

An expression module of WIPF1-coexpressed genes identifies patients with favorable prognosis in three tumor types

Eike Staub · Joern Groene · Maya Heinze · Detlev Mennerich · Stefan Roepcke · Irina Klamann · Bernd Hinzmann · Esmeralda Castanos-Velez · Christian Pilarsky · Benno Mann · Thomas Brümmendorf · Birgit Weber · Heinz-Johannes Buhr · André Rosenthal

Received: 25 November 2008 / Revised: 24 February 2009 / Accepted: 27 March 2009 / Published online: 28 April 2009
© The Author(s) 2009. This article is published with open access at Springerlink.com

Abstract Wiskott–Aldrich syndrome (WAS) predisposes patients to leukemia and lymphoma. WAS is caused by mutations in the protein WASP which impair its interaction with the WIPF1 protein. Here, we aim to identify a module of WIPF1-coexpressed genes and to assess its use as a prognostic signature for colorectal cancer, glioma, and breast cancer patients. Two public colorectal cancer microarray data sets were used for discovery and validation of the WIPF1 co-expression module. Based on expression of the WIPF1 signature, we classified more than 400 additional tumors with microarray data from our own experiments or from publicly available data sets according to their WIPF1 signature expression. This allowed us to separate patient populations for colorectal cancers, breast cancers, and

gliomas for which clinical characteristics like survival times and times to relapse were analyzed. Groups of colorectal cancer, breast cancer, and glioma patients with low expression of the WIPF1 co-expression module generally had a favorable prognosis. In addition, the majority of WIPF1 signature genes are individually correlated with disease outcome in different studies. Literature gene network analysis revealed that among WIPF1 co-expressed genes known direct transcriptional targets of c-myc, ESR1 and p53 are enriched. The mean expression profile of WIPF1 signature genes is correlated with the profile of a proliferation signature. The WIPF1 signature is the first microarray-based prognostic expression signature primarily developed for colorectal cancer that is instrumental in other tumor types: low expression of the WIPF1 module is associated with better prognosis.

Eike Staub and Joern Groene contributed equally as first author for this journal.

Electronic supplementary material The online version of this article (doi:10.1007/s00109-009-0467-y) contains supplementary material, which is available to authorized users.

E. Staub · S. Roepcke
Department of Computational Molecular Biology,
Max Planck Institute for Molecular Genetics,
Berlin, Germany

E. Staub (✉) · T. Brümmendorf · B. Weber
metaGen Pharmaceuticals i.L.,
Berlin, Germany
e-mail: eike.staub@nucleolus.net

J. Groene (✉) · M. Heinze · H.-J. Buhr
Department of General, Vascular and Thoracic Surgery,
Charité—Campus Benjamin Franklin,
Berlin, Germany

D. Mennerich
Boehringer Ingelheim Pharma GmbH and Co KG,
Biberach, Germany

Keywords Colorectal cancer · WIPF1 · Prognosis · Expression signature · Microarray

I. Klamann · B. Hinzmann · A. Rosenthal
Signature Diagnostics,
Potsdam, Germany

E. Castanos-Velez
Epigenomics,
Berlin, Germany

C. Pilarsky
Department of Visceral, Thoracic, and Vascular Surgery,
University Hospital Carl Gustav Carus Dresden,
Dresden, Germany

B. Mann
Department of Surgery, Augusta-Kranken-Anstalt GmbH,
Bochum, Germany

Introduction

The WIPF1 gene encodes the WASP/WASL interacting protein family member 1 that plays an important role in the organization of the actin cytoskeleton [1, 2]. The WIPF1-encoded protein WIP binds to a region of Wiskott–Aldrich syndrome protein (WASP) that is frequently mutated in patients with Wiskott–Aldrich syndrome (WAS) [3, 4], and WIP mutations themselves lead to an immunological disorder resembling Wiskott–Aldrich syndrome [5]. WAS is an X-linked recessive disease that predisposes to leukemia and lymphoma. The WIP protein is essential for WASP synthesis and probably acts as its chaperone [6]. Disruption of the WASP-WIP interaction by hereditary mutations leads to a rounded cell surface on immune cells, a conversion that is thought to coincide with a diminished capability to form immune synapses and reduction of NK cell cytotoxicity [7]. WIP is important for podosome formation in macrophages and cellular fusions in flies [8, 9], stressing its universal role in the design of cell membrane remodeling. Apart from its expression in diverse immune cells, several human tissues exhibit WIP expression. Little is known about the expression of WIPF1 in solid tumors. However, WIPF1 expression levels influence morphology and migration of fibroblasts [10]. This prompted us to investigate the expression characteristics of WIPF1 in colorectal tumors with the aim to study its potential for prognosis.

A multitude of microarray studies have been carried out during the past decade to gain a better understanding of basic colorectal cancer (CRC) biology [11–18]. Other CRC microarray studies led to the discovery of informative gene sets for the prediction of the response to therapy or tumor recurrence [19–23], diagnosis of tumor stage [24–27], lymph node metastasis [28–30], or liver metastasis [31,

32]. Until now, cross-validation of diagnostic or prognostic signatures with independent data sets has hardly been performed for colorectal cancer. This is probably due to the fact that published signatures are only overlapping to a small degree and that they are difficult to reproduce when originating from different laboratories and platforms (for a discussion see Groene et al. [24]). In addition, until recently, data sets with sufficient patient information were lacking in public databases which hindered cross-validation of signatures from different studies.

Here, we describe the identification of a set of genes that is co-expressed with WIPF1. It was discovered through re-analysis of two public microarray data sets on clinical colorectal cancer specimen that were deposited in the Gene Expression Omnibus (GEO) database (<http://www.ncbi.nlm.nih.gov/geo>). Whereas the first data set was used for discovery of WIPF1-coexpressed genes, the second data set served us for validation of the expression correlations. Using a simple classification algorithm trained on microarray data of WIPF1 co-expressed genes from the two studies, we identified patients with characteristic expression of the WIPF1 coexpression module in three further microarray data sets with information about survival or relapse of patients: a colorectal cancer data set of our own, a breast cancer data set, and a glioblastoma data set. Strikingly, patients with low expression of the WIPF1 signature have the best prognosis in all three data sets, in total comprising more than 400 patients. Based on an analysis of the WIPF1 coexpression module in the context of literature-based gene networks, we identified plausible regulatory mechanisms responsible for lower WIPF1 module expression in patients with better prognosis.

Materials and methods

Sample acquisition

For the generation of our own microarray data set, 62 CRC patients undergoing elective standard oncological resection at the Department of General, Vascular, and Thoracic Surgery, Campus Benjamin Franklin, Charité, were prospectively recruited. Several clinical characteristics of the patients were recorded (see Supplementary Table 1). The study was approved by the local ethical committee, and informed consent was obtained from all patients.

Laser-capture microdissection

Pre-processing of frozen tissue blocks by laser-capture microdissection was essentially performed as described in our previous publications [13, 14, 24, 33]. Briefly, all cancer specimens were snap frozen within 20 min follow-

Present Address:

E. Staub
Merck Serono R&D, Bio- and Chemoinformatics, Merck KGaA,
Darmstadt, Germany

Present Address:

T. Brümmendorf
Novartis Institutes for BioMedical Research,
Novartis Pharma AG,
Basel, Switzerland

Present Address:

S. Roepcke
Nycomed,
Konstanz, Germany

Present Address:

B. Weber
Ratiopharm,
Ulm, Germany

ing excision. After laser-capture microdissection frozen tissue specimens were serially cut into 6- to 8- μ m-thick sections which were mounted on a sterile 2.5 mm membrane. Slides were fixed in 70% ethanol. The sections were briefly stained with hematoxylin and eosin, dehydrated in ethanol, and dried for 10–15 min using an exsiccator. The membrane was turned around and fixed with adhesive tape on the other sterile slide. First slides served as a template on which the areas of tumor or normal epithelium were marked. On the consecutive section, these areas were microdissected using a laser microdissection system (SL, Jena, Germany and P.A.L.M. Microlaser Technologies AG Bernried, Germany) and capture transfer films (Arcturus GmbH, Moerfelden-Walldorf, Germany). For molecular analysis, up to 100,000 cells or approximately 30–60 mm² of tissue section areas were pooled and collected in ice-cooled tubes containing 100 ml of 98% guanidine thiocyanate (GTC) buffer and 2% beta-mercaptoethanol.

Messenger RNA preparation and DNA chip hybridization

PolyA mRNA from the microdissected specimens was prepared using the PolyA-tract 1000 kit (Promega, Heidelberg, Germany) according to the manufacturer's recommendations. For each sample, the cDNA synthesis and repetitive in vitro transcription were performed three times. The total amount of prepared mRNA from each sample was used. First strand cDNA synthesis was initiated using the Affymetrix T7-oligo-dT promoter–primer combination at 0.1 mM. The second strand cDNA synthesis was generated by internal priming. In vitro transcription was performed using the Megascript kit (Ambion, Huntington, UK) as recommended by Ambion. From the generated cRNA, a new first strand synthesis was initiated using 0.025 mM of a random hexamer as primer. After completion, the second strand synthesis was performed using the Affymetrix T7-oligo-dT promoter–primer combination. A second in vitro transcription was performed, and then the procedure was repeated one additional time. During the last in vitro transcription, biotin-labeled ribonucleotides were incorporated into the cRNA, as recommended by the Affymetrix protocol. Hybridization and detection of the labeled cRNA on the Affymetrix U133A Chip were performed according to Affymetrix standard protocol.

Microarray data pre-processing

Public expression data was downloaded from the Gene Expression Omnibus (GEO) database (<http://www.ncbi.nlm.nih.gov/projects/geo/>). In addition to our own data that was deposited in GEO with accession number GSE12945, we used four different data sets from this repository. The colorectal cancer data sets GSE5206 of the Aronow group

(see Kaiser et al. [17]), GSE7208 of Ayers and co-workers [20] and of our own served us for discovery and validation of the WIPF1 signature in CRC. The GSE2034 data set of Wang et al. served us for assessment of predictive power of

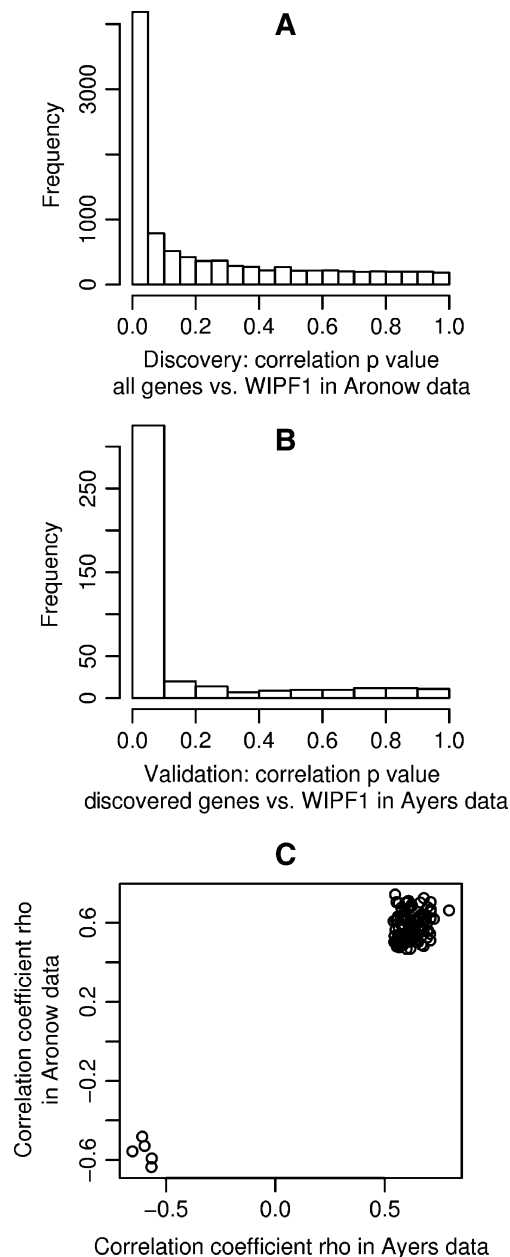
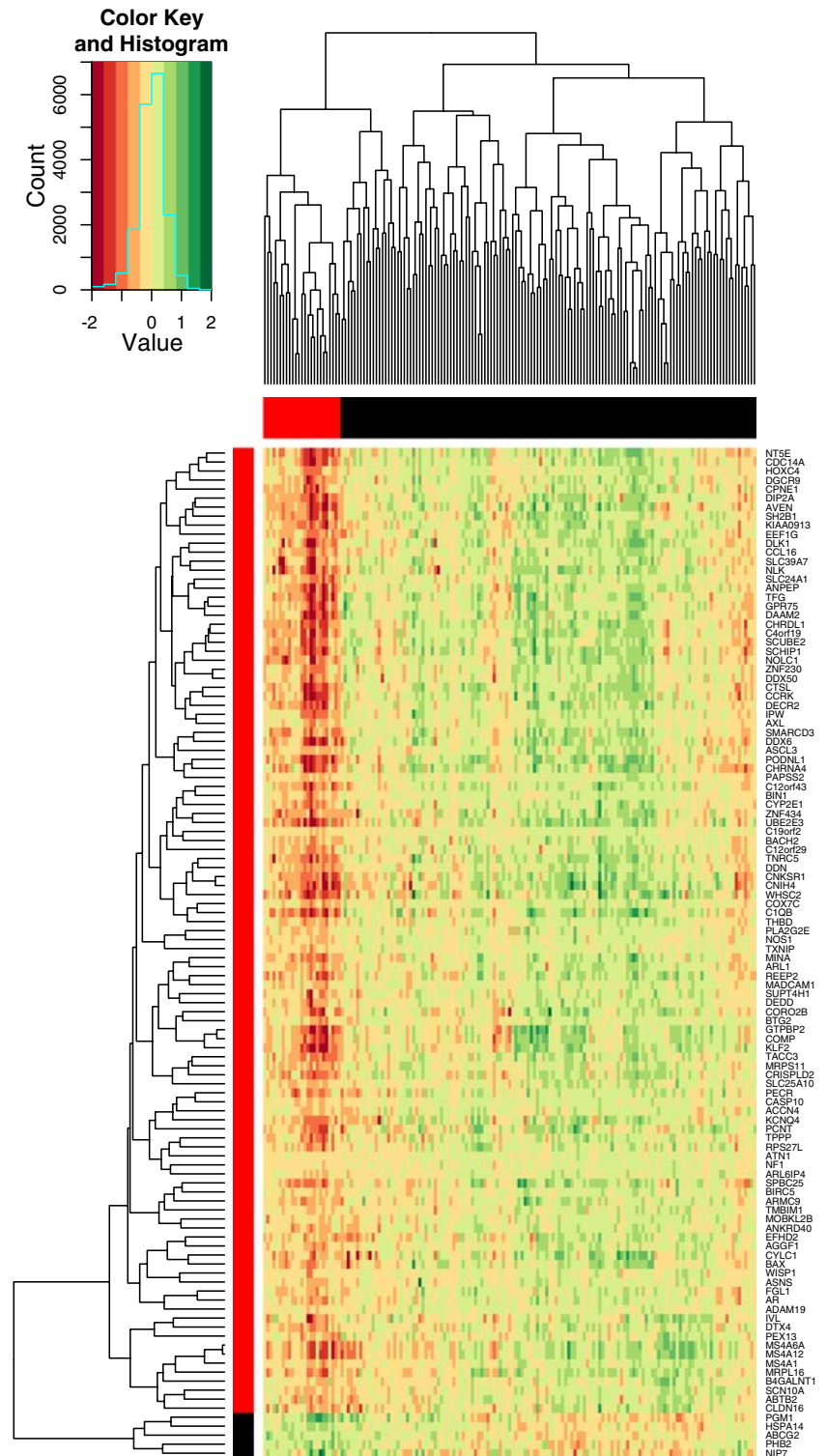


Fig. 1 Discovery and validation of the WIPF1 coexpression module. **a** The histogram shows the distribution of p values resulting from tests for correlation of the WIPF1 expression profile with expression profiles of each other gene in the Aronow data set. **b** The histogram shows the distribution of 430 p values resulting from correlation of expression profiles in the Ayers data set of the WIPF1 gene with each discovered gene in the Aronow data set. Note the strong deviation from the uniform distribution, indicating a high rate of successful validation. **c** A scatter plot of the correlation coefficients in Aronow and Ayers data of each of the 112 successfully validated genes reveals that most validated genes are positively correlated with the WIPF1 profile and only six genes are anti-correlated with WIPF1. All correlation coefficients are consistent in both data sets, i.e., they have the same sign

the WIPF1 signature for breast cancer patients [34]. The data set of Phillips et al. (GSE4271) served us for assessment of predictive power of the WIPF1 signature in high-grade glioma patients [35]. We refer to the original publications and the GEO database for patient and sample characteristics.

Fig. 2 Two-way hierarchical clustering of colorectal cancer expression data from WIPF1-coexpressed genes. The data set results from fusing the Aronow and Ayers data sets and comprises 112 genes and 159 patients. *Top*: The dendrogram results from hierarchical clustering of tumors. *Light gray* indicates low expression, *dark gray* indicates high expression (log-transformed mean-centered). Five clusters of tumors are evident: cluster A is colored in *red*, all other clusters are colored in *black*. *Left*: The dendrogram results from hierarchical clustering of genes. Two main clusters are evident: the larger cluster X is colored in *red*, the smaller cluster Y in *black*. Note the characteristic down-regulation of genes from cluster X in tumors from cluster A. Other tumors have anti-correlated or indifferent expression profiles



For our own microarray experiment, we used algorithms implemented in the freely available statistical software package R (<http://www.r-project.org/>) and its public package repositories CRAN (<http://cran.r-project.org/>) and the bioinformatics R package repository Bioconductor (<http://www.bioconductor.org/>). Preprocessing: Raw expression

data were condensed to probe set-wise intensity values using the RMA algorithm. For experiment normalization, all colorectal cancer data sets were pre-processed using the same data transformations. If not already done, the raw data were log transformed. Independently of the original authors' pre-processing, we quantile-normalized each data set on the probe set level. Then we filtered out the 10% probesets with lowest median expression and 10% probesets with lowest variance to enrich informative probesets in an unbiased way. We restricted the further analysis to probesets passing this filter in each of the three data sets. Probeset annotations for Affymetrix expression microarrays (most importantly gene symbols) were retrieved from the Affymetrix web site (version 22). We condensed probeset signals on the gene symbol level by averaging across all remaining probesets of a gene. On the gene level, each data set was finally mean centered. Breast cancer and glioma data were processed in the same way: but here no additional probesets were filtered out due to low variance or mean expression signals in these data sets. Instead, for mapping of probeset expression intensities onto the gene level, we considered all probesets that were used for probeset-to-gene mapping during processing of the three CRC data sets. Using the applied normalization scheme, we intend to ensure that Affymetrix microarray data from the five different studies are comparable (we note that our study does not include a cross-platform comparison as all data sets were generated using Affymetrix gene chips).

Expression data mining

Expression data mining was carried out using the statistics software *R* supplemented with diverse packages from the *CRAN* or *Bioconductor* projects. Analysis of the correlation of two expression profiles was evaluated with Pearson correlation coefficients determined with the function *cor.test* in package *stats*. Using the same function we determined *p* values for the significance of the deviation of the correlation coefficient from 0. The average expression profile of a multigene expression signature (proliferation signature by Rosenwald et al. [36]), here denoted as *signature centroid*, was determined using averaging across signature genes for each patient. For tumor class discovery, we applied hierarchical clustering using the *heatmap.2* function of the *gplots* package in *R*. The distance matrices for row and column clustering were determined using pairwise *correlation distances* ($d=0.5(1-\text{cor}(x,y))$) of the gene-wise mean-centered expression intensities of genes and samples, respectively. For clustering, we used the *complete linkage* hierarchical clustering algorithm.

In the preceding paragraph, we described how we identify tumor classes in a training data set using unsupervised analysis (clustering). Using classification

(supervised analysis), we then attempted to detect the tumor classes that we identified on a primary (training) data set in tumors of secondary (test) data sets from independent studies. As a classification algorithm, we used *k*-nearest neighbor classification as implemented in the *R* package *class* (function *knn* with $k=9$). The classifiers were trained on tumor expression profiles of the training data and then directly applied on test data sets. A prerequisite for our classifier to work properly on the test data (here solely external data sets from independent studies) is that training and test data are sufficiently normalized which in our study should be ensured by log-transforming the expression values followed by gene-wise mean centering. Across-study normalization based on intra-study mean centering of log-transformed expression intensities from Affymetrix chips was already shown by Lusa et al. to be a pre-processing strategy that can be the basis of good classifier performance when the aim is to construct gene expression-based predictors for tumor classes across studies [37]. However, the authors also stated that best classifier performance can only be expected if the fraction of tumor classes in the different data sets is comparable. In our own validation of our methodology, we found that *k*-nearest neighbor-based predictors of estrogen receptor status in breast cancers based on an estrogen-responsive set of genes achieves prediction accuracies on external data sets of ~90% on average on gene-wise mean-centered expression data (based on four Affymetrix U133A microarray data sets, data available upon request, manuscript in preparation). Classifier performance was still at 76% when the

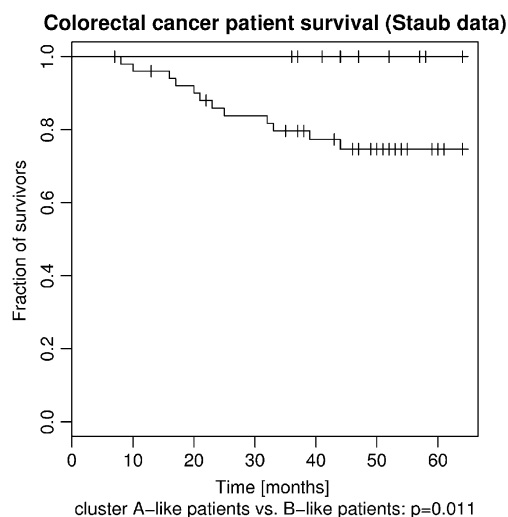


Fig. 3 Differences in survival between colorectal cancer patients of the “cluster B” type in an independent microarray data set. Tumors that resembled the cluster “A” type were predicted by machine learning in an independent data set of 62 tumors with microarray expression profiles. Nine cluster “A”-like patients all survived until the study end. The difference in survival compared with 53 remaining patients is significant with $p=0.011$ in the logrank test

tumor classes were not balanced (e.g., 1:8 in test data). We note that we did not attempt to assess classifier accuracy using cross-validation on the primary data because we have derived the tumor classes by data mining in the complete

primary data set: it is obvious that a numerical difference between classes exists. Therefore, the assessment of classifier performance using cross-validation could lead to a serious overestimation of classifier accuracy. The reason

Table 1 Summary of WIPF1 correlation and survival/relapse association for 38 core genes of the WIPF1 module and WIPF1 itself

Gene	WIPF1 correlation <i>p</i> Value Aronow data	WIPF1 correlation <i>p</i> Value Ayers data	WIPF1 correlation <i>p</i> Value Staub data	Cox regression Wald test <i>p</i> Value Wang data	Cox regression Wald test <i>p</i> Value Phillips data
AGGF1	1.01E-09	6.40E-06	6.66E-03	4.74E-01	7.74E-01
AVEN	1.87E-10	1.04E-08	9.21E-08	1.23E-04	6.37E-04
AXL	3.15E-07	4.18E-06	1.27E-04	9.10E-02	2.23E-05
BIN1	9.45E-08	4.09E-08	1.57E-04	8.26E-02	1.01E-02
C12orf29	9.28E-07	4.99E-07	9.31E-04	4.75E-02	7.31E-06
C12orf43	9.05E-12	1.61E-09	3.27E-03	2.46E-02	1.35E-04
C1QB	4.44E-16	6.73E-07	9.96E-04	1.76E-02	1.56E-02
CCL16	8.44E-15	9.04E-10	3.02E-04	4.16E-03	9.65E-04
CHRDL1	1.22E-09	1.17E-07	2.10E-03	1.34E-03	5.83E-03
CYLC1	9.27E-12	1.91E-06	5.33E-04	1.16E-03	2.38E-01
DDX50	3.69E-07	3.60E-09	1.35E-08	5.63E-04	7.08E-03
DECR2	4.89E-12	4.55E-07	8.30E-07	8.97E-04	1.46E-04
DEDD	7.73E-12	4.91E-08	6.10E-11	1.48E-03	2.74E-02
DTX4	4.00E-09	3.28E-07	7.26E-04	7.73E-01	5.25E-05
EFHD2	3.97E-08	2.55E-06	3.42E-03	2.05E-01	2.42E-02
GTPBP2	4.64E-12	3.01E-10	1.26E-05	1.55E-05	1.63E-02
HSPA14	1.18E-12	2.92E-06	1.84E-03	6.09E-01	6.07E-03
IPW	9.49E-10	2.44E-07	1.64E-05	2.38E-03	3.69E-02
MADCAM1	1.39E-08	2.19E-06	1.16E-03	3.89E-01	1.22E-02
MINA	0.00E+00	6.98E-06	8.23E-05	9.65E-02	5.60E-03
MOBK2B	8.13E-12	4.70E-06	5.86E-03	1.61E-01	9.03E-01
MRPL16	1.48E-09	9.52E-10	8.58E-04	1.36E-01	4.49E-02
MS4A12	1.36E-13	2.77E-10	2.02E-04	4.64E-01	1.02E-03
NF1	1.31E-14	2.16E-07	6.94E-03	3.51E-01	6.15E-02
NIP7	1.92E-09	2.03E-08	7.09E-03	4.60E-01	8.45E-01
NLK	0.00E+00	3.29E-09	6.99E-05	7.86E-03	3.52E-02
NT5E	3.61E-12	4.34E-09	1.05E-04	5.92E-06	4.19E-03
PGM1	3.93E-07	3.17E-07	4.95E-03	4.04E-01	8.73E-01
PLA2G2E	1.99E-11	9.69E-06	6.35E-04	2.24E-01	1.04E-03
SCUBE2	4.93E-08	4.20E-06	1.94E-03	7.52E-04	2.49E-03
SH2B1	1.29E-11	2.68E-08	1.28E-06	5.99E-03	6.13E-03
SLC24A1	1.06E-09	5.12E-08	4.70E-03	2.26E-05	1.92E-04
SLC39A7	1.17E-07	4.49E-06	1.72E-04	3.57E-03	9.01E-05
SPBC25	1.28E-11	5.89E-06	2.90E-05	1.48E-03	2.39E-03
THBD	1.02E-14	1.78E-06	3.81E-03	1.79E-03	8.01E-04
TXNIP	8.99E-08	1.85E-07	9.15E-04	2.85E-01	1.25E-01
UBE2E3	1.55E-15	1.31E-07	5.94E-04	4.08E-04	3.28E-02
ZNF230	1.60E-09	3.87E-09	1.90E-09	6.37E-04	1.03E-03
WIPF1	–	–	–	6.92E-06	7.80E-02

A core set of 38 genes of the WIPF1 module (discovered in the Aronow data set) for which WIPF1 coexpression could be double validated in the Ayers and Staub data sets at stringent thresholds are listed here together with the *p* values of their WIPF1 correlation tests and of their survival/relapse association tests. A list of results for the full set of 112 genes of the WIPF1 module is given in Supplementary Table 2.

is that even if test cases in cross-validation are not used for classifier training, they were already included in the initial clustering analysis that led to the assignment of class labels: a violation of a main principle of cross validation—the independence of training and test data.

Patient survival and relapse were visualized with Kaplan–Meier curves determined with the *survfit* function in the *survival* package. The logrank test as implemented in the *coxph* function of the *survival* package was used to assess the significance of differences in survival/relapse times between patient groups. The significance of association of continuous variables, here gene expression intensities, with survival/relapse were tested using Cox regression and Wald tests on the model coefficients and their variances as implemented in the *coxph* function. For survival analyses, we used robust estimates of Cox model coefficient variances

(parameter *robust=T*). If not otherwise indicated, default parameter settings were used in the functions mentioned above.

Literature-based gene networks were investigated using the metaCore software from GeneGo (Saxony Road, #104, Encinitas, CA 92024, USA). Subnetworks of genes with functional links based on literature evidence were screened for enrichment of genes co-expressed with WIPF1. Those networks with significant enrichment of WIPF1 genes were further investigated for enrichment of Gene Ontology categories. The significance of enrichment of either user supplied gene lists (like the gene list of the WIPF1 signature) or genes lists associated with Gene Ontology (GO) terms in gene lists of literature subnetworks is given by hypergeometric *p* values (with the complete set of human genes with literature information as a reference set).

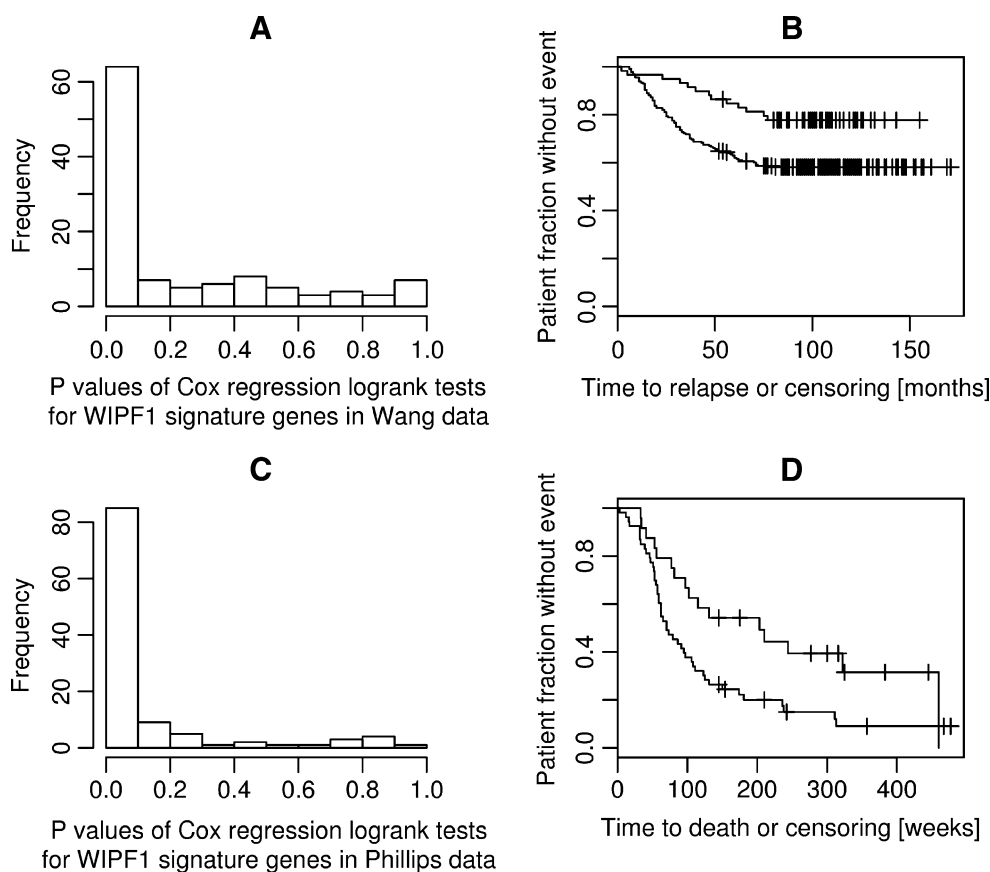
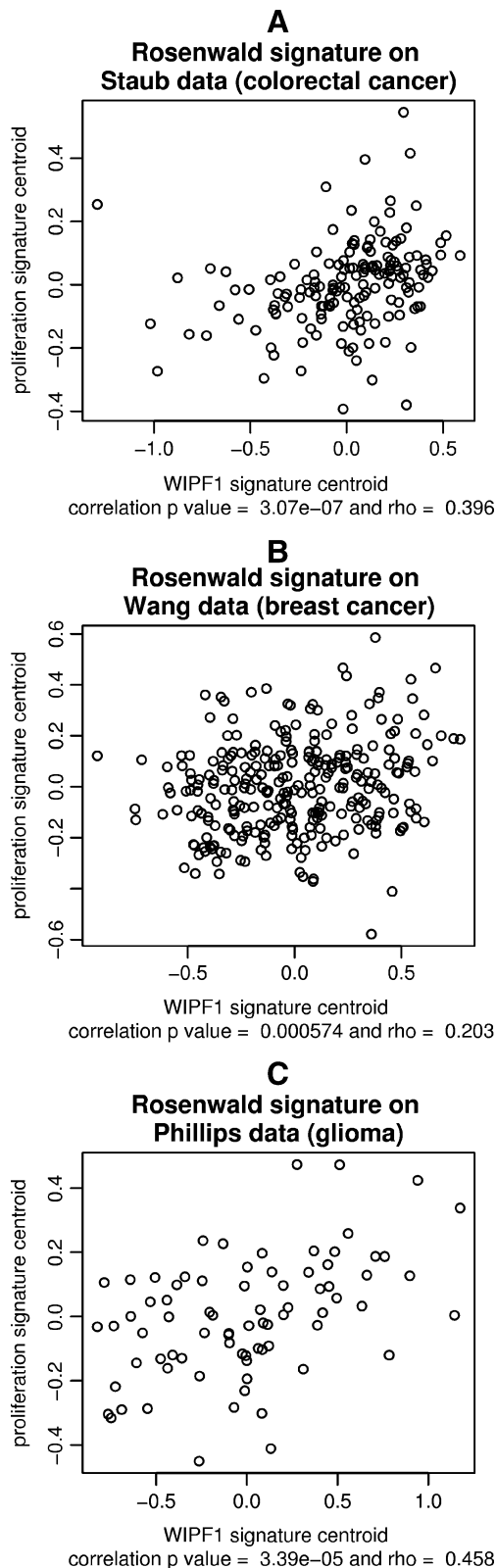


Fig. 4 The WIPF1 module identifies breast cancer and glioma patients with better prognosis. Histogram (**a**) visualizes the distribution of *p* values resulting from logrank tests for the association of profiles of single genes of the WIPF1 module with relapse in the Wang data on breast cancers. **b** The Kaplan–Meier curves for breast cancer patients assigned to the groups cluster “A” (upper curve) or cluster “B”. Histogram (**c**) visualizes the distribution of *p* values resulting from logrank tests for the association of single expression profiles of the WIPF1 coexpressed genes with relapse in the Phillips

data on gliomas. **d** The Kaplan–Meier curves for glioma patients assigned to the groups “A” (upper curve) or “B”. Note the strong deviation from the uniform distribution (that would result from chance association) in plots (**a**) or (**c**) and the consistently higher fractions of survivors among “A” patients in (**b**) and (**d**). These plots visualize the high fraction of genes of the WIPF1 module that are individually associated with survival in breast and brain cancers and the survival significance of the complete module



◀ **Fig. 5** Similarity of the WIPF1 and proliferation signatures. We determined average profiles (signature centroids) of the WIPF1 and the proliferation signatures in three data sets. Only 107 of 112 WIPF1 module genes with concordant lower expression in cluster X of the WIPF1 signature were considered for this analysis. Signature centroids were determined using averaging over all genes for each patient. The scatter plots visualize the strength of the correlation between proliferation signature and WIPF1 signature. Coordinates of each data point correspond to a single patient's averaged mean-centered expression values for both signatures, **a** for the Staub colorectal cancer data set, **b** for the Wang breast cancer data set, and **c** for the Phillips glioma expression data set

Results

Discovery and validation of a WIPF1 coexpression module

We used microarray data from a study of 100 colorectal cancer specimens of mixed stage and subtype by Kaiser et al. for the identification of WIPF1 correlated genes (Gene Expression Omnibus (GEO) database identifier GSE5206) [17]. We correlated the expression profiles of the WIPF1 gene across all experiments to the expression profiles of each other gene that passed initial quality criteria. A histogram of these p values is shown in Fig. 1a. The strong deviation from a uniform distribution suggests that a large number of genes are correlated with WIPF1 expression in this data set. A threshold for significantly correlated expression was set at $p < 1e-6$ resulting in 430 candidate genes. For validation of the WIPF1 coexpression of these genes, we employed a data set of Ayers et al. (GSE7208) on 59 colorectal tumors [20]. We correlated the expression patterns of the 430 discovered genes to the WIPF1 profile. This led to the p value distribution shown in Fig. 1b that exhibits an even stronger deviation from the uniform distribution than for the discovery gene set, thus documenting the enrichment of true WIPF1 coexpressed genes. Genes, 112 out of 430, were correlated with $p < 1e-5$. A scatter plot of correlation coefficients in Ayers and Aronow data (Fig. 1c) confirmed that for all 112 genes, the correlation coefficients have the same sign, meaning that expression changes in the data sets have the same direction. In the following, these 112 genes were considered as the WIPF1 coexpression module, and their associated expression profiles constitute the WIPF1 signature.

Prediction of low-risk colorectal cancers using the WIPF1 signature

We fused the discovery and validation data of the WIPF1 module resulting in a data set with 112 genes and 159 patients. Then, we performed hierarchical clustering on both, genes and patients (see Fig. 2). It revealed two major clusters of genes and five major clusters of patients. Patient cluster "A" (Fig. 2, marked in red) comprises 25 tumors

Table 2 Summaries of top literature subnetworks enriched with genes from the WIPF1 coexpression module

No	General Molecular Network	GO Processes	Total nodes	Root nodes	<i>p</i> -Value
1	ADAM19, SLC25A10, CDC14a, UBE2E3, TXNIP (VDUP1), ...	Sulfate transport (8.1%; 7.943e-06), cell division (18.9%; 1.334e-05), mitosis (16.2%; 1.414e-05), M phase of mitotic cell cycle (16.2%; 1.598e-05), M phase (18.9%; 1.854e-05)	50	13	4.38E-26
2	REA, NLK, Chordin-like 1, Copine-1, ...	BMP signaling pathway (11.6%; 8.328e-08), positive regulation of osteoblast differentiation (9.3%; 1.255e-06), regulation of osteoblast differentiation (9.3%; 7.241e-06), transmembrane receptor protein serine/threonine kinase signaling pathway (11.6%; 1.533e-05), developmental process (67.4%; 2.095e-05)	50	12	1.49E-23
3	Neurofibromin, TXNIP (VDUP1), REA, DEDD, DEDD2, ...	Regulation of apoptosis (45.5%; 2.153e-13), regulation of programmed cell death (45.5%; 2.710e-13), regulation of developmental process (52.3%; 2.099e-12), Ras protein signal transduction (22.7%; 1.044e-11), negative regulation of cellular process (54.5%; 1.944e-11)	50	10	9.99E-19
No	Transcriptional Regulation Network	GO Processes	Total nodes	Root nodes	<i>p</i> -Value
4	c-Myc	Positive regulation of mitotic cell cycle (25.0%; 1.015e-05), cell cycle (62.5%; 3.803e-05), regulation of mitotic cell cycle (37.5%; 5.661e-05), regulation of cell cycle (50.0%; 5.920e-05), positive regulation of cell cycle (25.0%; 3.432e-04)	9	8	4.02E-22
5	ESR1	Response to hormone stimulus (57.1%; 2.915e-05), response to endogenous stimulus (57.1%; 3.184e-05), response to organic nitrogen (28.6%; 1.001e-04), response to steroid hormone stimulus (42.9%; 2.113e-04), negative regulation of hydrolase activity (28.6%; 2.580e-04)	8	7	2.31E-19
6	p53	Response to organic nitrogen (28.6%; 1.001e-04), positive regulation of cell cycle (28.6%; 2.580e-04), regulation of apoptosis (57.1%; 4.295e-04), nucleic acid-protein covalent cross-linking (14.3%; 4.446e-04), RNA-protein covalent cross-linking (14.3%; 4.446e-04)	8	7	2.31E-19

Here we show information about the top literature subnetworks with significant enrichment for genes of the WIPF1 module using two modes of analysis in the metacore software. Networks 1, 2, and 3 were derived using a large literature network considering all types of molecular interactions. Networks 4, 5, and 6 resulted from an enrichment analysis of subnetworks centered around transcription factors. Column “GO processes” shows Gene Ontology (GO) categories that are enriched in a subnetwork. Significance of enrichment of gene groups (be it WIPF1 coexpressed genes or genes associated with GO categories) in subnetworks was assessed using hypergeometric tests

that exhibit a characteristic down-regulation of the larger gene cluster. In contrast, other tumor clusters (collectively labeled as “B”) show an up-regulation of the larger gene cluster or indifferent expression. Using patient cluster labels as the prediction target, we trained a *k*-nearest neighbor classifier (*k*=9) for prediction of WIPF1 cluster labels in further independent data sets.

We were interested in the question whether the cluster labels can predict certain clinical parameters in an independent colorectal cancer data set. We determined expression profiles

for 62 colorectal cancer patients for which information about survival has been tracked. Direct application of the *k*-nearest neighbor classifier (trained on gene-wise mean-centered combined Aronow/Ayers data) on the 62 tumor expression profiles led to the assignment of cluster label “A” for nine patients and cluster label “B” for the remaining 53 patients (for a more detailed explanation/justification of the application of the *k*-nearest neighbor classifier across data sets we refer to the “Materials and methods” section). We compared the survival expectancy of the three groups using Kaplan–

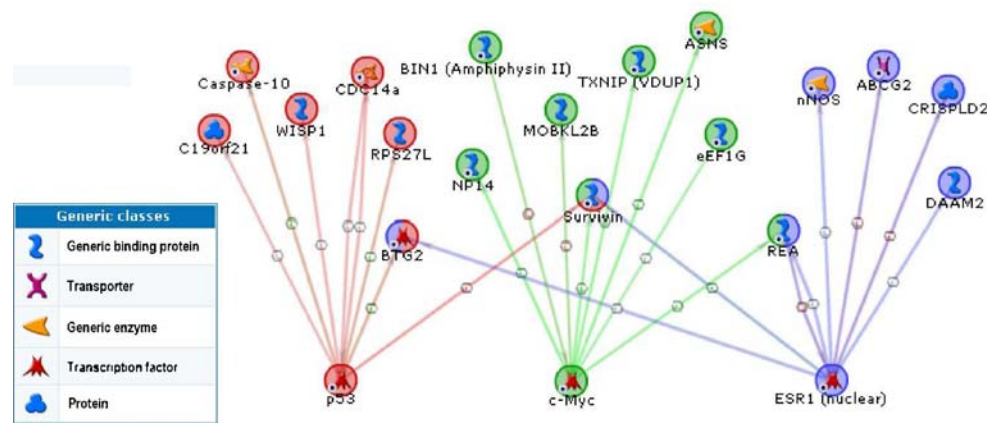


Fig. 6 Fused transcriptional regulation network of genes of the WIPF1 module that are regulated by c-myc, ESR1 or p53. Here we show the fused network of WIPF1 co-expressed genes that are linked to c-myc, ESR1 or p53 according to literature evidence. The legend on

the left depicts the graphical symbols that describe the type of protein in the network. The *arrows* indicate the direction of the regulation. The *colors of circles* around gene names represent the coupling to transcription factors: *green* for c-myc, *red* for p53, *blue* for ESR1

Meier plots (see Fig. 3). We found that all nine cluster “A” colorectal cancer patients have survived until the end of the study. The differences in the survival curves between clusters “A” and “B” were already significant ($p=0.011$), although the numbers of patients are rather small. These results suggest that the WIPF1 expression module in combination with the applied scheme of normalization/classification also has potential for the prediction of low-risk cancers in further independent studies.

The WIPF1 signature has prognostic predictivity for brain and breast tumors

We fitted Cox regression models for genes of the WIPF1 signature and tested their individual association with survival using Wald tests on microarray expression studies of 286 breast cancers (Wang et al., GSE2034) and 77 high grade-gliomas (Phillips et al., GSE4271). We found that a high fraction (50 of 112) of genes of the WIPF1 signature were significantly associated with survival ($p<0.05$) in both studies. A list of 38 genes for which significant correlation of WIPF1 expression across patients was validated in Aronow, Ayers, and Staub data sets is given in Table 1 together with results of statistical tests on WIPF1 correlation and survival/relapse association.

We next applied the k -nearest neighbor classifier trained on Aronow/Ayers data directly on the data sets of breast cancers and high-grade gliomas to stratify these patient collectives. Among the microarray profiles of 286 breast tumor patients of a data set of Wang et al. (GSE2034), we identified 59 of the cluster “A” type. These patients had a remarkably better prognosis regarding time to relapse, the difference between cluster “A” patients, and others being significant at $p=0.0045$ (see Fig. 4a and b). In the glioma

patient set of Phillips et al. (GSE4271), 24 out of 77 patients were assigned the cluster “A” type by classification. The cluster “A” glioma patients had a better prognosis ($p=0.012$, see Fig. 4c and d). In conclusion, the application of the WIPF1 k -nearest neighbor classifier on two different tumor sets further validates the predictive value of the WIPF1 coexpression module for the identification of patients with favorable prognosis.

The WIPF1 signature has links to c-myc, p53, proliferation, and apoptosis

Because of the multitude of genes that are correlated with WIPF1 (see Fig. 1), we hypothesized that a dominating cancer-driving mechanism is linked to WIPF1 expression. It was frequently reported that patient prognosis is associated with the expression of a so-called *proliferation* signature. This term circumscribes a large set of signatures that can be pulled from a large set of genes deregulated during the cell cycle. To our knowledge, a link between the proliferation signature and colorectal cancer has not been established so far. We found that the WIPF1 signature does not have a substantial overlap to proliferation signatures. To investigate an indirect link to the transcriptional program of proliferation, we tested the proliferation signature described by Rosenwald et al. for correlation with our WIPF1 signature [36]. Correlation tests of signature centroids revealed that both signatures are significantly correlated in three data sets, the colorectal cancer data generated during this study, the Wang data on breast cancer, and the Phillips data on gliomas (see Fig. 5). This suggests a regulatory link of the WIPF1 signature with typical proliferation genes.

We next addressed the question which functional themes are dominating the list of 112 WIPF1-coexpressed genes.

We investigated the genes of the WIPF1 module by mining of literature networks using the metacore software. We identified several gene networks that were highly enriched with genes of the WIPF1 module (see Table 2) and are therefore suited to highlight the function of the WIPF1 module. First, we took all different types of functional links into account (protein–protein interactions, enzymatic reactions, transcriptional regulation, etc.) for building of literature subnetworks. We identified three subnetworks that are enriched with gene functions in cancer-relevant processes like mitosis and apoptosis as important themes among WIPF1 signature-induced networks (see Table 2). By focusing the analysis on literature networks centered at transcription factors, three subnetworks stood out with 21 genes linked via the oncogenic transcription factors c-myc, ESR1, and p53 (see Table 2 and Fig. 6). For these three factors, there are 17 genes with literature evidence that c-myc, ESR1, and/or p53 regulate their transcription. Thus, literature network analysis could link the WIPF1 signature to several functional themes related to cancer biology with high significance: mitosis and proliferation, apoptosis, the estrogen receptor, c-myc, and p53.

Discussion

The hypothesis that the WIPF1 gene is important for cancer development was based on two facts. First, its encoded WIP protein interacts with the Wiskott–Aldrich syndrome protein WASP through a surface that is affected by a WASP mutation, and Wiskott–Aldrich syndrome predisposes to leukemia and lymphoma. Second, expression levels of the WIP protein influence the migratory and differentiation properties of fibroblasts. It has not been studied so far how the expression of the WIPF1 gene is regulated and if its coexpression neighborhood provides an additional link to cancer. Here, we show that there exists a module of genes that is coexpressed with WIPF1 in colorectal cancers. The majority of genes in this module show a characteristic down-regulation in several cancer patients with longer survival time or time to relapse, also in other cancer types than colorectal cancer. We found that the module genes do not overlap with the frequently rediscovered “proliferation” signature that is regulated during cancer cell mitosis. Instead, a large number of genes of the WIPF1 coexpression module have poorly characterized functions. Only single genes link directly to cancer-relevant processes like proliferation and apoptosis. However, we could show that the expression profile of the WIPF1 signature correlates significantly with expression profile of the Rosenwald proliferation signature. Literature networks revealed that the link of the WIPF1 module to proliferation can partly be explained by the fact that a large fraction of WIPF1 module

genes are known transcriptional targets of oncogenic transcription factors like c-myc, ESR1, or p53. In this context, it is interesting to note that estrogen receptor signaling is not only of importance for breast cancers, but also apparently able to modulate the aggressiveness of prostate cancers [38]. It is tempting to hypothesize that keeping the expression of the WIPF1 module in a low expression state is causative for a less aggressive cancer phenotype, e.g., by inhibition of WIPF1/WASP-related cytoskeletal remodeling that coincides with a reduced ability of cells to migrate and metastasize.

In conclusion, we presented a module of WIPF1-coexpressed genes. The expression signature of this module could be used to identify patients with better prognosis with respect to relapse or survival in expression data sets of three different tumor types, colorectal cancer, breast cancer, and high-grade glioma. The WIPF1 coexpressed genes seem to be linked to proliferation and apoptosis possibly by regulation through c-myc, ESR1, and p53. We propose the WIPF1 signature as an alternative predictor of breast, brain, and colorectal cancer prognosis.

Acknowledgements We thank Anja von Heydebreck for critical reading of the manuscript and valuable comments.

Disclosure of potential conflict of interests The authors declare that they have no conflicting interests related to this study.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Moreau V, Frischknecht F, Reckmann I et al (2002) A complex of N-WASP and WIP integrates signalling cascades that lead to actin polymerization. *Nat Cell Biol* 2:441–448
- Ramesh N, Antón IM, Hartwig JH et al (1997) WIP, a protein associated with Wiskott–Aldrich syndrome protein, induces actin polymerization and redistribution in lymphoid cells. *Proc Natl Acad Sci U S A* 94:14671–14676
- Volkman BF, Prehoda KE, Scott JA et al (2002) Structure of the N-WASP EVH1 domain-WIP complex: insight into the molecular basis of Wiskott–Aldrich Syndrome. *Cell* 111:565–576
- Peterson FC, Deng Q, Zetl M et al (2007) Multiple WASP-interacting protein recognition motifs are required for a functional interaction with N-WASP. *J Biol Chem* 282:8446–8453
- Curcio C, Pannellini T, Lanzardo S et al (2007) WIP null mice display a progressive immunological disorder that resembles Wiskott–Aldrich syndrome. *J Pathol* 211:67–75
- de la Fuente MA, Sasahara Y, Calamito M et al (2007) WIP is a chaperone for Wiskott–Aldrich syndrome protein (WASP). *Proc Natl Acad Sci U S A* 104:926–931
- Krzewski K, Chen X, Strominger JL (2008) WIP is essential for lytic granule polarization and NK cell cytotoxicity. *Proc Natl Acad Sci U S A* 105:2568–2573

8. Massarwa R, Carmon S, Shilo B et al (2007) WIP/WASp-based actin-polymerization machinery is essential for myoblast fusion in *Drosophila*. *Dev Cell* 12:557–569
9. Tsuboi S (2007) Requirement for a complex of Wiskott-Aldrich syndrome protein (WASP) with WASP interacting protein in podosome formation in macrophages. *J Immunol* 178:2987–2995
10. Lanzardo S, Curcio C, Forni G et al (2007) A role for WASP Interacting Protein, WIP, in fibroblast adhesion, spreading and migration. *Int J Biochem Cell Biol* 39:262–274
11. Alon U, Barkai N, Notterman DA et al (1999) Broad patterns of gene expression revealed by clustering analysis of tumour and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A* 96:6745–6750
12. Tsafirir D, Bacolod M, Selvanayagam Z et al (2006) Relationship of gene expression and chromosomal abnormalities in colorectal cancer. *Cancer Res* 66:2129–2137
13. Staub E, Groene J, Heinze M et al (2007) Genome-wide expression patterns of invasion front, inner tumour mass and surrounding normal epithelium of colorectal tumours. *Mol Cancer* 6:79
14. Staub E, Grone J, Mennerich D et al (2006) A genome-wide map of aberrantly expressed chromosomal islands in colorectal cancer. *Mol Cancer* 5:37
15. Reichling T, Goss KH, Carson DJ et al (2005) Transcriptional profiles of intestinal tumours in Apc(Min) mice are unique from those of embryonic intestine and identify novel gene targets dysregulated in human colorectal tumours. *Cancer Res* 65:166–176
16. Notterman DA, Alon U, Sierk AJ et al (2001) Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Res* 61:3124–3130
17. Kaiser S, Park Y, Franklin JL et al (2007) Transcriptional recapitulation and subversion of embryonic colon development by mouse colon tumour models and human colon cancer. *Genome Biol* 8:R131
18. Kitahara O, Furukawa Y, Tanaka T et al (2001) Alterations of gene expression during colorectal carcinogenesis revealed by cDNA microarrays after laser-capture microdissection of tumour tissues and normal epithelia. *Cancer Res* 61:3544–3549
19. Inoue Y, Shirane M, Miki C et al (2004) Gene expression profiles of colorectal carcinoma in response to neo-adjuvant chemotherapy. *Int J Oncol* 25:1641–1649
20. Ayers M, Fargnoli J, Lewin A, Wu Q, Platero JS (2007) Discovery and validation of biomarkers that respond to treatment with brivanib alaninate, a small-molecule VEGFR-2/FGFR-1 antagonist. *Cancer Res* 67:6899–6906
21. Mariadason JM, Arango D, Shi Q et al (2003) Gene expression profiling-based prediction of response of colon carcinoma cells to 5-fluorouracil and camptothecin. *Cancer Res* 63:8791–8812
22. Ghadimi BM, Grade M, Difilippantonio MJ et al (2005) Effectiveness of gene expression profiling for response prediction of rectal adenocarcinomas to preoperative chemoradiotherapy. *J Clin Oncol* 23:1826–1838
23. Arango D, Laiho P, Kokko A et al (2005) Gene-expression profiling predicts recurrence in Dukes' C colorectal cancer. *Gastroenterology* 129:874–884
24. Groene J, Mansmann U, Meister R et al (2006) Transcriptional census of 36 microdissected colorectal cancers yields a gene signature to distinguish UICC II and III. *Int J Cancer* 119:1829–1836
25. Friederichs J, Rosenberg R, Mages J et al (2005) Gene expression profiles of different clinical stages of colorectal carcinoma: toward a molecular genetic understanding of tumour progression. *Int J Colorectal Dis* 20:391–402
26. Frederiksen CM, Knudsen S, Laurberg S et al (2003) Classification of Dukes' B and C colorectal cancers using expression arrays. *J Cancer Res Clin Oncol* 129:263–271
27. Bertucci F, Salas S, Eysteris S et al (2004) Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters. *Oncogene* 23:1377–1391
28. Grade M, Hörmann P, Becker S et al (2007) Gene expression profiling reveals a massive, aneuploidy-dependent transcriptional deregulation and distinct differences between lymph node-negative and lymph node-positive colon carcinomas. *Cancer Res* 67:41–56
29. Croner RS, Peters A, Brueckl WM et al (2005) Microarray versus conventional prediction of lymph node metastasis in colorectal carcinoma. *Cancer* 104:395–404
30. Croner RS, Förtsch T, Brückl WM et al (2008) Molecular signature for lymphatic metastasis in colorectal carcinomas. *Ann Surg* 247:803–810
31. Ki DH, Jeung H, Park CH et al (2007) Whole genome analysis for liver metastasis gene signatures in colorectal cancer. *Int J Cancer* 121:2005–2012
32. Yamasaki M, Takemasa I, Komori T et al (2007) The gene expression profile represents the molecular nature of liver metastasis in colorectal cancer. *Int J Oncol* 30:129–138
33. Alldinger I, Dittert D, Peiper M et al (2005) Gene expression analysis of pancreatic cell lines reveals genes overexpressed in pancreatic cancer. *Pancreatol* 5:370–379
34. Wang Y, Klijn JGM, Zhang Y et al (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 365:671–679
35. Phillips HS, Kharbanda S, Chen R et al (2006) Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* 9:157–173
36. Rosenwald A, Wright G, Wiestner A et al (2003) The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell* 3:185–197
37. Lusa L, McShane LM, Reid JF et al (2007) Challenges in projecting clustering results across gene expression-profiling datasets. *J Natl Cancer Inst* 99(22):1715–1723
38. Setlur SR, Mertz KD, Hoshida Y et al (2008) Estrogen-dependent signaling in a molecularly distinct subclass of aggressive prostate cancer. *J Natl Cancer Inst* 100(11):815–825