THE LANCET Oncology

Supplementary webappendix

This webappendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Vermeulen J, Preter KD, Naranjo A, et al. Predicting outcomes for children with neuroblastoma using a multigene-expression signature: a retrospective SIOPEN/COG/GPOH study. *Lancet Oncol* 2009; published online June 9, 2009. DOI:10.1016/S1470-2045(09)70154-8.

Supplemental material and data:

1. Gene prioritization for inclusion in a robust prognostic signature

In order to prioritize robust prognostic marker genes we applied a unique re-analysis approach (unpublished data; De Preter et al.).

Four large[1–4] (including more than 90 patients per study) and three small[5–7] (including fewer than 90 patients per study) published studies were used for selecting the robust prognostic markers. Datasets were either downloaded from the NCBI GEO database[2,4] (GSE2283 and GSE3960), or from the EBI ArrayExpress database[1,6,7] (E-TABM-38, E-MEXP-669 and E-MEXP-83), or from the authors' website[3] (http://www.imt.uni-marburg.de/microarray/download.html).

In order to make the data from the different micro-array platforms maximally comparable, annotation information of the probes was updated using the MatchMiner tool[8] for the custom made cDNA or oligonucleotide arrays[1,3,4] and using the latest version of the R packages hgu95av2 and hgu133plus2 for the Affymetrix array data[2,5–7]. Probe identification numbers were converted into gene symbols to enable straightforward comparison of the gene lists between the different studies.

Updated clinical information for the patients were obtained upon request from the authors[2–5,7] or were available[1,6]. For the Berwanger and Ohira studies, only overall survival data were available.

In order to train a prognostic model with robust prognostic markers patients from the different studies were selected and divided in 2 clearly defined uniform risk groups. The low-risk subgroup was defined by INSS stage 1, 2 or 4S without *MYCN* amplification with progression-free survival time[1,2,5–7] (or overall survival time[3,4]) of at least 1000 days and the high-risk subgroup comprised patients older than 12 months at diagnosis with INSS stage 4 tumours (irrespective of *MYCN* status) or with INSS stage 2 and 3 tumours with *MYCN* amplification that died from disease.

A complete 10-times-repeated 10-fold cross validation using the PAM algorithm[1,9] was performed on the patients belonging to one of the two clearly defined risk groups from the four larger studies separately, in order to identify robust prognostic markers. For each dataset we selected the probes that were included in at least 65 of the 100 cross validation gene lists as these genes are likely to be the ones with the highest prognostic value[1]. For the three smaller datasets, prognostic genes were selected by a single PAM analysis. The resulting prognostic gene lists from the seven studies showed substantial overlap.

In addition to this re-analysis we performed an extensive literature screening for single candidate prognostic genes (~800 abstracts).

Finally, we composed a list of 59 prognostic genes that were independently identified in at least two of the seven prognostic gene sets or literature gene list (Supplemental Table 1). The occurrence of the 59 genes in at least 2 of the 7 prognostic gene lists or literature list makes them robust platform independent, prognostic markers.

2. RNA extraction and amplification

Total RNA extraction of NB tumour samples was performed in different national reference laboratories by silica gel-based membrane purification methods (RNeasy Mini kit or MicroRNeasy kit, Qiagen), by phenol-based (TRIzol reagent, Invitrogen, or Tri Reagent product, Sigma), or by chaotropic solution-based isolation methods (Perfect Eukaryotic RNA kit, Eppendorf) according to the manufacturer's instructions.

Starting from 20 ng of total RNA, a validated sample pre-amplification method was applied, generating approximately 6 µg of cDNA, sufficient to measure more than 1000 target genes (WT-Ovation, NuGEN).

3. Assessment of RNA purity and integrity

In order to assess the RNA quality of the 711 collected tumour samples, we used 30 ng of each RNA isolate to perform two PCR-based assays (5'-3' mRNA integrity assay to determine a 5'-3'-delta-Cq, and a SPUD assay for the detection of enzymatic inhibitors in nucleic acid preparations[10]) and a capillary gel electrophoresis analysis (high sensitivity chips, Experion, Bio-Rad) to establish an RNA quality index (RQI). Based on these tests, we retained approximately 80% of samples (579/711) with acceptable quality[11] (RQI > 5 and absence of enzymatic inhibitors). Impact of RNA quality on performance will be published elsewhere (unpublished data; Vermeulen et al.).

4. High-throughput real-time quantitative PCR based gene expression

A real-time quantitative polymerase chain reaction (qPCR) assay was designed for each of the 59 prognostic genes and five reference genes by PrimerDesign and went through an extensive in silico validated analysis using BLAST and BiSearch specificity, amplicon secondary structure, SNP presence, and splice variant analysis.[12] The mean amplification efficiency was 95% (± 4%) (Supplemental Data, Vermeulen.rdml).

Real time qPCR was performed on a high throughput 384-well plate instrument (LC480, Roche). PCR plates were prepared using a 96-well head pipetting robot (Sciclone ALH 3000, Calliper). Real time qPCR amplifications were performed in 8 µl containing 4 µl 2X SYBR Green I master mix (Roche), 0·4 µl forward and reverse primer (5 µM each), 0·2 µl nuclease-free water, and 3 µl cDNA (corresponding to 4·5 ng unamplified cDNA). The cycling conditions were comprised of 3 min polymerase activation at 95 °C and 40 cycles of 15 s at 95 °C and 30 s at 60°C, followed by a dissociation curve analysis from 60 °C to 95 °C. To detect and correct

inter-run variation and allow data comparison with different labs, we included a dilution series of absolute standards consisting of 55 bp oligonucleotides (Biolegio) run in parallel with patient samples (Supplemental Data, Vermeulen.rdml).

For data pre-processing, all samples without signal for a particular gene were set to the minimum Cq of the gene across all samples. The Cq values were converted to relative quantities and converted to $\log_2$ values. Relative gene expression levels were then normalised using the geometric mean of five reference sequences (*HPRT1*, *SDHA*, *UBC*, *HMBS*, and *AluSq*).[13] Data handling and calculations (normalization, rescaling, inter-run calibration, and error propagation) were done in qBase*Plus* version 1·1 (http://www.qbaseplus.com)[14] (Supplemental Data, Vermeulen.rdml).

5. Multigene expression signature

For establishment of the multigene expression classifier, the SIOPEN tumour samples were divided into a training set and a test set. The training set was comprised of 30 samples from two patient subgroups with maximally divergent clinical courses randomly selected: 15 low-risk patients with INSS stage 1, 2, or 4S without *MYCN* amplification and with progression-free survival time of at least 1000 days (treated according to the INES NB99·s (n=9), INES NB99·2 (n=3), or LNESG1 (n=3) protocols) and 15 deceased high-risk patients older than 12 months at diagnosis with INSS stage 4 tumour (irrespective of the *MYCN* gene status) or with INSS stage 2 or 3 tumour with *MYCN* amplification (treated according to the HR-NBL1 protocol). The multigene expression signature was built for these 30 training samples using the Prediction Analysis of Microarrays (PAM) method[9] in the R statistical language (Bioconductor package MCRestimate[15]). This analysis resulted in a classifier of which the expression levels best characterized each risk group enabling class prediction of test patients. Important to note is that this phase of the study did not represent a classical training. Gene selection was performed previously on nearly 700 cases (unpublished data; De Preter et al.). The 30 training patients were only used to establish the classifier, after which it could be tested on the remaining SIOPEN samples and validated in a blind study on the COG samples.

6. Case-control study on the COG cohort

The COG sampling was initially spiked to include more events because random sampling would not have provided sufficient power because there would have been too few events. Due to the use of this sampling technique that generated a non-representative cohort of the general neuroblastoma population in terms of outcome, Kaplan-Meier analyses and Cox proportional hazards models would have yielded biased survival results. Multivariate logistic regression analyses were performed instead to determine if the multigene expression signature was a significant independent predictor after controlling for currently used risk factors. A case-control study was set up where case was defined as failure (relapse, progression, or death of disease for PFS, and death for OS) prior to two years and control as non-failure prior to two years in patients with at least two years of follow-up. Controls and cases were selected two to one to increase the sample size and power. From the available controls with complete data, two controls were selected at random for each case with complete data. The final logistic regression model used 139 controls and 70 cases with complete data for PFS; for OS, each individual model was fit with complete data for 74 controls for 37 cases. For this type of analysis, logistic regression does not lead to biased coefficient estimates.[16]

Known risk factors age, INSS stage, *MYCN* status, and ploidy were tested using a backward selection approach, and variables with $p<0{\cdot}05$ were retained in the model. Instead of Shimada histology, underlying pathological components MKI and grade were tested in order to avoid confounding with age.[17]

Progression-free survival (PFS) time was calculated from the date of diagnosis until the first occurrence of tumour progression/relapse or the date of last follow-up if no event occurred. Overall survival (OS) was calculated from the date of diagnosis to death or date of last follow-up if the patient survived. Deaths due to toxicity were censored.

P-values < 0·05 were deemed statistically significant.

7. RDML file

The Real-time PCR Data Markup Language (RDML) is a structured and universal data standard for exchanging quantitative PCR (qPCR) data (http://www.rdml.org)[18]. The Vermeulen.rdml file can be downloaded from http://medgen.ugent.be/jvermeulen.

References

1 Oberthuer A, Berthold F, Warnat P, et al. Customized oligonucleotide microarray gene expression-based classification of neuroblastoma patients outperforms current clinical risk stratification. *J Clin Oncol* 2006; **24:** 5070–8.

2 Wang Q, Diskin S, Rappaport E, et al. Integrative genomics identifies distinct molecular classes of neuroblastoma and shows that multiple genes are targeted by regional alterations in DNA copy number. *Cancer Res* 2006; **66:** 6050–62.

3 Berwanger B, Hartmann O, Bergmann E, et al. Loss of a FYN-regulated differentiation and growth arrest pathway in advanced stage neuroblastoma. *Cancer Cell* 2002; **2:** 377–86.

4 Ohira M, Oba S, Nakamura Y, et al. Expression profiling using a tumor-specific cDNA microarray predicts the prognosis of intermediate risk neuroblastomas. *Cancer Cell* 2005; **7:** 337–50.
5 Schramm A, Schulte JH, Klein-Hitpass L, et al. Prediction of clinical outcome and biological characterization of neuroblastoma by expression profiling. *Oncogene* 2005; **24:** 7902–12.
6 De Preter K, Vandesompele J, Heimann P, et al. Human fetal neuroblast and neuroblastoma transcriptome analysis confirms neuroblast origin and highlights neuroblastoma candidate genes. *Genome Biol* 2006; **7:** R84.
7 McArdle L, McDermott M, Purcell R, et al. Oligonucleotide microarray analysis of gene expression in neuroblastoma displaying loss of chromosome 11q. *Carcinogenesi*s 2004; **25:** 1599–609.
8 Bussey KJ, Kane D, Sunshine M, et al. MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biol* 2003; **4:** R27.
9 Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A* 2002; **99:** 6567–72.
10 Nolan T, Hands RE, Ogunkolade W, Bustin SA. SPUD: a quantitative PCR assay for the detection of inhibitors in nucleic acid preparations. *Anal Biochem* 2006; **351:** 308–10.
11 Fleige S, Pfaffl MW. RNA integrity and the effect on the real-time qRT-PCR performance. *Mol Aspects Med* 2006; **27:** 126–39.
12 Lefever S, Vandesompele J, Speleman F, Pattyn F. RTPrimerDB: the portal for real-time PCR primers and probes. *Nucleic Acids Res* 2009; **37:** D942–5.
13 Vandesompele J, De Preter K, Pattyn F, et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* 2002; **3:** RESEARCH0034.
14 Hellemans J, Mortier G, De Paepe A, Speleman F, Vandesompele J. qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biol* 2007; **8:** R19.
15 Ruschhaupt M, Huber W, Poustka A, Mansmann U. A compendium to ensure computational reproducibility in high-dimensional classification tasks. *Stat Appl Genet Mol Biol* 2004; **3:** Article37.
16 Allison P (1995). *Survival analysis using the SAS system: a practical guide.* SAS Institute Inc, Cary, NC.
17 London WB, Shimada H, d'Amore E, et al. Age, tumor grade, and MKI are independently predictive of outcome in neuroblastoma. Paper presented at the American Society of Clinical Oncology Annual Meeting, Chicago, June 2007.
18 Lefever S, Hellemans J, Pattyn F, et al. RDML: structured language and reporting guidelines for real-time quantitative PCR data. *Nucleic Acids Res* 2009; **37:** 2065-9.

Supplemental Table 1: Prognostic gene signature content

symbol	accession number	sequence definition	chromosomal position	higher expressed in HR or LR tumours
AHCY	NM_000687	S-adenosylhomocysteine hydrolase (AHCY), mRNA.	20cen-q13.1	HR
AKR1C1	NM_001353	aldo-keto reductase family 1, member C1 (dihydrodiol dehydrogenase 1; 20-alpha (3-alpha)-hydroxysteroid dehydrogenase) (AKR1C1), mRNA.	10p15-p14	LR
ARHGEF7	NM_145735	Rho guanine nucleotide exchange factor (GEF) 7 (ARHGEF7), transcript variant 2, mRNA.	13q34	LR
BIRC5	NM_001168	baculoviral IAP repeat-containing 5 (survivin) (BIRC5), transcript variant 1, mRNA.	17q25	HR
CADM1	NM_014333	cell adhesion molecule 1	11q23.2	LR
CAMTA1	NM_015215	calmodulin binding transcription activator 1 (CAMTA1), mRNA.	1p36.31-23	LR
CAMTA2	NM_015099	calmodulin binding transcription activator 2 (CAMTA2), mRNA.	17p13.2	LR
CD44	NM_001001392	CD44 molecule (Indian blood group)	11p13	LR
CDCA5	NM_080668	cell division cycle associated 5 (CDCA5), mRNA.	11q12.1	HR
CDKN3	NM_005192	cyclin-dependent kinase inhibitor 3	14q22	HR
CHD5	NM_015557	chromodomain helicase DNA binding protein 5 (CHD5), mRNA.	1p36.31	LR
CLSTN1	NM_001009566	calsyntenin 1 (CLSTN1), transcript variant 1, mRNA.	1p36.22	LR
CPSG3	NM_004386	chondroitin sulfate proteoglycan 3 (neurocan) (CSPG3)(NCAN), mRNA.	19p12	HR
DDC	NM_000790	dopa decarboxylase (aromatic L-amino acid decarboxylase)	7p11	LR
DPYSL3	NM_001387	dihydropyrimidinase-like 3 (DPYSL3), mRNA.	5q32	LR
ECEL1	NM_004826	endothelin converting enzyme-like 1 (ECEL1), mRNA.	2q36-q37	LR
ELAVL4	NM_021952	Embryonic lethal, abnormal vision, Drosophila)-like 4 (Hu antigen D)	1p34	LR
EPB41L3	NM_012307	erythrocyte membrane protein band 4.1-like 3 (EPB41L3), mRNA.	18p11.32	LR
EPHA5	NM_004439	EPH receptor A5 (EPHA5), transcript variant 1, mRNA.	4q13.1	LR
EPN2	NM_014964	epsin 2 (EPN2), transcript variant 2, mRNA.	17p11.2	LR
FYN	NM_153048	proto-oncogene tyrosine-protein kinase fyn	6q21	LR
GNB1	NM_002074	guanine nucleotide binding protein (G protein), beta polypeptide 1 (GNB1), mRNA.	1p36.33	LR
HIVEP2	NM_006734	human immunodeficiency virus type I enhancer binding protein 2 (HIVEP2), mRNA.	6q23-q24	LR
INPP1	NM_002194	inositol polyphosphate-1-phosphatase (INPP1), mRNA.	6q22-q23	LR
MAP2K4	NM_003010	mitogen-activated protein kinase kinase 4 (MAP2K4), mRNA.	17p11.2	LR
MAP7	NM_003980	microtubule-associated protein 7 (MAP7), mRNA.	6q23.3	LR
MAPT	NM_016835	microtubule-associated protein tau (MAPT), transcript variant 1, mRNA.	17q21.1	LR
MCM2	NM_004526	MCM2 minichromosome maintenance deficient 2, mitotin (S. cerevisiae) (MCM2), mRNA.	3q21	HR
MRPL3	NM_007208	mitochondrial ribosomal protein L3 (MRPL3), nuclear gene encoding mitochondrial protein, mRNA.	3q21-q23	HR
MTSS1	NM_014751	metastasis suppressor 1 (MTSS1), mRNA.	8p22	LR
MYCN	NM_005378	v-myc myelocytomatosis viral related oncogene, neuroblastoma derived	2p24.1	HR
NHLH2	NM_005599	nescient helix loop helix 2 (NHLH2), mRNA.	1p12-p11	HR
NME1	NM_198175	non-metastatic cells 1, protein (NM23A) expressed in (NME1), transcript variant 1, mRNA.	17q21.3	HR

NRCAM	NM_001037132	neuronal cell adhesion molecule (NRCAM), transcript variant 1, mRNA.	7q31.1-q31.2	LR
NTRK1	NM_001012331	Neurotrophic tyrosine kinase receptor type 1	1q21-q22	LR
ODC1	NM_002539	ornithine decarboxylase 1 (ODC1), mRNA.	2p25	HR
PAICS	NM_001079525.1	phosphoribosylaminoimidazole carboxylase, phosphoribosylaminoimidazole succinocarboxamide synthetase	4q12	HR
PDE4DIP	NM_014644	phosphodiesterase 4D interacting protein (myomegalin) (PDE4DIP), transcript variant 1, mRNA.	1q12	LR
PIK3R1	NM_181523	phosphoinositide-3-kinase, regulatory subunit 1 (p85 alpha) (PIK3R1), transcript variant 1, mRNA.	5q13.1	LR
PLAGL1	NM_002656	pleiomorphic adenoma gene-like 1 (PLAGL1), transcript variant 1, mRNA.	6q24-q25	LR
PLAT	NM_033011	plasminogen activator, tissue	8p12	LR
PMP22	NM_000304	peripheral myelin protein 22 (PMP22), transcript variant 1, mRNA.	17p12-p11.2	LR
PRAME	NM_006115	preferentially expressed antigen in melanoma	22q11.22	HR
PRDM2	NM_012231	PR domain containing 2, with ZNF domain (PRDM2), transcript variant 1, mRNA.	1p36.21	LR
PRKACB	NM_182948	protein kinase, cAMP-dependent, catalytic, beta	1p36.1	LR
PRKCZ	NM_002744	protein kinase C, zeta (PRKCZ), transcript variant 1, mRNA.	1p36.33-p36.2	LR
PTN	NM_002825	pleiotrophin (heparin binding growth factor 8, neurite growth-promoting factor 1) (PTN), mRNA.	7q33-q34	LR
PTPRF	NM_002840	protein tyrosine phosphatase, receptor type, F (PTPRF), transcript variant 1, mRNA.	1p34	LR
PTPRH	NM_002842	protein tyrosine phosphatase, receptor type, H (PTPRH), mRNA.	19q13.4	LR
PTPRN2	NM_002847	protein tyrosine phosphatase, receptor type, N polypeptide 2 (PTPRN2), transcript variant 1, mRNA.	7q36	LR
QPCT	NM_012413	glutaminyl-peptide cyclotransferase (glutaminyl cyclase) (QPCT), mRNA.	2p22.2	LR
SCG2	NM_003469	secretogranin II (chromogranin C)	2q35-q36	LR
SLC25A5	NM_001152	solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 5 (SLC25A5), mRNA.	Xq24-q26	HR
SLC6A8	NM_005629	solute carrier family 6 (neurotransmitter transporter, creatine), member 8 (SLC6A8), mRNA.	Xq28	HR
SNAPC1	NM_003082	small nuclear RNA activating complex, polypeptide 1, 43kDa (SNAPC1), mRNA.	14q22	HR
TNFRSF25	NM_148965	tumor necrosis factor receptor superfamily, member 25 (TNFRSF25), transcript variant 1, mRNA.	1p36.2	LR
TYMS	NM_001071	thymidylate synthetase (TYMS), mRNA.	18p11.32	HR
ULK2	NM_014683	unc-51-like kinase 2 (C. elegans) (ULK2), mRNA.	17p11.2	LR
WSB1	NM_134265	WD repeat and SOCS box-containing 1	17q11.1	LR