenesis across human tissues encouraged us to examine the sequence variation patterns of mature miRNAs
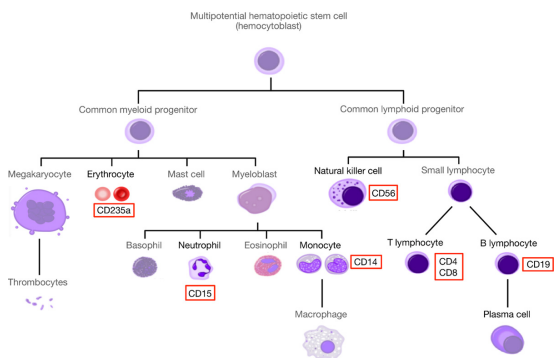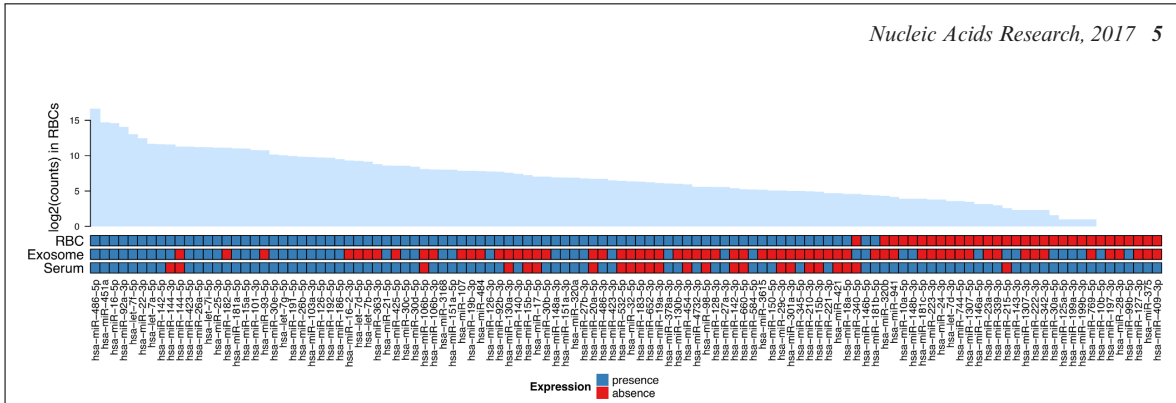
Figure 2. The presence of miRNAs in erythrocytes (RBCs), exosomes and serum. The barplot shows the $\log_2$ transformed median read counts of the most abundant miRNAs in RBCs. The heatmap represents presence (blue) or absence (red) of miRNAs in specific blood compounds. Only the miRNAs which were detected (expression value > 5) in at least 85% of the samples in at least one of the blood compound were considered as being present.

within different peripheral blood compounds. Accordingly, we used mirAligner to identify trimming variations at the 5′ and 3′ ends, nucleotide (nt) substitutions and non-templated 3′ additions in miRNA sequences. In order to avoid sequencing errors, we consider a mismatch as a real substitution, when the fraction of reads having that substitution was >2% within the mapped reads to a certain miRNA sequence. Additionally, we only kept those substitutions which mapped uniquely to one miRNA molecule.

By comparing miRNA sequences obtained from our dataset to the archetype sequences deposited in miRBase, we identified sequence variations in 69.0% (1454/2106) of detected miRNAs. The most common miRNA sequence variation types were non-templated 3′ addition and 3′ trimming, which were observed in 77.0% (27539/35769) and in 75.6% (27032/35769) of isomiR sequences, respectively. Conversely, 5′ trimming (33.6%, 12025/35769) and nt-substitutions (6.2%, 2229/35 769) within the mature miRNA sequence were the least common variants. About one third (33.9%, 755/2229) of nt-substitutions occurred in the seed sites of miRNAs. More than one type of sequence variation was observed in 74.0% (26482/35769) of isomiRs.

To get deeper insights into isomiR variation preferences, we looked at variation distributions in our investigated blood compounds. The unique miRNA sequences having 5′ trimming variations were similarly distributed across the blood compounds, except that in neutrophils (CD15+), serum and exosomes this variation at position −1 with respect to reference sequences was more prevalent than in other cell types (Figure 3A). Surprisingly, 3′ trimming distributions showed clear lineage-specific patterns across the cell types. The frequency of unique sequences having 3′ trimming modification at positions −3 and −2 was higher in lymphoid than in myeloid lineage, whereas from position −1 to 3 it changed, and unique sequences having 3′-trimming variation became more abundant in the myeloid lineage. Meanwhile, the 3′ trimming distributions in serum and exosomes showed similar patterns, and in most of the positions were distinct from RBCs and whole blood samples (Figure 3B). The nt-substitutions were distributed similarly across the blood compounds, except for serum at positions 1 and 4 and for exosomes from position 12 to 14. Interestingly, the overall distribution of substitutions within

miRNA sequences revealed a consistent pattern of positions in which nucleotides are less frequently substituted, pointing toward the nucleotides which are the most important in the base-pairing during miRNA–target interaction, i.e. the ones within the seed site (2–8 nt) of miRNAs (Figure 3C). The fractions of unique sequences having 3′ non-templated nucleotides were similar in neutrophils (CD15+), RBCs, serum and exosomes. Unique sequences having this modification were highly frequent at position 1 and vice versa at position 2 of the above-mentioned compounds when compared to all the others (Figure 3D).

We were interested to further explore whether the anucleate compounds such as RBCs, serum and exosomes could also share the similar composition of the added nucleotides, therefore we decided to look at the non-templated 3′ additions more closely. Remarkably, the composition of 3′ added nucleotides, the same as 3′ trimming preference, was lineage-specific. The composition of non-templated nucleotide additions clustered into three distinct groups, corresponding to myeloid lineage, lymphoid lineage and noncellular blood compounds, including serum and exosomes (Figure 4). These results suggest that most of the RBC-miRNAs probably originate from RBCs (or their progenitors) and that the majority of miRNAs found in serum and exosomes are not the products of hemolysis.

Differential expression analysis identifies blood cell specific miRNA transcription signatures

To determine whether miRNA expression profiles (precisely miRNA-arms) could distinguish cell types, we first performed multidimensional scaling analysis which revealed a consistent clustering within and between cell types with only a very small proportion of outliers (see Figure 5). Even the distinction of lymphoid and myeloid blood cells is clearly visible with the red blood cell fraction forming a separated cluster. Also, the already mentioned heterogeneous neutrophil cells form a consistent cluster pointing towards a consistent sorting result.

Analysing the repertoire of previously known miRNAs listed in miRBase v21, 382 of the currently annotated 2578 miRNAs were found to be expressed in >85% of the samples in at least one of the cell types. Pairwise differential expression analysis between different cell types revealed
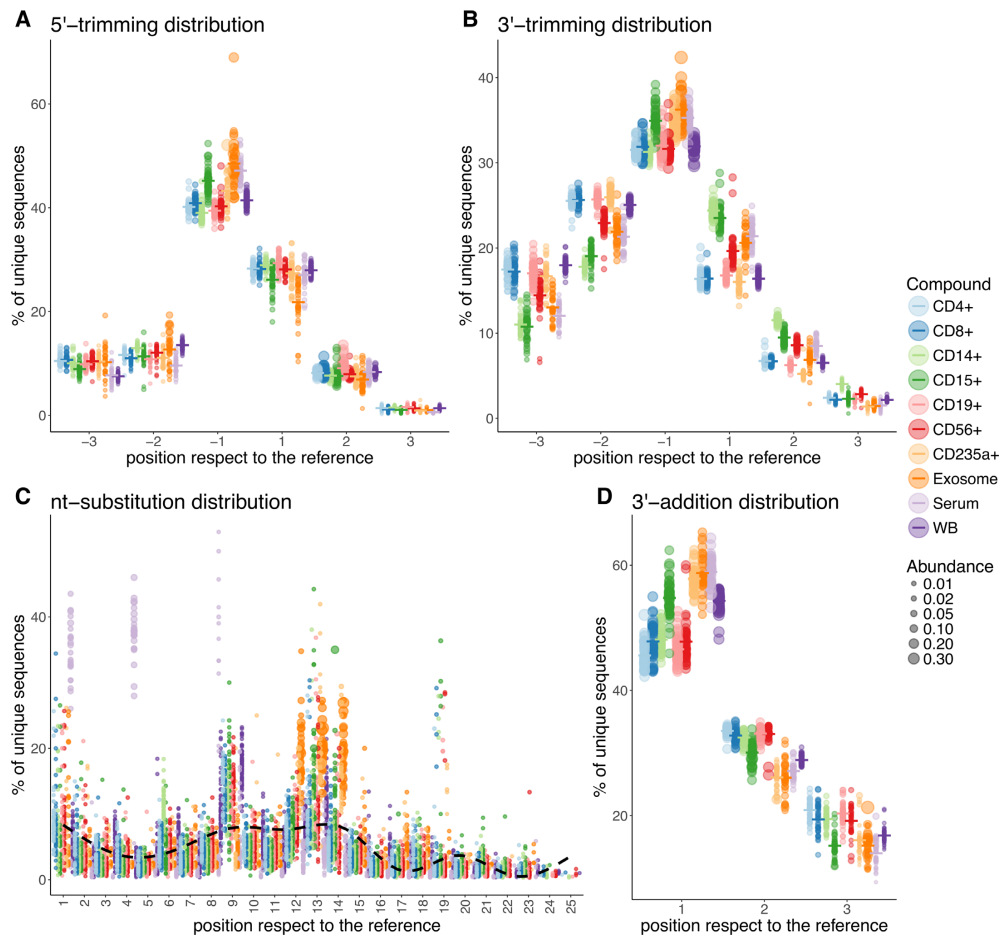
6 Nucleic Acids Research, 2017

Figure 3. The distributions of isomiR modification types across different blood compounds. (**A**) 5′ end trimming modification distribution; (**B**) 3′ end trimming modification distribution; (**C**) nucleotide substitution modification distribution; (**D**) 3′ end addition modification distribution. Each dot represents individual sample colored by group. The size of dots indicates the abundance which corresponds to the relative number of molecule copies per modification type and position present in a given sample. The dash (–) symbol indicates the mean frequency of unique sequences per group. The smoothing in figure C was performed using the generalized additive model (GAM) method.

224 of the 382 arms being significantly differentially expressed (FDR corrected P-value < 0.001; |$\log_2$FC| > 1) between the two major blood cell lineages (lymphoid versus myeloid). We excluded the RBC data from this expression analysis, as RBCs per definition are anucleate and do not express genes. As depicted in Figure 6A, 77 out of 224 differentially expressed miRNAs are up-regulated in lymphoid cells, whereas 167 miRNAs are up-regulated in the myeloid lineage. Furthermore, by focusing on the subtypes of lymphoid and myeloid cells, we found that 149 miRNAs were significantly different between natural killer cells (CD56+) and small lymphocytes (CD4+, CD8+ and CD19+), 119 miRNAs between T-lymphocytes (CD4+ and CD8+) and B-lymphocytes (CD19+) and 193 between CD14+ and CD15+ cells. For more details reference is made to http://134.245.63.235/ikmb-tools/bloodmiRs.

We next focused on miRNAs that were significantly higher expressed in one given cell type when compared to every other cell type (see Figure 6B). Those miRNAs could be considered part of the cell type-specific miRNA-arm expression signature. Of the 382 miRNAs that were expressed in at least one of the investigated cell types, 136 miRNAs followed the above pattern (FDR corrected P-value < 0.001 and $\log_2$FC > 1). For CD4+ (T helper cells), CD8+ (cytotoxic T cells), CD19+ (B cells) and CD56+ (NK cells) we identified 9, 3, 10 and 18 specific miRNAs, respectively, whereas for CD14+ (monocytes) and CD15+ (neutrophils) we detected 25 and 50 specific miRNAs, respectively. The majority of these miRNAs were derived from
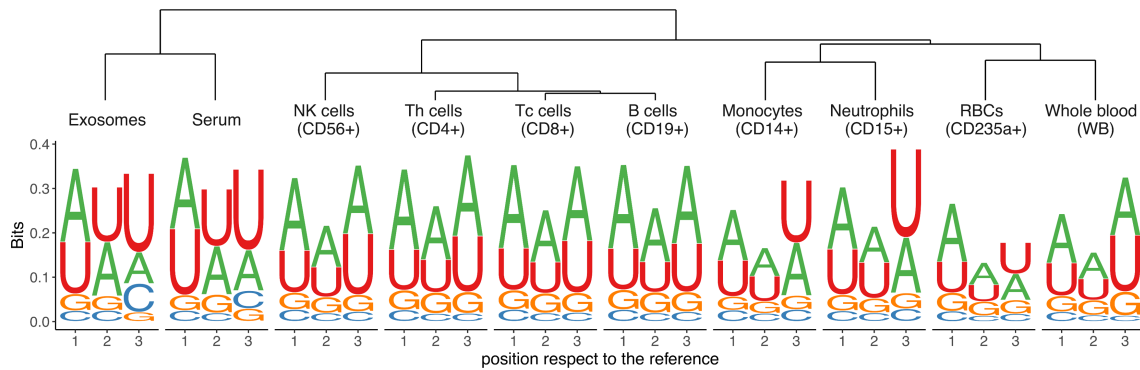
Figure 4. The composition of 3′ added nucleotides across the blood compounds. A hierarchical average linkage clustering was used to generate a dendrogram based on relative frequencies of bases per position of the unique sequences, which then were visualized as sequence logos.

intronic regions of protein coding genes. Genomic context and miRNA family information of cell type specific miRNA-arms can be found in Supplementary Table S2.

To get an overview on miRNAs that are present in human blood but have not been covered by our cell sorting approach, we additionally generated miRNA expression data for whole blood samples. We found 417 miRNAs that were expressed in 85% of the samples in at least one of the cell types or whole blood samples. Compared to the above-mentioned 382 miRNA species of the cell type specific datasets, we detected additional 35 miRNAs which are either specific for other blood cell types or originate from other sources (Supplementary Table S3).

Distinct isomiRs yield higher cell type specificity than miRNA-arms

Taking into consideration that isomiRs can make a significant contribution to miRNA representation, we repeated the pairwise differential expression analysis between different cell types and/or cell type groups by using isomiR counts. Applying the previously described threshold, we identified 2538 expressed isomiRs arising from 309 distinct miRNA-arms. Pairwise comparison between lymphoid and myeloid lineages revealed 1862 differentially expressed isomiRs deriving from 281 arms, 912 of those isomiRs being up-regulated and 950 being down-regulated. Analogous comparisons of the subtypes of lymphoid and myeloid cells displayed 1204 (derived from 211 miRNA-arms) differentially expressed isomiRs between natural killer cells (CD56+) and small lymphocytes (CD4+, CD8+ and CD19+), 773 (from 179 arms) isomiRs between T-lymphocytes (CD4+ and CD8+) and B-lymphocytes (CD19+) and 1420 (from 246 arms) between CD14+ and CD15+ cells. For detailed information reference is made to http://134.245.63.235/ikmb-tools/bloodmiRs.

As expected, when looking at higher resolution we were able to detect considerably higher number of cell type specific isomiRs than by looking only at the miRNA-arms. In total, we identified 800 isomiRs which were uniquely significantly up-regulated in one of the cell types when compared to every other cell type. For lymphoid lineage CD4+ (T helper cells), CD8+ (cytotoxic T cells), CD19+ (B cells) and CD56+ (NK cells) we identified 31, 5, 57 and 212 cell type specific isomiRs, respectively, whereas for myeloid lineage CD14+ (monocytes) and CD15+ (neutrophils) cells we detected 206 and 289 cell type specific isomiRs, respectively (Supplementary Figure S2). The number of identified isomiRs per cell type was highly correlated with the number of identified cell type specific miRNA-arms (Pearson's $r = 0.9$).
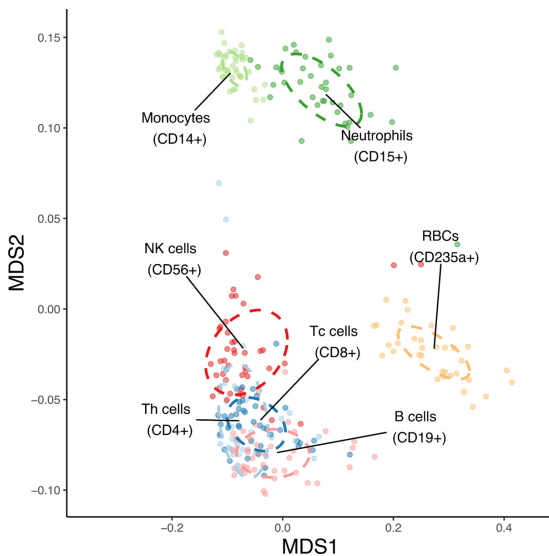
Figure 5. The similarity structure of human blood cell miRNA transcriptomes. MDS plot showing three clearly resolved clusters corresponding to lymphoid cells (NK cells, B cells, T cells and Th cells), myeloid cells (monocytes and neutrophils) and anucleate erythrocytes. The analysis was performed on miRNA count data using Spearman's correlation distance (1 – correlation coefficient). The dots represent samples coloured by group, while the centre of ellipses corresponds to the group mean and the shapes are defined by the covariance within group.
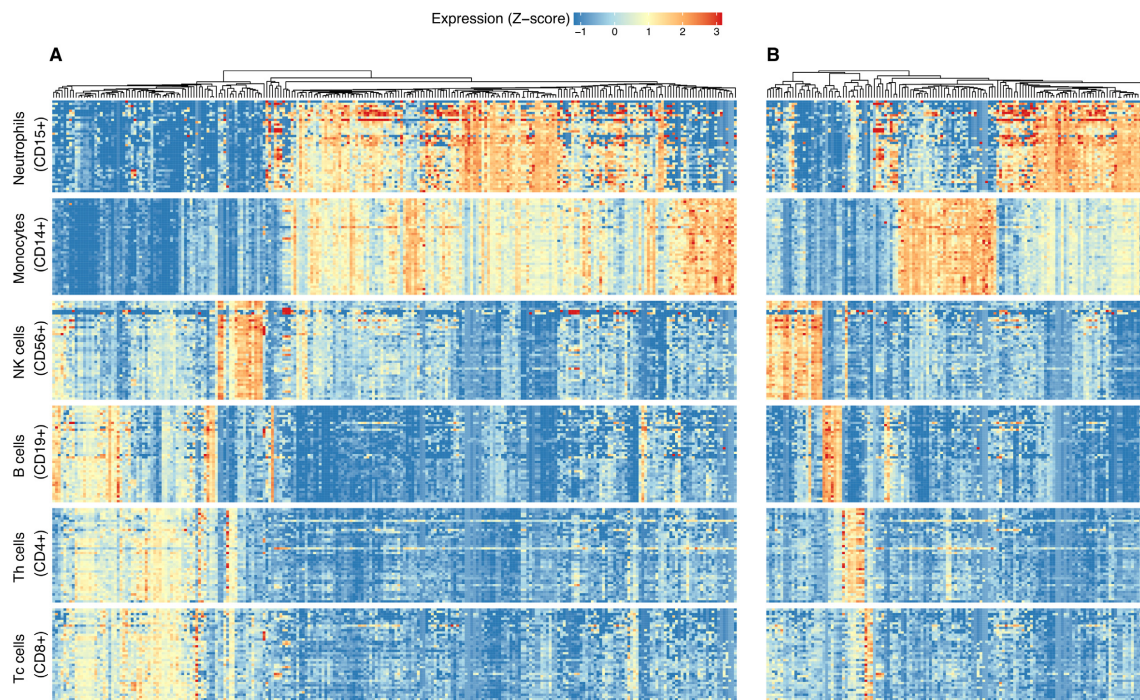
Figure 6. (**A**) Differentially expressed miRNAs between lymphoid and myeloid cell lineage; (**B**) expression levels of blood cell lineage-specific miRNAs. Each row of the heatmaps represents a sample corresponding to the one of the blood cell types, and each column represents an individual miRNA-arm. All miRNAs were statistically differentially expressed (FDR $\leq$ 0.001 and |log$_2$FC| > 1). The Z-score in the heatmaps represents standardized normalized expression values. The unsupervised (agglomerative) hierarchical clustering of miRNAs was performed using Spearman's correlation distance (1 − correlation coefficient) as metric and average linkage clustering as linkage criterion.

Modification types of cell type specific isomiRs were similarly distributed as in the general population of identified isomiRs. The most common were 3′ trimming and 3′ addition modifications, and the least common were 5′ trimming and nucleotide substitutions (Table 1). Interestingly, on average, only ∼56% of isomiRs per cell type were arising from previously identified cell type specific miRNAs, meaning that the rest were not captured when only looking at the miRNA-arm level. The complete list of cell type specific miRNA-arms and isomiRs identified in this study is provided in Supplementary Table S4.

Identification of novel, blood cell specific miRNAs

In addition to measuring the expression of known miRNAs, we also investigated whether novel miRNAs are hidden in the generated sRNA-Seq data. For that purpose, pooled data for each blood cell type were analyzed using the de novo prediction functionality of miRDeep2. Overall, we obtained 716 novel miRNA candidates predicted by miRDeep2. After removing false positives and performing several filtering steps (Supplementary Figure S3), we retained 413 most probable novel miRNA candidates and ranked the sequences according to their similarity of known miRNA features using novo-miRank. Of these 413 novel miRNA candidates, 52 molecules shared the same seed sequence with known miRNAs from the taxonomic family of primates (*Hominidae*). Characteristics such as genome context (intronic, UTR or intergenic), nucleotide composition, minimum free energy (MFE) structure etc. of the novel candidates are provided in the Supplementary Table S5. Additionally, precursor structures, read signatures and prediction scores of novel candidates are provided in the Supplementary File S1. Moreover, 50 out of 197 common novel miRNAs were validated by the hybridization method in recent study by Fehlmann *et al.* (56).

Next, we focused on differential expression of novel miRNAs in cell types and quantified novel miRNA candidates in a sample-wise manner. We detected 76 out of 413 candidates in at least 85% of the samples in at least one cell type, and 26 of those were identified as cell type specific (Supplementary Figure S4). Interestingly, with our thresholds, unlike in previous reports, we were not able to identify novel miRNA candidates within the RBC fraction.

DISCUSSION

Most sRNA-based biomarker studies today investigate archetype miRNA or miRNA-arm expression in either serum, exosomes or whole blood, while only a small number of studies focus on sub-compartments of blood, such as the peripheral mononuclear cell fraction (PBMCs) or even

Table 1. Summary of cell type specific isomiRs modification types

Cell type	Number of isomiRs	5′-trimming isomiRs	Seed substituted isomiRs	3′-trimming isomiRs	3′ nt-added isomiRs	nt-substituted isomiRs	Multi-modified isomiRs
Monocytes (CD14+)	206	44 (21.4%)	1 (0.5%)	112 (54.4%)	82 (39.8%)	1 (0.5%)	70 (34%)
Neutrophils (CD15+)	289	52 (18%)	0 (0%)	166 (57.4%)	142 (49.1%)	4 (1.4%)	99 (34.3%)
B cells (CD19+)	57	7 (12.3%)	0 (0%)	42 (73.7%)	39 (68.4%)	0 (0%)	32 (56.1%)
T helper cells (CD4+)	31	0 (0%)	0 (0%)	26 (83.9%)	15 (48.4%)	0 (0%)	14 (45.2%)
NK cells (CD56+)	212	52 (24.5%)	0 (0%)	161 (75.9%)	135 (63.7%)	0 (0%)	136 (64.2%)
Cytotoxic T cells (CD8+)	5	0 (0%)	0 (0%)	4 (80%)	1 (20%)	0 (0%)	1 (20%)
Total	800	155 (19.4%)	1 (0.1%)	511 (63.9%)	414 (51.7%)	5 (0.6%)	352 (44%)

single cell types. Moreover, even fewer studies are focusing on isomiR expression in the aforementioned compounds. Nearly all studies, including our own, report differentially expressed miRNAs characteristic for a certain condition or a disease, but in most of the cases it is not further investigated what are the variations in their sequences, and which tissues or cell types contribute to the expression pattern. The herein presented detailed miRNA expression catalogue for different cellular and non-cellular components of healthy human peripheral blood enables the researcher to assign expression-signatures of single or groups of miRNAs and isomiRs to cell types and, thus, gather hypotheses about putative functionality of the respective miRNA (or its predicted target genes) in the study context. While previous reports on miRNAs in sorted human blood cells are mostly based on very few samples ($n \leq 5$, see below) or focused on developmental aspects such as hematopoiesis (57), a comprehensive miRNA expression catalogue for a broad range of mature blood cell types currently only exist for mice (58). Moreover, all of the mentioned studies were focusing only on archetype miRNAs and did not explore the sequence variations and modifications of mature miRNAs.

Given our data, we were able to detect a substantial number of miRNAs outside transcriptionally active blood compounds. We identified 271 known miRNA species within RBCs. The results of a paper by Doss *et al.* (27) are nicely mirrored in our RBC profiles, although we are able to expand the catalogue of highly abundant miRNAs in RBCs with the most abundant miRNAs being present with median read counts >1000 reads per sample.

Analysis of miRNA sequence heterogeneity showed that a vast majority of mature miRNAs, which were found to be expressed in the peripheral blood compounds, are carrying different and often composite sequence modifications. Within our dataset, we identified sequence variations in 69.0% of detected miRNAs. As already indicated by previous studies in various human tissues, 3′-end modifications were the predominant category (10,13,59–61).

Comparing the distributions of different types of miRNA modifications across the studied blood compounds, we found that 3′ trimming modifications were distributed in a lineage-specific manner – the closer the hematopoietic progenitors are, the higher are their similarities in the preference of 3′ trimming modification. This phenomenon was not observed in 5′ trimming distributions of isomiRs sequences, where the frequencies of modification were distributed similarly across the cell types, except for neutrophils (CD15+) and RBCs (CD235a+) in which the

isomiRs having earlier trimmed nucleotide were more frequent than in another cell types. Similar findings of 5′ trimming preference were observed in serum and exosomes. These results support the assumption that, in contrast to isomiRs with 3′ variability, 5′ isomiRs because of seed-shifting have effects on target selection and the choice of miRNA-5p/-3p strand for AGO binding (12,18), and possibly due to that, the intra-group variability of 5′ sequence variation is lower than in the 3′ end.

One particularly interesting finding is that the overall distribution of nt-substitution modification within isomiR sequences showed a consistent pattern of positions in which the nucleotides are less frequently substituted, thus, highlighting the nucleotides at positions 2–8 (seed site) and 15–19, respectively, which are known to be important for target site recognition (62,63). All together this explains the reduced heterogeneity at those sites.

Another surprising observation relates to our finding that compositions of 3′ added non-templated nucleotides in the blood compounds are also lineage-specific. The frequencies of non-templated nucleotides clustered into three main groups, corresponding to myeloid, lymphoid and non-cellular compartments (serum and exosomes). Non-templated additions of isomiRs found in whole blood clustered together with RBCs, which is meaningful because RBCs are the most abundant cells in blood. It is worth pointing out that 3′ end uridylated isomiRs are relatively more common in serum and exosomes than in cells—this finding is consistent with results by Koppers-Lalic *et al.*, where they showed that 3′ end adenylated miRNAs are relatively enriched in B-lymphocytes, whereas 3′ end uridylated isoforms appear overrepresented in exosomes (30). Another study by Gutiérrez-Vázquez *et al.* demonstrated that 3′ addition of uridine promotes degradation of these uridylated miRNAs after T-lymphocytes activation (31). Moreover, it has been reported that non-templated nucleotide additions can affect miRNA stability, target identification and targeting power (30,60).

Several recent studies already reported miRNA expression data on single blood cell types, such as natural killer cells (64), B-lymphocytes (65), T-lymphocytes (29) and RBCs (27). Two other projects screened the miRNA-transcriptome using microarray technology in several blood cell types (26,28). Comparing the results of both studies to our miRNA-arm expression data, the previously reported cell type specific miRNA signatures could only be partially confirmed, as many miRNAs previously reported to be cell type specific are actually found in other blood cell types

in our data. All of six cell type specific miRNAs from Allantaz *et al.* were confirmed within our dataset (26). The miRNA profile of monocytes and B-cells reported by Leidinger *et al.* were also completely verified by our findings. The other profiles (T-cells, NK-cells and granulocytes), however, contained only about half of cell type specific miRNAs that could be confirmed in our data. These differences most probably arise from distinct miRNA profiling platforms (66) or even different blood cell purification techniques (67). In addition, our data complements the already known miRNA profiles with numerous other miRNAs, previously unknown to be specific for certain blood cell types. The same holds true for studies investigating only a single blood cell type. Although these reports do not compare different cellular blood compounds, but rather cells in activated and resting states, thus, they do not report cell type specific miRNAs; however, a considerable fraction of the named miRNAs is reflected by our cell type specific profiles. As nearly one half of miRNAs that have previously been reported to be cell type specific are actually present in more than one cell type in our data, the best explanation for these discrepancies is the sample size and profiling platform. All recent studies were conducted on $n \leq 5$ samples using microarrays compared to $n > 40$ samples using NGS in our case, which increases the statistical power substantially and overcomes possible inter-individual variability.

Finally, we determined cell type specific isomiRs of the human peripheral blood cells. In comparison to miRNA-arms, we found a substantially higher number of cell type specific isomiRs. The modification types of these molecules were commonly distributed as in the general isomiR population. An interesting observation is that almost a half of cell type specific isomiRs were arising from miRNA-arms which previously, in our dataset, were not assigned as cell type specific, showing that a big proportion of molecules is diluted when looking only at the miRNA-arm level. However, up to now there are no reports referring to differential expression of isomiRs within different blood cells, thus making it impossible to compare and draw conclusions about these results.

Taken together, we do not claim that our results are exhaustive or cover the entire miRNA-repertoire of healthy human blood, but they provide the most comprehensive contribution to date (with $n > 40$ individuals) toward a complete miRNA inventory of human peripheral blood. Upcoming reports from bigger research initiatives such as the international human epigenome program (68) will hopefully amend the data with additional cell types and more exhaustive data on miRNA and especially, on isomiRs in circulation and exosomes. Moreover, recent technological developments enabling analyses on the single-cell level will—as soon as they are applicable for miRNA analysis—likely also make a significant contribution towards completing the miRNA map of human blood.

AVAILABILITY

The web-tool is available at http://134.245.63.235/ikmb-tools/bloodmiRs.

ACCESSION NUMBERS

Raw sequencing reads and quantified read-count data have been deposited at NCBI Gene Expression Omnibus (GEO) (36) under the accession number GSE100467.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

REFERENCES

1. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
2. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–D73.
3. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
4. Friedman,R.C., Farh,K.K.-H., Burge,C.B. and Bartel,D.P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.
5. Iorio,M. V., Ferracin,M., Liu,C.-G., Veronese,A., Spizzo,R., Sabbioni,S., Magri,E., Pedriali,M., Fabbri,M., Campiglio,M. *et al.* (2005) MicroRNA gene expression deregulation in human breast cancer. *Cancer Res.*, **65**, 7065–7070.
6. Lu,J., Getz,G., Miska,E.A., Alvarez-Saavedra,E., Lamb,J., Peck,D., Sweet-Cordero,A., Ebert,B.L., Mak,R.H., Ferrando,A.A. *et al.* (2005) MicroRNA expression profiles classify human cancers. *Nature*, **435**, 834–838.
7. Calin,G.A. and Croce,C.M. (2006) MicroRNA signatures in human cancers. *Nat. Rev. Cancer*, **6**, 857–866.
8. Ma,L., Teruya-Feldstein,J. and Weinberg,R.A. (2007) Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. *Nature*, **449**, 682–688.
9. Ikeda,S., Kong,S.W., Lu,J., Bisping,E., Zhang,H., Allen,P.D., Golub,T.R., Pieske,B. and Pu,W.T. (2007) Altered microRNA expression in human heart disease. *Physiol. Genomics*, **31**, 367–373.
10. Lee,L.W., Zhang,S., Etheridge,A., Ma,L., Martin,D., Galas,D. and Wang,K. (2010) Complexity of the microRNA repertoire revealed by next-generation sequencing. *RNA*, **16**, 2170–2180.
11. Telonis,A.G., Magee,R., Loher,P., Chervoneva,I., Londin,E. and Rigoutsos,I. (2017) Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types. *Nucleic Acids Res.*, **45**, 2973–2985.
12. Neilsen,C.T., Goodall,G.J. and Bracken,C.P. (2012) IsomiRs—the overlooked repertoire in the dynamic microRNAome. *Trends Genet.*, **28**, 544–549.
13. Loher,P., Londin,E.R. and Rigoutsos,I. (2014) IsomiR expression profiles in human lymphoblastoid cell lines exhibit population and gender dependencies. *Oncotarget*, **5**, 8790–8802.
14. Telonis,A.G., Magee,R., Loher,P., Chervoneva,I., Londin,E. and Rigoutsos,I. (2017) Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types. *Nucleic Acids Res.*, **6**, 20707.
15. Telonis,A.G., Loher,P., Jing,Y., Londin,E. and Rigoutsos,I. (2015) Beyond the one-locus-one-miRNA paradigm: microRNA isoforms enable deeper insights into breast cancer heterogeneity. *Nucleic Acids Res.*, **43**, 9158–9175.
16. Warnefors,M., Liechti,A., Halbert,J., Valloton,D. and Kaessmann,H. (2014) Conserved microRNA editing in mammalian evolution, development and disease. *Genome Biol.*, **15**, R83.

17. Llorens,F., Bañez-Coronel,M., Pantano,L., del Río,J.A., Ferrer,I., Estivill,X. and Martí,E. (2013) A highly expressed miR-101 isomiR is a functional silencing small RNA. *BMC Genomics*, **14**, 104.

18. Tan,G.C., Chan,E., Molnar,A., Sarkar,R., Alexieva,D., Isa,I.M., Robinson,S., Zhang,S., Ellis,P., Langford,C.F. *et al.* (2014) 5′ isomiR variation is of functional and evolutionary importance. *Nucleic Acids Res.*, **42**, 9424–9435.

19. Mitchell,P.S., Parkin,R.K., Kroh,E.M., Fritz,B.R., Wyman,S.K., Pogosova-Agadjanyan,E.L., Peterson,A., Noteboom,J., O'Briant,K.C., Allen,A. *et al.* (2008) Circulating microRNAs as stable blood-based markers for cancer detection. *Proc. Natl. Acad. Sci.*, **105**, 10513–10518.

20. Lawrie,C.H., Gal,S., Dunlop,H.M., Pushkaran,B., Liggins,A.P., Pulford,K., Banham,A.H., Pezzella,F., Boultwood,J., Wainscoat,J.S. *et al.* (2008) Detection of elevated levels of tumour-associated microRNAs in serum of patients with diffuse large B-cell lymphoma. *Br. J. Haematol.*, **141**, 672–675.

21. Chen,X., Ba,Y., Ma,L., Cai,X., Yin,Y., Wang,K., Guo,J., Zhang,Y., Chen,J., Guo,X. *et al.* (2008) Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. *Cell Res.*, **18**, 997–1006.

22. Brase,J.C., Wuttig,D., Kuner,R. and Sültmann,H. (2010) Serum microRNAs as non-invasive biomarkers for cancer. *Mol. Cancer*, **9**, 306.

23. McManus,D.D. and Ambros,V. (2011) Circulating microRNAs in cardiovascular disease. *Circulation*, **124**, 1908–1910.

24. Sita-Lumsden,A., Dart,D.A., Waxman,J. and Bevan,C.L. (2013) Circulating microRNAs as potential new biomarkers for prostate cancer. *Br. J. Cancer*, **108**, 1925–1930.

25. Pritchard,C.C., Kroh,E., Wood,B., Arroyo,J.D., Dougherty,K.J., Miyaji,M.M., Tait,J.F. and Tewari,M. (2012) Blood cell origin of circulating microRNAs: a cautionary note for cancer biomarker studies. *Cancer Prev. Res.*, **5**, 492–497.

26. Allantaz,F., Cheng,D.T., Bergauer,T., Ravindran,P., Rossier,M.F., Ebeling,M., Badi,L., Reis,B., Bitter,H., D'Asaro,M. *et al.* (2012) Expression profiling of human immune cell subsets identifies miRNA-mRNA regulatory relationships correlated with cell type specific expression. *PLoS One*, **7**, e29979.

27. Doss,J.F., Corcoran,D.L., Jima,D.D., Telen,M.J., Dave,S.S. and Chi,J.-T. (2015) A comprehensive joint analysis of the long and short RNA transcriptomes of human erythrocytes. *BMC Genomics*, **16**, 952.

28. Leidinger,P., Backes,C., Meder,B., Meese,E. and Keller,A. (2014) The human miRNA repertoire of different blood compounds. *BMC Genomics*, **15**, 474.

29. Mitchell,C.J., Getnet,D., Kim,M.-S., Manda,S.S., Kumar,P., Huang,T.-C., Pinto,S.M., Nirujogi,R.S., Iwasaki,M., Shaw,P.G. *et al.* (2015) A multi-omic analysis of human naïve CD4+ T cells. *BMC Syst. Biol.*, **9**, 75.

30. Koppers-Lalic,D., Hackenberg,M., Bijnsdorp,I. V., van Eijndhoven,M.A.J., Sadek,P., Sie,D., Zini,N., Middeldorp,J.M., Ylstra,B., de Menezes,R.X. *et al.* (2014) Nontemplated nucleotide additions distinguish the small RNA composition in cells from exosomes. *Cell Rep.*, **8**, 1649–1658.

31. Gutiérrez-Vázquez,C., Enright,A.J., Rodríguez-Galán,A., Pérez-García,A., Collier,P., Jones,M.R., Benes,V., Mizgerd,J.P., Mittelbrunn,M., Ramiro,A.R. *et al.* (2017) 3′ Uridylation controls mature microRNA turnover during CD4 T-cell activation. *RNA*, **23**, 882–891.

32. Nair,V.S., Pritchard,C.C., Tewari,M. and Ioannidis,J.P.A. (2014) Design and analysis for studying microRNAs in human disease: a primer on -Omic technologies. *Am. J. Epidemiol.*, **180**, 140–152.

33. Pritchard,C.C., Cheng,H.H. and Tewari,M. (2012) MicroRNA profiling: approaches and considerations. *Nat. Rev. Genet.*, **13**, 358–369.

34. Keller,A. and Meese,E. (2016) Can circulating miRNAs live up to the promise of being minimal invasive biomarkers in clinical settings? *Wiley Interdiscip. Rev. RNA*, **7**, 148–156.

35. Hafner,M., Renwick,N., Farazi,T.A., Mihailović,A., Pena,J.T.G. and Tuschl,T. (2012) Barcoded cDNA library preparation for small RNA profiling by next-generation sequencing. *Methods*, **58**, 164–170.

36. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

37. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10.

38. Pantano,L., Estivill,X. and Marti,E. (2010) SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. *Nucleic Acids Res.*, **38**, e34.

39. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

40. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

41. R Core Team (2017) R: A Language and Environment for Statistical Computing.

42. Wickham,H. (2009) ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, NY.

43. Gu,Z., Eils,R. and Schlesner,M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.

44. Kang,W. and Friedländer,M.R. (2015) Computational prediction of miRNA genes from small RNA sequencing data. *Front. Bioeng. Biotechnol.*, **3**, 7.

45. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.

46. Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.

47. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

48. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

49. Friedländer,M.R., Mackowiak,S.D., Li,N., Chen,W. and Rajewsky,N. (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.*, **40**, 37–52.

50. Cunningham,F., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2014) Ensembl 2015. *Nucleic Acids Res.*, **43**, D662–D669.

51. Wheeler,T.J., Clements,J., Eddy,S.R., Hubley,R., Jones,T.A., Jurka,J., Smit,A.F.A. and Finn,R.D. (2013) Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res.*, **41**, D70–D82.

52. Backes,C., Meder,B., Hart,M., Ludwig,N., Leidinger,P., Vogel,B., Galata,V., Roth,P., Menegatti,J., Grässer,F. *et al.* (2016) Prioritizing and selecting likely novel miRNAs from NGS data. *Nucleic Acids Res.*, **44**, e53.

53. Chang,W., Cheng,J., Allaire,J., Xie,Y. and McPherson,J. (2017) shiny: Web Application Framework for R.

54. Wickham,H. and Romain,F. (2016) dplyr: A Grammar of Data Manipulation.

55. Metpally,R.P.R., Nasser,S., Malenica,I., Courtright,A., Carlson,E., Ghaffari,L., Villa,S., Tembe,W. and Van Keuren-Jensen,K. (2013) Comparison of analysis tools for miRNA high throughput sequencing using nerve crush as a model. *Front. Genet.*, **4**, 20.

56. Fehlmann,T., Backes,C., Kahraman,M., Haas,J., Ludwig,N., Posch,A.E., Würstle,M.L., Hübenthal,M., Franke,A., Meder,B. *et al.* (2017) Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs. *Nucleic Acids Res.*, **13**, 1084–1088.

57. Chen,L., Kostadima,M., Martens,J.H.A., Canu,G., Garcia,S.P., Turro,E., Downes,K., Macaulay,I.C., Bielczyk-Maczynska,E., Coe,S. *et al.* (2014) Transcriptional diversity during lineage commitment of human blood progenitors. *Science*, **345**, 1251033.

58. Petriv,O.I., Kuchenbauer,F., Delaney,A.D., Lecault,V., White,A., Kent,D., Marmolejo,L., Heuser,M., Berg,T., Copley,M. *et al.* (2010) Comprehensive microRNA expression profiling of the hematopoietic hierarchy. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 15443–15448.

59. Karali,M., Persico,M., Mutarelli,M., Carissimo,A., Pizzo,M., Singh Marwah,V., Ambrosio,C., Pinelli,M., Carrella,D., Ferrari,S. *et al.* (2016) High-resolution analysis of the human retina miRNome

reveals isomiR variations and novel microRNAs. *Nucleic Acids Res.*, **44**, 1525–1540.

60. Burroughs,A.M., Ando,Y., de Hoon,M.J.L., Tomaru,Y., Nishibu,T., Ukekawa,R., Funakoshi,T., Kurokawa,T., Suzuki,H., Hayashizaki,Y. *et al.* (2010) A comprehensive survey of 3′ animal miRNA modification events and a possible role for 3′ adenylation in modulating miRNA targeting effectiveness. *Genome Res.*, **20**, 1398–1410.

61. Gyvyte,U., Juzenas,S., Salteniene,V., Kupcinskas,J., Poskiene,L., Kucinskas,L., Jarmalaite,S., Stuopelyte,K., Steponaitiene,R., Hemmrich-Stanisak,G. *et al.* (2017) MiRNA profiling of gastrointestinal stromal tumors by next generation sequencing. *Oncotarget*, **5**, 37225–37238.

62. Kim,D., Sung,Y.M., Park,J., Kim,S., Kim,J., Park,J., Ha,H., Bae,J.Y., Kim,S. and Baek,D. (2016) General rules for functional microRNA targeting. *Nat. Genet.*, **48**, 1517–1526.

63. Khorshid,M., Hausser,J., Zavolan,M. and van Nimwegen,E. (2013) A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nat. Methods*, **10**, 253–255.

64. Ni,F., Guo,C., Sun,R., Fu,B., Yang,Y., Wu,L., Ren,S., Tian,Z. and Wei,H. (2015) MicroRNA transcriptomes of distinct human NK cell populations identify miR-362-5p as an essential regulator of NK cell function. *Sci. Rep.*, **5**, 9993.

65. Jima,D.D., Zhang,J., Jacobs,C., Richards,K.L., Dunphy,C.H., Choi,W.W.L., Au,W.Y., Srivastava,G., Czader,M.B., Rizzieri,D.A. *et al.* (2010) Deep sequencing of the small RNA transcriptome of normal and malignant human B cells identifies hundreds of novel microRNAs. *Blood*, **116**, e118–e127.

66. Mestdagh,P., Hartmann,N., Baeriswyl,L., Andreasen,D., Bernard,N., Chen,C., Cheo,D., D'Andrade,P., DeMayo,M., Dennis,L. *et al.* (2014) Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. *Nat. Methods*, **11**, 809–815.

67. Schwarz,E.C., Backes,C., Knörck,A., Ludwig,N., Leidinger,P., Hoxha,C., Schwär,G., Grossmann,T., Müller,S.C., Hart,M. *et al.* (2016) Deep characterization of blood cell miRNomes by NGS. *Cell. Mol. Life Sci.*, **73**, 3169–3181.

68. Ebert,P. and Bock,C. (2015) Improving reference epigenome catalogs by computational prediction. *Nat. Biotechnol.*, **33**, 354–355.

69. Häggström,M. Simplified hematopoiesis. *Wikimedia Commons*.
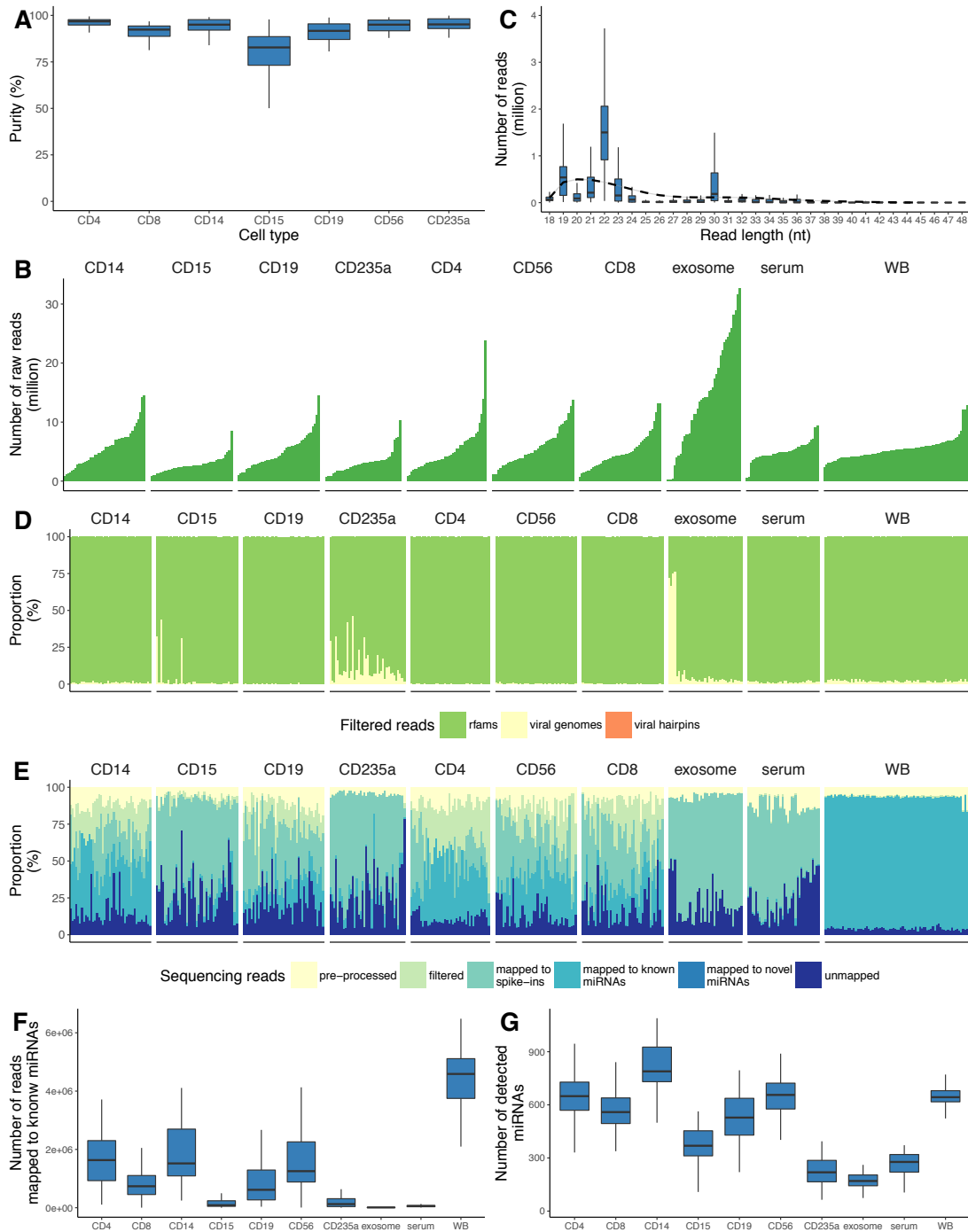
Supplementary Figure S2.9: Overview of the small RNA transcriptome data. A) The purity of blood cell populations after sorting by MACS; **B)** Raw reads per sample within different blood compounds; **C)** Read length distribution of pre-filtered sequencing reads; **D)** Composition of filtered reads within different blood compounds. The reads were filtered against non-miRNA small RNAs from Rfam (green), viral genomes from RefSeq (yellow) and viral precursors from miRBase (orange); **E)** Overall composition of the initial reads per sample in distinct blood compounds. Due to a small fraction of novel miRNAs, their proportions within most of the samples are not visible in graphs; **F)** Number of reads mapped to known miRNAs per distinct blood compound; G) Number of mature miRNAs detected per individual blood compound. **Note:** Samples in graphs B, D and E are aligned in the same order.

Supplementary Figure S2.10: Heatmap representing the expression levels of blood cell lineage-specific isomiRs. Each row of the heatmap corresponds to the one of the samples per blood cell types, and each column corresponds to an isomiR. All molecules were significantly up-regulated (FDR p-value < 0.001; |log2FC| > 1) exclusively in one of the cell types. The unsupervised hierarchical clustering of isomiRs was performed using Spearman's correlation distance as metric and average linkage clustering as linkage criterion.
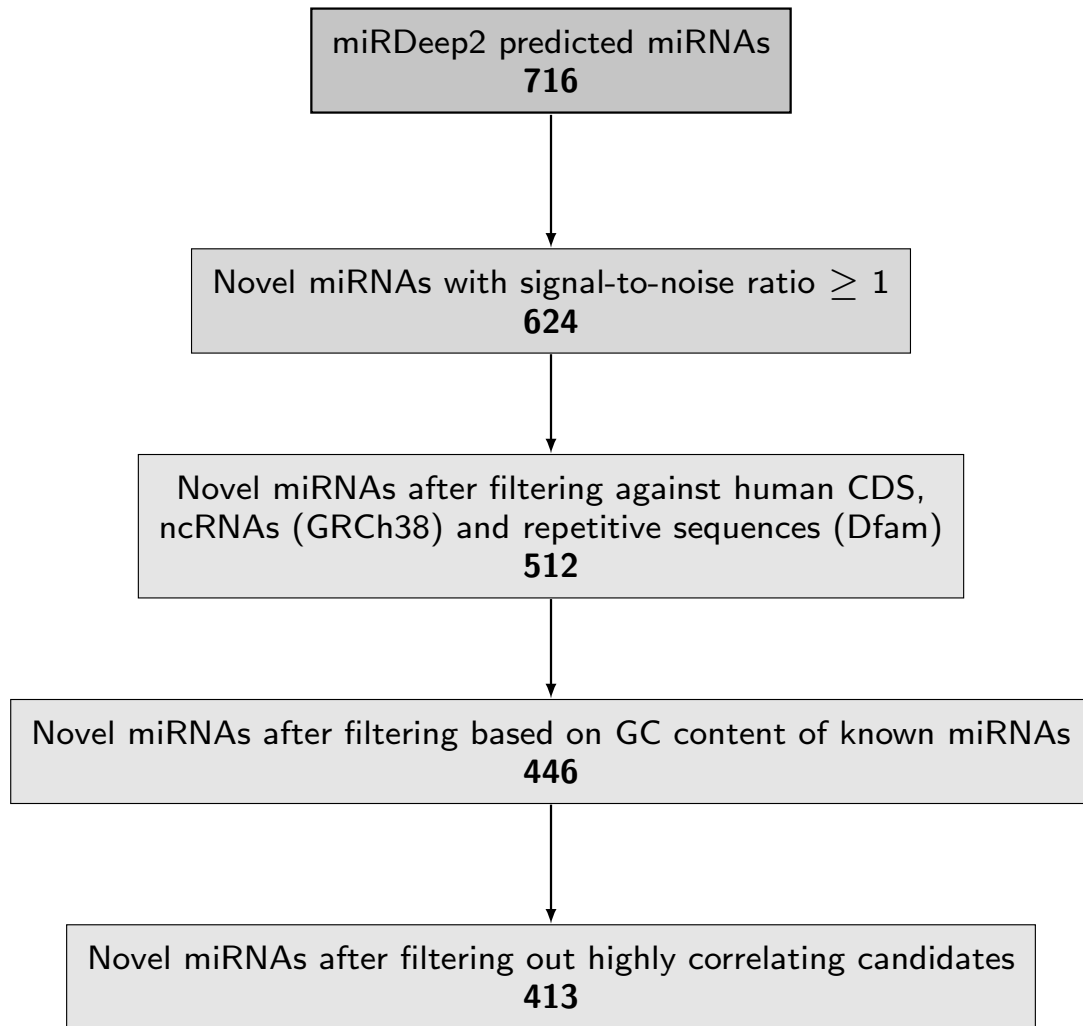
Supplementary Figure S2.11: Filtering pipeline of novel miRNA candidates.
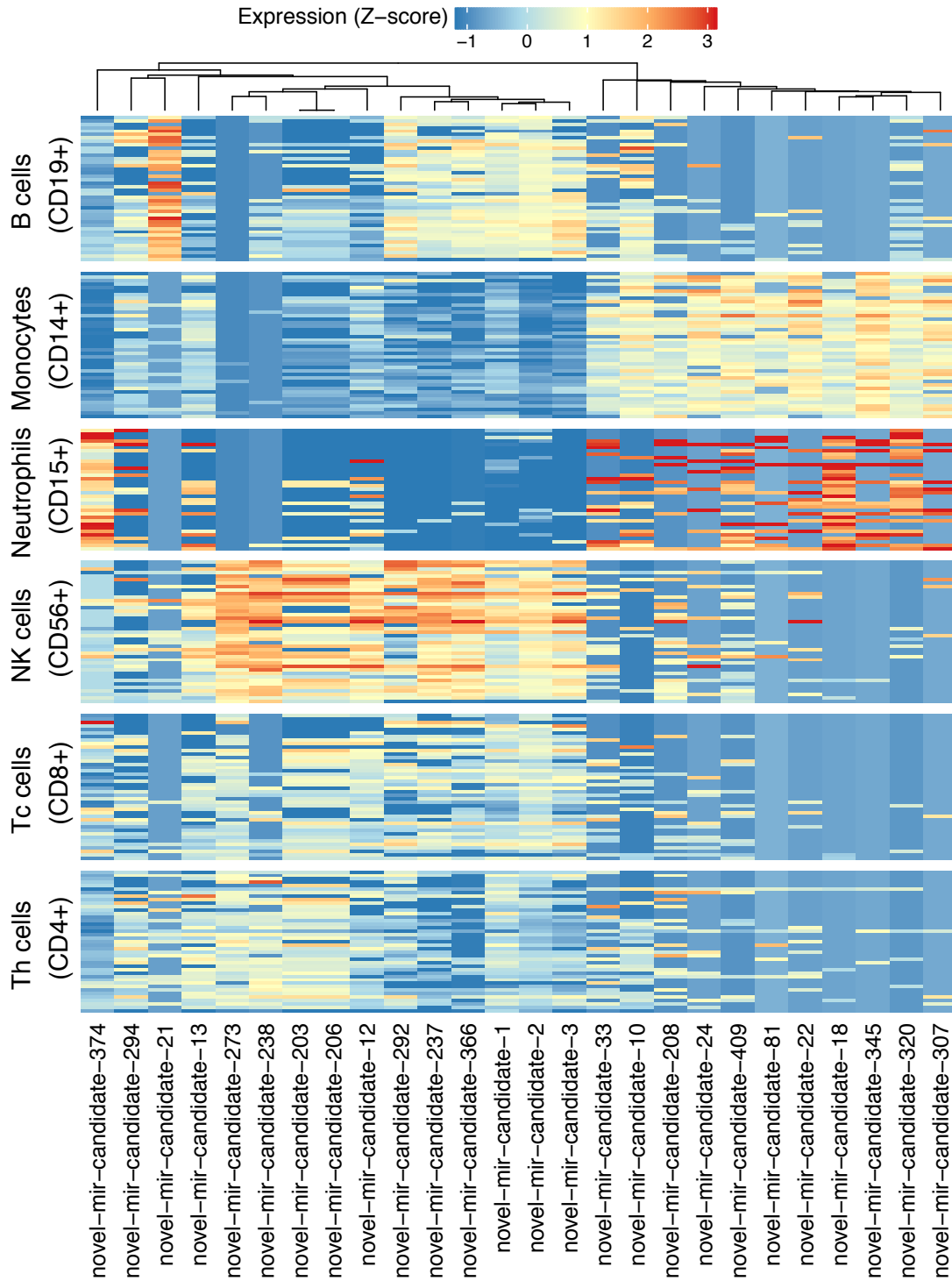
Supplementary Figure S2.12: Expression of blood cell specific novel miRNA candidates in different blood cells. In the heatmap, rows represent samples and columns represent novel miRNA candidate transcripts which are exclusively significantly up-regulated (FDR p-value < 0.001; |log2FC| > 1) in one of the cell types. The Z-score represents standardized normalized expression values. The unsupervised hierarchical clustering was performed using Spearman's correlation distance as metric and average linkage clustering as linkage criterion.

3 Discussion

The diagnosis of inflammatory bowel disease still remains a clinical challenge and the most accurate diagnostic procedure involves a combination of clinical, endoscopic, histological and radiological parameters. At present, neither genetic nor serological testing is recommended for routine clinical practice. However, a multitude of biomarkers has been proposed to complement diagnostics, particularly in case of diagnostic uncertainty. In the scope of this work we evaluated whether systematic miRNA or miRNA variant expression profiling, in conjunction with state-of-the-art machine learning techniques, is suitable as a non-invasive tool for diagnostics of IBD.

Based on microarray technology, expression levels of 863 miRNAs have been determined for whole blood samples from patients (40 CD and 36 UC, repectively) as well as healthy and inflammatory controls (108 HC and 130 IC, repectively; see Article II). Subsequently, we identified comprehensive profiles of miRNA expression patterns being distinctive for each of the traits under consideration. Then we conducted mathematical modeling to generate computer-based tools utilizing these profiles for diagnostic phenotype prediction. In this way we were able to reveal highly accurate classification models that distinguish CD and UC among each other as well as from different types of controls. Relying on a minimal set of not more than 16 miRNAs being sufficient for sensitive and specific classification, these models hold great promises and should be further evaluated in independent sample panels.

We employed support vector machines (SVMs) to solve the given diagnostic problems. Regularized instances of these models, such as the elastic SCAD SVM, incorporate penalties for model complexity to prevent overfitting and to provide sparse solutions. In the scope of the here-presented study this property has been used to obtain small sets of miRNAs being suitable for diagnostic application. It is expected that miRNAs being informative for a particular classification problem are likewise selected when utilizing another modeling strategy. In this work the miRNA signature selected using the elastic SCAD SVM has been confirmed by comparably high classification performance of random forests as an independent classification approach. To obtain more even comparable results, future studies might consider regularized random forests as proposed by Deng and Runger [1].

To obtain a model applicable with high accuracy to independent data, for each of the given diagnostic problems we chose the sparsest median performing elastic SCAD SVM along with the corresponding miRNA signature. Both, the regularization approach and the comprehensive holdout sampling decrease the model's probability of being overfitted to the dataset generated for this study. However, due to correlating expression profiles it is expected that models with matching accuracy potentially incorporate differing miRNAs. For the same reason more complex signatures may exist which merely incorporate additional highly correlated miRNAs.

To derive hypothetical biological functions of the here-presented miRNA signatures, we screened current databases for experimentally validated miRNA-target interactions. Notably, a considerable fraction of the target genes has been implicated in intestinal diseases. Many of these targets have been identified in recent IBD GWAS. However, most of the genetic variation detected does not correlate (and thus not interfere) with miRNA regulatory binding sites. Solely binding sites of hsa-mir-99b, located within the 3'-UTRs of RAVER2, represent good candidates for further experimental investigation. More complete data on genetic variation, particularly within the 3'-UTRs of IBD-related genes, is needed for more comprehensive analyses

of miRNA target genes. In a recent article reviewing genetic studies in IBD, Liu and Anderson conclude that most of the identified GWAS loci reside in non-coding regions of the human genome [2]. A vast number of these non-coding variants is assumed to affect gene regulation, which miRNAs play an crucial role in. Future studies on the here-presented miRNA signatures might therefore give valuable insights into the etiology of inflammatory bowel disease.

Relying on predefined annealing probes, microarray technology solely allows the detection of annotated miRNAs and isomiRs, respectively. Next generation sequencing (NGS) technology, instead, enables the quantification of the entire set of isomiRs being expressed in a given sample. With our biomarker study utilizing NGS we provide the first discriminative analysis of global sequencing-based isomiR expression in the context of IBD (see Article III). We examined isomiR expression profiles of 515 individuals, including 75 cases of untreated IBD, 271 cases of treated IBD, 124 healthy (HC) and 45 symptomatic controls (SC). Based on the resulting data, we identified isomiRs that exhibit differential expression and likewise allow accurate distinction between the investigated traits. We report standard SVMs to be able to predict disease phenotypes with a remarkably high accuracy. Penalized SVMs showed comparably lower performance and/or stability and were therefore investigated in less detail. Due to their sparsity, however, they might provide a foundation for further functional research.

Previously being dismissed as experimental artifacts, isomiRs are now considered to result from alterations in miRNA biogenesis. Recent studies show these alterations to vary with respect to gender and ethnicity as well as tissue and cell type. With our NGS-based study we ultimately provide evidence for the existence of blood-born isomiRs being specific for IBD-related traits. We furthermore show altered isomiR expression being dependent on the patients' treatment regimen. All comparisons between the treatment groups have been conducted with respect to healthy controls. This approach enables the identification of treatment-induced differences specific for a given disease. In contrast, differences being treatment-induced but unrelated to the diseases, have not been considered to ensure the neglection of unwanted biological and technical variation. Accordingly, reliability of the results has been increased at the cost of overall comprehensiveness of the study. However, future investigations might consider them to elucidate the role of isomiRs with respect to treatment response.

Current research on circulating biomarkers usually focuses on miRNAs in either whole blood, plasma or serum, while only a small number of studies investigates hematopoietic cells. Moreover, even fewer studies are investigating isomiR expression in these particular cells. As part of this work we present the currently most comprehensive miRNA expression catalogue of cellular (NK cells, B lymphocytes, cytotoxic T lymphocytes, T helper cells, monocytes, neutrophils and erythrocytes) and non-cellular fractions of healthy human peripheral blood (serum, exosomes; see Article IV). Therewith, we provide a foundation for decomposing blood-born miRNA expression patterns into its cell-type-specific components.

Moreover, the availability of this unique resource facilitated comprehensive analysis of the cell-type-specific features of hematopoietic miRNA landscape. We identified sets of miRNAs being specifically DE in each of the investigated blood compounds. Notably, only subsets of these miRNA signatures could replicate previous findings. They rather complement reported signatures by adding miRNAs, previously unknown to be specific for a certain blood cell type. The same holds true for earlier studies investigating only a single blood cell type. Although these reports do not compare different cellular blood compounds and therefore do not investigate cell-type-specific expression, a considerable fraction of the named miRNAs is reflected by our cell-type-specific profiles. Discrepancies between all the studies, however, can probably be explained by varying miRNA profiling platforms [3] as well as varying blood cell purification techniques [4]. Furthermore, they mirror differences in size and homogeneity of the investigated study samples. Completing our catalogue of expression profiles, we identified 271, 90 and 51 miRNA species in RBCs, serum and exosomes, respectively. With these findings we finally confirm previous reports on miRNAs being detectable

outside the transcriptionally active components of human blood [5, 6].

Cell-type-specific signatures established on the basis of isomiRs instead of archetype miRNAs turn out to be significantly larger in size. This suggests that considering sequence variation adds significantly to the distinguishability of cell-type-specific miRNA profiles. However, the description of isomiRs within different blood cells represents a unique feature of our study. We detected sequence modifications in 69.0% of the miRNA variants. A combination of more than one type of sequence variation has been observed in 74.0% of the isomiRs. We identified non-templated 3' additions (observed in 77.0% of isomiR sequences) to be distributed in a lineage-specific manner. Thus, the frequency distribution of added nucleotides clustered into three main groups, corresponding to myeloid, lymphoid and non-cellular compartments. Non-templated additions detected in RBCs, in turn, formed an additional cluster with whole blood. The observed frequencies are in line with previous reports on nucleotide distribution in B and T lymphocytes [7, 8]. It has been reported that non-templated nucleotide additions modulate miRNA stability as well as targetin efficiency. According to the clustering, these modulations likewise are expected to show a lineage-specific pattern [7, 9]. Additional lineage specificity we discovered for the frequencies of 3' trimming modifications (detectable in 75.6% of isomiR sequences). Particularly, the fraction of unique sequences at positions -3 and -2 has been estimated to be higher in lymphoid than in myeloid cells. The fraction at positions -1 to 3, in turn, has been estimated to be higher in myeloid than in lymphoid cells. Elucidating the functional significance of these observations needs further investigation. One particularly interesting finding relates to the overall distribution of nt-substitution (detectable in 6.2% of isomiR sequences). We observe decreased frequencies of this modification type at positions 2–8 and 15–19, respectively, known to play a crucial role in the recognition of miRNA target sites [10, 11].

Certainly, the clinically most relevant objectives of this work include the accurate distinction of the major IBD phenotypes among each other as well as from healthy controls. Employing microarray-based miRNA expression profiling we estimated these tasks to be solvable with median balanced accuracies of up to 91.70% (CD vs. UC) and 98.10% (CD or UC vs. HC), respectively (see Article II). For sequencing-based isomiR expression profiling, we generated solutions performing with median balanced accuracies of 78.57%/78.83% (untreated/treated CD vs. UC) and 100.00%/98.28% (untreated/treated CD or UC vs. HC), respectively (see Article III). Hence, performance estimates for the distinction from healthy controls by far exceed those available for established biomarkers (ACCA, ASCA, ALCA, AMCA, anti-C and anti-L). The same holds true for our performance estimates for the distinction of CD from UC being compared to established biomarkers (ACCA, ALCA, AMCA, anti-C, anti-L, anti-OmpC and pANCA). Remarkably, they deceed the performance of ASCA or ASCA+ANCA. However, in contrast to our findings, the estimates provided by literature about these biomarkers exhibit substantial instability (balanced accuracy ranging from 59.50 to 86.00%, see Article I). Notably, conducting a valid comparison of performance estimates assumes the availability of a representative study cohort of asymptotic size for which every biomarker of interest has been evaluated. The lack of such a dataset introduces uncertainty while classifying the clinical value of the here-presented candidates. Accordingly, a conclusive assessment needs further investigations.

Pairwise distinction of the traits under consideration has been implemented utilizing different types of support vector machines (SVMs). Aiming for sparse solutions incorporating miRNA signatures of minimal size, the microarray-based study (Article II) employed penalized models, such as ridge, LASSO, elastic net, SCAD and elastic SCAD penalized SVMs. Allowing for more comprehensive signatures, the sequencing-based study (Article III) then focused on unpenalized models, including linear, radial and polynomial SVMs. We provide evidence for SVMs being a valuable tool for non-invasive prediction of inflammatory phenotypes, irrespective of the penalty and kernel function being used. However, the most accurate models are not sparse.

Unlike earlier studies the here-presented investigations have not been restricted to binomial classification

problems (see Article I). We rather evaluated joint discrimination of arbitrary combinations of the given traits (Article II: CD, UC, IC and HC; Article III: CD, UC, SC and HC). The predictive performance of the resulting multinomial classification models thereby has been estimated to be remarkably high, for both the microarray- as well as the sequencing-based analysis. In favor of generalizability, we only considered the main phenotypes. However, extensions to multinomial models further stratifying with respect to clinical parameters, such as disease location and disease activity, do not represent a technical challenge.

Cortes and Vapnik introduced SVMs as binomial classifiers [12]. Other groups introduced formulations of the model that allow multinomial classification problems to be solved in a single constrained optimization [13, 14]. However, we employed the classical strategy of reducing single multinomial classification problems to sets of binomial classification problems, providing the basis for subsequent majority voting [15]. This method features numerous favorable properties. Thus, it enables the reuse of single binomial models for the construction of a multitude of multinomial models. Furthermore, it does not restrict the binomial models to be of a particular type. In the course of this work we combined SVMs employing different kernels. However, the chosen strategy allows a combination of models of arbitrary nature and thereby introduces the possibility for post-hoc improvement of here-presented phenotype predictions. Notably, our approach likewise allows an improvement by incorporation of unrelated explanatory variables. These might include genomic, epigenomic or metagenomic information.

The model-based approach to (multinomial) prediction of inflammatory phenotypes represents a unique feature of this work. Direct comparisons to earlier studies are aggravated for this reason. However, miRNA expression patterns being replicably observed in independent datasets are expected to allow for the derivation of models of comparable quality. In accordance to this assumption we rely on consistency of expression patterns in regard to previous work to evaluate the validity of our models. Thus, our microarray data confirms 45.0%, 71.4% and 75.0% of deregulated miRNAs being reported by Wu et al. [16], Duttagupta et al. [17] and Chen et al. [18]. Agreement of expression patterns observed by Zahm et al. [19] and Paraskevi et al. [20], in turn, amounts to 100%. The significance of our sequencing-based findings we evaluate in terms of overlapping sets of differentially expressed isomiRs and miRNA arms, respectively. Based on enrichment analysis we provide evidence for non-random overlaps between our study and earlier works, including Zahm et al. [19], Paraskevi et al. [20] and Schaefer et al. [21]. Employing the same methodology we confirm our biomarker studies with one another (see Article II and III).

Nearly all biomarker studies, including our own, employ varying abundance of miRNAs as an indicator for differential expression being associated with disease. However, accumulating evidence suggests those differences likewise being attributable to the cell type composition of the tissue under investigation. Whereas the extent of cell-type-specific miRNA expression varies depending on the tissue type, it certainly can not be neglected. First evidence for the blood cell composition contributing to the levels of miRNAs in circulation has been provided by Pritchard et al., questioning the validity of a substantial fraction of proposed cancer biomarkers [5]. Since then, only a few research projects elaborated on the miRNA transcriptomes of cellular components of human peripheral blood [6, 22–24]. Ultimately, our group presented a catalogue of miRNA/isomiR expression pattern being specific for blood cell lineages (lymphoid and myeloid) as well as individual blood cell types (CD4+, CD8+, CD14+, CD15+, CD19+, CD56+ and CD235a+) (see Article IV). Still, this resource represents the most comprehensive contribution toward a complete miRNA inventory of the blood. Along with cell count data it will facilitate the discrimination between expression changes being attributable to blood cell composition and disease. Models incorporating this additional layer of information might have increased clinical utility. However, further studies are needed to verify this hypothesis.

All the data presented throughout this thesis has been designed, sampled, reviewed and analyzed based on the premise of obtaining generalizable results. Depending on availability of the corresponding meta-data, known sources of technical and biological variability have been accounted for. However, unknown sources

of variability might exist. This holds true especially for inflammatory, auto-immune diseases as being investigated in the scope of this work, since intra- and inter-individual miRNA-expression variability might be altered by numerous comorbidities. Further research is likewise needed to investigate environmental factors affecting miRNA expression. Cohorts being characterized in even more detail on both, the data and the meta-data level, might enable the identification of these sources and alleviate translation to clinical practice.

References

[1] H. Deng and G. Runger. "Gene selection with guided regularized random forest". In: *Pattern Recognition* 46.12 (2013), pp. 3483–3489.

[2] J. Z. Liu and C. a. Anderson. "Genetic studies of Crohn's disease: past, present and future." In: *Best practice & research. Clinical gastroenterology* 28.3 (2014), pp. 373–86.

[3] P. Mestdagh, N. Hartmann, L. Baeriswyl, et al. "Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study." In: *Nature methods* 11.8 (2014), pp. 809–15.

[4] E. C. Schwarz, C. Backes, A. Knörck, et al. "Deep characterization of blood cell miRNomes by NGS." In: *Cellular and molecular life sciences : CMLS* 73.16 (2016), pp. 3169–81.

[5] C. C. Pritchard, E. Kroh, B. Wood, et al. "Blood cell origin of circulating microRNAs: a cautionary note for cancer biomarker studies." In: *Cancer prevention research* 5.3 (2012), pp. 492–7.

[6] J. F. Doss, D. L. Corcoran, D. D. Jima, et al. "A comprehensive joint analysis of the long and short RNA transcriptomes of human erythrocytes." In: *BMC genomics* 16 (2015), p. 952.

[7] D. Koppers-Lalic, M. Hackenberg, I. V. Bijnsdorp, et al. "Nontemplated nucleotide additions distinguish the small RNA composition in cells from exosomes." In: *Cell reports* 8.6 (2014), pp. 1649–1658.

[8] C. Gutiérrez-Vázquez, A. J. Enright, A. Rodríguez-Galán, et al. "3' Uridylation controls mature microRNA turnover during CD4 T-cell activation." In: *RNA* 23.6 (2017), pp. 882–891.

[9] A. M. Burroughs, Y. Ando, M. J. L. de Hoon, et al. "A comprehensive survey of 3' animal miRNA modification events and a possible role for 3' adenylation in modulating miRNA targeting effectiveness." In: *Genome research* 20.10 (2010), pp. 1398–410.

[10] D. Kim, Y. M. Sung, J. Park, et al. "General rules for functional microRNA targeting." In: *Nature genetics* 48.12 (2016), pp. 1517–1526.

[11] M. Khorshid, J. Hausser, M. Zavolan, et al. "A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets". In: *Nature Methods* 10.3 (2013), pp. 253–255.

[12] C. Cortes and V. Vapnik. "Support-Vector Networks". In: *Machine Learning* 20 (1995), pp. 273–297.

[13] J. Weston and C. Watkins. "Support Vector Machines for Multi-Class Pattern Recognition". In: *Proceedings of the European Symposium on Artificial Neural Networks* (1999), pp. 219–22.

[14] K. Crammer and Y. Singer. "On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines". In: *Journal of Machine Learning Research* 2 (2001), pp. 265–292.

[15] J. H. Friedman. *Another approach to polychotomous classification.* Tech. rep. Stanford University, 1996.

[16] F. Wu, N. J. Guo, H. Tian, et al. "Peripheral blood microRNAs distinguish active ulcerative colitis and Crohn's disease." In: *Inflammatory bowel diseases* 17.1 (2011), pp. 241–50.

[17] R. Duttagupta, S. DiRienzo, R. Jiang, et al. "Genome-wide maps of circulating miRNA biomarkers for ulcerative colitis." In: *PloS one* 7.2 (2012), e31241.

[18] W.-x. Chen, L.-h. Ren, and R.-h. Shi. "Implication of miRNAs for inflammatory bowel disease treatment: Systematic review". In: *World Journal of Gastrointestinal Pathophysiology* 5.2 (2014), pp. 63–70.

[19] A. M. Zahm, M. Thayu, N. J. Hand, et al. "Circulating microRNA is a biomarker of pediatric Crohn disease." In: *Journal of pediatric gastroenterology and nutrition* 53.1 (2011), pp. 26–33.

[20] A. Paraskevi, G. Theodoropoulos, I. Papaconstantinou, et al. "Circulating MicroRNA in inflammatory bowel disease." In: *Journal of Crohn's & colitis* 6.9 (2012), pp. 900–4.

[21] J. S. Schaefer, T. Attumi, A. R. Opekun, et al. "MicroRNA signatures differentiate Crohn's disease from ulcerative colitis". In: *BMC Immunology* 16.1 (2015), pp. 1–13.

[22] F. Allantaz, D. T. Cheng, T. Bergauer, et al. "Expression Profiling of Human Immune Cell Subsets Identifies miRNA-mRNA Regulatory Relationships Correlated with Cell Type Specific Expression". In: *PLoS ONE* 7.1 (2012), e29979.

[23] P. Leidinger, C. Backes, B. Meder, et al. "The human miRNA repertoire of different blood compounds." In: *BMC genomics* 15 (2014), p. 474.

[24] C. J. Mitchell, D. Getnet, M.-S. Kim, et al. "A multi-omic analysis of human naïve CD4+ T cells." In: *BMC systems biology* 9 (2015), p. 75.

4 Curriculum vitae

Personal details

Name:	Matthias Hübenthal
Address:	Kronshagener Weg 38, 24116 Kiel, Germany
Email:	matthiashuebenthal@gmx.net

Education

10/2011–03/2019:	PhD studies of bioinformatics, Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, Kiel, Germany
10/2001–06/2011:	Studies of bioinformatics, Institute of Computer Science, Martin-Luther-University Halle-Wittenberg, Halle, Germany

Research articles

[1] M. Hübenthal, G. Hemmrich-Stanisak, F. Degenhardt, et al. "Sparse Modeling Reveals miRNA Signatures for Diagnostics of Inflammatory Bowel Disease". In: *PLOS ONE* 10.10 (2015), e0140155. (12 citations).

[2] S. Juzėnas, G. Venkatesh, M. Hübenthal, et al. "A comprehensive, cell specific microRNA catalogue of human peripheral blood". In: *Nucleic Acids Research* 45.16 (2017), pp. 9290–9301. (23 citations).

[3] D. Ellinghaus, L. Jostins, S. L. Spain, et al. "Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci". In: *Nature Genetics* 48.5 (2016), pp. 510–518. (174 citations).

[4] J. Wang, L. B. Thingholm, J. Skiecevičienė, et al. "Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota". In: *Nature Genetics* 48.11 (2016), pp. 1396–1406. (147 citations).

[5] H. Westerlind, M.-R. Mellander, F. Bresso, et al. "Dense genotyping of immune-related loci identifies HLA variants associated with increased risk of collagenous colitis". In: *Gut* 66.3 (2017), pp. 421–428. (21 citations).

[6] T. Fehlmann, C. Backes, M. Kahraman, et al. "Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs". In: *Nucleic Acids Research* 45.15 (2017), pp. 8731–8744. (10 citations).

[7] U. Gyvyte, S. Juzėnas, V. Salteniene, et al. "MiRNA profiling of gastrointestinal stromal tumors by next-generation sequencing". In: *Oncotarget* 8.23 (2017). (10 citations).

[8] H. Westerlind, F. Bonfiglio, M.-R. Mellander, et al. "HLA Associations Distinguish Collagenous From Lymphocytic Colitis". In: *The American Journal of Gastroenterology* 111.8 (2016), pp. 1211–1213. (6 citations).

[9] L. Wienbrandt, J. C. Kässens, M. Hübenthal, et al. "Fast Genome-Wide Third-order SNP Interaction Tests with Information Gain on a Low-cost Heterogeneous Parallel FPGA-GPU Computing Architecture". In: *Procedia Computer Science* 108 (2017), pp. 596–605. (4 citations).

[10] L. Thingholm, M. Rühlemann, J. Wang, et al. "Sucrase-isomaltase 15Phe IBS risk variant in relation to dietary carbohydrates and faecal microbiota composition". In: *Gut* 68.1 (2019), pp. 177–178. (3 citations).

[11] F. Bonfiglio, T. Zheng, K. Garcia-Etxebarria, et al. "Female-specific Association Between Variants on Chromosome 9 and Self-reported Diagnosis of Irritable Bowel Syndrome". In: *Gastroenterology* 155.1 (2018), pp. 168–179. (3 citations).

[12] T. Fehlmann, C. Backes, J. Alles, et al. "A high-resolution map of the human small non-coding transcriptome". In: *Bioinformatics* (2017). (3 citations).

[13] L. Wienbrandt, J. C. Kässens, M. Hübenthal, et al. "1,000x Faster than PLINK: Genome-Wide Epistasis Detection with Logistic Regression Using Combined FPGA and GPU Accelerators". In: Lecture Notes in Computer Science 10862 (2018). Ed. by Y. Shi, H. Fu, Y. Tian, et al., pp. 368–381. (1 citations).

[14] E. S. Jung, K.-w. Choi, S. W. Kim, et al. "ZNF133 Is Associated with Infliximab Responsiveness in Patients with Inflammatory Bowel Diseases," in: *Journal of Gastroenterology and Hepatology* (2019).

[15] L. Wienbrandt, J. C. Kässens, M. Hübenthal, et al. "1,000x faster than PLINK: Combined FPGA and GPU accelerators for logistic regression-based detection of epistasis". In: *Journal of Computational Science* 30 (2019), pp. 183–193.

[16] F. Degenhardt, M. Wendorff, M. Wittig, et al. "Construction and benchmarking of a multi-ethnic reference panel for the imputation of HLA class I and II alleles". In: *Human Molecular Genetics* (2018), ddy443.

[17] S. Mucha, H. Baurecht, N. Novak, et al. "Exome-wide association study reveals DOK2 as novel atopic dermatitis susceptibility gene harbouring protective susceptibility variants of low frequency". (in preparation).

[18] L. B. Thingholm, M. C. Rühlemann, M. Koch, et al. "Obese individuals with and without type 2 diabetes show different gut microbial functional capacity and composition". (in preparation).

Books and book chapters

[1] M. Hübenthal, S. Lipinski, A. Franke, et al. "Molecular Genetics of Inflammatory Bowel Disease". In: ed. by M. D'Amato, C. Hedin, and J. D. Rioux. Springer Science+Business Media New York, 2019. Chap. B001: MiRNAs in inflammatory bowel disease. (in press).

Patents

[1] T. Brefort, A. Franke, G. Hemmrich-Stanisak, et al. "MiRNAs as non-invasive biomarkers for inflammatory bowel disease". EP Patent 3212803B1 (granted). 2018.

[2] T. Brefort, A. Franke, G. Hemmrich-Stanisak, et al. "MiRNAs as non-invasive biomarkers for inflammatory bowel disease". US Patent 20170306407A1 (published). 2017.

[3] T. Brefort, A. Franke, G. Hemmrich-Stanisak, et al. "MiRNAs as non-invasive biomarkers for inflammatory bowel disease". WO Patent 2016066288A1 (published). 2016.

Prizes and prize nominations

[1] "Coworking award. Healthcare Hackathon 2018, September 14–15, 2018, Kiel, Germany".

[2] "Best workshop paper award. International Conference on Computational Science, June 11–13, 2018, Wuxi, China".

[3] "Poster prize. Nordic conference on personalized medicine, May 30 – June 1, 2018, Nyborg, Denmark".

[4] "Poster prize nomination. Annual Meeting of the AGD 2014, October 10–11, 2014, Potsdam, Germany".

Conference contributions

[1] G. Hemmrich-Stanisak, M. Hübenthal, M. Spehlmann, et al. "Towards blood-born microRNA biomarkers for inflammatory bowel disease". In: *Regulatory & Non-Coding RNAs Conference, August 28 – September 1, 2012, Cold Spring Harbor, New York, USA*.

[2] G. Hemmrich-Stanisak, M. Spehlmann, A. Elsharawy, et al. "Blood-born diagnostic microRNA biomarkers for inflammatory bowel disease". In: *ASHG 2012 Annual Meeting, November 6–10, 2012, San Francisco, California, USA*.

[3] M. Hübenthal, G. Hemmrich-Stanisak, and A. Franke. "Joint classification and miRNA biomarker selection in inflammatory bowel disease". In: *International 4th Cluster Symposium "Inflammation at Interfaces", February 22–23, 2013, Hamburg, Germany*.

[4] F. Rieder, A. Franke, R. Lopez, et al. "Genetic Risk Profiling Alone or in Combination with Serum Anti-Microbial Antibodies for the Stratification of Complicated Crohn's Disease Courses". In: *Digestive Disease Week 2013, May 18–21, 2013, Orlando, Florida, USA*.

[5] M. Hübenthal, G. Hemmrich-Stanisak, Z. G. Du, et al. "Using blood-born miRNA profiles and machine-learning techniques to predict inflammatory phenotypes". In: *ASHG 2013 Annual Meeting, October 22–26, 2013, Boston, Massachusetts, USA*.

[6] Z. G. Du, M. Hübenthal, W. Lieb, et al. "Analysis of next generation sequencing derived smallRNA data: a comparison of current software tools". In: *HiTSeq 2014, High Throughput Sequencing Algorithms and Applications, ISMB 2014 Special Interest Group Meeting, July 11–12, 2014, Boston, Massachusetts, USA*.

[7] S. Juzėnas, J. Skieceviciene, J. Kupčinskas, et al. "Micro-RNA as biomarkers for early diagnosis of premalignant and malignant colon diseases". In: *Summer School on "Systems Medicine", September 7–11, 2014, Kiel, Germany*.

[8]   M. Hübenthal, A. Keller, A. Elsharawy, et al. "Using miRNAs and sparse modeling to predict inflammatory phenotypes". In: *Annual Meeting of the AGD 2014, October 10–11, 2014, Potsdam, Germany.*

[9]   Z. G. Du, M. Hübenthal, M. Paulsen, et al. "Defining the transcriptional landscape of microRNAs in human peripheral blood". In: *ASHG 2014 Annual Meeting, October 18–22, 2014, San Diego, California, USA.*

[10]   M. Hübenthal and A. Franke. "Genome-wide study of anti-TNF response in IBD". In: *UEG Week 2014, October 18–22, 2014, Vienna, Austria.*

[11]   F. Hadizadeh, M. Henstrom, M. Belheouane, et al. "A genome-wide study of the impact of human genetic variation on gut microbiota composition". In: *International 5th Cluster Symposium "Inflammation at Interfaces", February 26–28, 2015, Kiel, Germany.*

[12]   S. Juzėnas, J. Skieceviciene, J. Kupčinskas, et al. "MiRNA expression patterns in colon of active and inactive ulcerative colitis". In: *International 5th Cluster Symposium "Inflammation at Interfaces", February 26–28, 2015, Kiel, Germany.*

[13]   M. Hübenthal, G. Hemmrich-Stanisak, F. Degenhardt, et al. "Using miRNA expression profiling and sparse modeling for diagnostics of inflammatory bowel disease". In: *International 5th Cluster Symposium "Inflammation at Interfaces", February 26–28, 2015, Kiel, Germany.*

[14]   M. Hübenthal, G. Hemmrich-Stanisak, F. Degenhardt, et al. "Using sparse modeling to construct miRNA signatures for diagnostics of inflammatory bowel disease". In: *International RTG Symposium "Genes, Environment and Inflammation", June 1–3, 2015, Kiel, Germany.*

[15]   K. Chaichoompu, I. Cleynen, R. Fouladi, et al. "Detecting patient subgroups using reduced set of disease-related markers with iterative pruning Principal Component Analysis (ipPCA)". In: *IGES 2015 Meeting, October 4–6, 2015, Baltimore, Maryland, USA.*

[16]   M. Hübenthal, G. Hemmrich-Stanisak, F. Degenhardt, et al. "Using miRNA expression profiling and sparse machine learning methods for diagnostics of inflammatory bowel disease". In: *ASHG 2015 Annual Meeting, October 6–10, 2015, Baltimore, Maryland, USA.*

[17]   G. Venkatesh, M. Hübenthal, S. Juzėnas, et al. "The transcriptional landscape of microRNAs in human blood". In: *Advances in Genome Biology and Technology (AGBT) General Meeting, February 10–13, 2016, Orlando, Florida, USA.*

[18]   M. Hübenthal, S. Juzėnas, S. Zeißig, et al. "Sparse sequencing-based profiles of blood-born miRNAs for diagnostics of inflammatory bowel disease". In: *e:Med Meeting 2016 on Systems Medicine, October 4–6, 2016, Kiel, Germany.*

[19]   S. Juzėnas, J. Skieceviciene, V. Salteniene, et al. "MiRNA expression patterns in colon of active and inactive ulcerative colitis". In: *12th Congress of ECCO, February 15–18, 2017, Barcelona, Spain.*

[20]   L. Wienbrandt, J. C. Kässens, M. Hübenthal, et al. "Fast Genome-Wide Third-order SNP Interaction Tests with Information Gain on a Low-cost Heterogeneous Parallel FPGA-GPU Computing Architecture". In: *International Conference on Computational Science, June 12–14, 2017, Zurich, Switzerland.*

[21]   U. Gsell, M. Hübenthal, S. Schreiber, et al. "Popgen-OSSA: Development of an Organ Specific Self Assessment (OSSA) for interdisciplinary documentation of patient reported clinical outcomes". In: *Annual European Congress of Rheumatology, June 14–17, 2017, Madrid, Spain.*

[22]   L. Wienbrandt, J. C. Kässens, M. Hübenthal, et al. "Ultra-fast 2-way and 3-way SNP Interaction Tests on FPGAs and GPUs". In: *ISMB/ECCB, July 21–25, 2017, Prague, Czech Republic.*

[23]   S. Juzėnas, G. Venkatesh, M. Hübenthal, et al. "High-resolution miRNome analysis of the human peripheral blood". In: *ISMB/ECCB, July 21–25, 2017, Prague, Czech Republic.*

[24]   M. Hübenthal, S. Juzėnas, S. Zeißig, et al. "Sparse sequencing-based profiles of isomiRs for non-invasive diagnostics of inflammatory phenotypes". In: *ISMB/ECCB, July 21–25, 2017, Prague, Czech Republic.*

[25] G. Streleckiene, J. Kupčinskas, S. Juzėnas, et al. "Role of miR-20B-5p, miR-451a-5p and miR-1468-5p in gastric cancer". In: *25th UEG Week, October 28 – November 1, 2017, Barcelona, Spain.*

[26] L. Wienbrandt, J. C. Kässens, M. Hübenthal, et al. "1,000x Faster than PLINK: Genome-Wide Epistasis Detection with Logistic Regression Using Combined FPGA and GPU Accelerators". In: *Research in Computational Molecular Biology (RECOMB) 2018, April 21–24, 2018, Paris, France.*

[27] R. Böhm, H. Bruckmüller, M. Hübenthal, et al. "Preliminary results of the clinical trial "genotype-phenotype correlation of drug-metabolizing enzymes"". In: *Nordic conference on personalized medicine, May 30 – June 1, 2018, Nyborg, Denmark.*

[28] L. Wienbrandt, J. C. Kässens, M. Hübenthal, et al. "1,000x Faster than PLINK: Genome-Wide Epistasis Detection with Logistic Regression Using Combined FPGA and GPU Accelerators". In: *International Conference on Computational Science, June 11–13, 2018, Wuxi, China.*

[29] F. Degenhardt, G. Mayr, G. Boucher, et al. "Trans-ethnic analysis of the human leukocyte antigen region for ulcerative colitis reveals common disease signatures". In: *ASHG 2018 Annual Meeting, October 16–20, 2018, San Diego, California, USA.*

[30] D. Ellinghaus, L. Wienbrandt, J. C. Kässens, et al. "1,000x Faster than PLINK: Genome-Wide Epistasis Detection with Logistic Regression Using Combined FPGA and GPU Accelerators". In: *ASHG 2018 Annual Meeting, October 16–20, 2018, San Diego, California, USA.*

[31] S. Juzėnas, M. Hübenthal, S. Zeißig, et al. "Sequencing-based hematopoietic miRNA landscape reveals common and distinct features of autoimmune inflammatory phenotypes (DOP27)". In: *14th Congress of ECCO, March 6–9, 2019, Copenhagen, Denmark.*

[32] E. S. Jung, K.-w. Choi, S. W. Kim, et al. "ZNF133 is associated with infliximab responsiveness in patients with inflammatory bowel diseases using whole-exome sequencing (DOP55)". In: *14th Congress of ECCO, March 6–9, 2019, Copenhagen, Denmark.*

Conference attendances

[1] "International Symposium on Integrative Bioinformatics, August 20–22, 2008, Lutherstadt Wittenberg, Germany".

[2] "German Conference on Bioinformatics 2009, September 28–30, 2009, Halle (Saale), Germany".

[3] "International symposium "Predictive Genetic Testing, Risk Communication and Risk Perception", November 21–22, 2011, Robert Koch Institute, Berlin, Germany".

[4] "Annual Meeting of the AGD 2012, October 12–13, 2012, Potsdam, Germany".

[5] "International 4th Cluster Symposium "Inflammation at Interfaces", February 22–23, 2013, Hamburg, Germany".

[6] "Dritter Workshop zur Genetischen Epidemiologie "Genetic Variants, Omics, and Individualized Prevention", April 10–12, 2013, Grainau, Germany".

[7] "ASHG 2013 Annual Meeting, October 22–26, 2013, Boston, Massachusetts, USA".

[8] "Annual Meeting of the AGD 2014, October 10–11, 2014, Potsdam, Germany".

[9] "UEG Week 2014, October 18–22, 2014, Vienna, Austria".

[10] "International 5th Cluster Symposium "Inflammation at Interfaces", February 26–28, 2015, Kiel, Germany".

[11] "International RTG Symposium "Genes, Environment and Inflammation", June 1–3, 2015, Kiel, Germany".

[12] "ASHG 2015 Annual Meeting, October 6–10, 2015, Baltimore, Maryland, USA".

[13] "DAGStat 2016, March 14–18, 2016, Göttingen, Germany".

[14] "e:Med Meeting 2016 on Systems Medicine, October 4–6, 2016, Kiel, Germany".

[15] "ISMB/ECCB, July 21–25, 2017, Prague, Czech Republic".

Activities and memberships

[1] "Associate membership, Cluster of Excellence 306 "Inflammation at Interfaces"".

[2] "Peer reviewing, "Genomics, Proteomics & Bioinformatics"".

[3] "Peer reviewing, "Human Mutation"".

Workshop attendances

[1] "Agfa ORBIS, December 13, 2011, UKSH Akademie, Kiel, Germany".

[2] "3. Workshop der Deutschen i2b2-Usergroup, February 27, 2012, TMF e.V., Berlin, Germany".

[3] "Training in Genetischer Epidemiologie, April 19–21, 2012, University of Lübeck, Lübeck, Germany".

[4] "TMF school extra "Logical Reasoning in Human Genetics", September 24–28, 2012, TMF e.V., Berlin, Germany".

[5] "Inflammation at Interfaces Indian Summer School "A Comprehensive View of Experimental Approaches in Inflammation Research", November 19–21, 2013, Kiel, Germany".

[6] "Workshop Theoretical Biology, February 28, 2014, Plön, Germany".

Trainings and coachings

[1] "Feedback culture. Helga Hänsler, Kiel, Germany".

[2] "Oral presentation skills. Carla Westerheide, ProSciencia, Lübeck, Germany".

[3] "Coaching. Helga Hänsler, Kiel, Germany".