SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

Cross-sectional study of human coding- and non-coding RNAs in progressive stages of *Helicobacter pylori* infection

Sergio Lario[1,2,7] ✉, María J. Ramírez-Lázaro[1,2,7], Aintzane González-Lahera[1,3], José L. Lavín[3], Maria Vila-Casadesús[1,4], María E. Quílez[2], Anna Brunet-Vega[5,7], Juan J. Lozano[1,4,7], Ana M. Aransay[1,3,7] & Xavier Calvet[1,2,6]

Helicobacter pylori **infects 4.4 billion individuals worldwide and is considered the most important etiologic agent for peptic ulcers and gastric cancer. Individual response to** ***H. pylori*** **infection is complex and depends on complex interactions between host and environmental factors. The pathway towards gastric cancer is a sequence of events known as Correa's model of gastric carcinogenesis, a stepwise inflammatory process from normal mucosa to chronic-active gastritis, atrophy, metaplasia and gastric adenocarcinoma. This study examines gastric clinical specimens representing different steps of the Correa pathway with the aim of identifying the expression profiles of coding- and non-coding RNAs that may have a role in Correa's model of gastric carcinogenesis. We screened for differentially expressed genes in gastric biopsies by employing RNAseq, microarrays and qRT-PCR. Here we provide a detailed description of the experiments, methods and results generated. The datasets may help other scientists and clinicians to find new clues to the pathogenesis of** ***H. pylori*** **and the mechanisms of progression of the infection to more severe gastric diseases. Data is available via ArrayExpress.**

Background & Summary

Helicobacter pylori is one of the most successful human bacterial pathogens, infecting 4.4 billion individuals worldwide[1]. Infection can induce gastric pathologies ranging from chronic gastritis in all infected individuals to peptic ulcers (in 15–20% of patients) and gastric cancer (0.5–1% of patients)[2].

Individual response to *H. pylori* infection is complex and depends on a combination of environmental factors, genetic background, host response and strain virulence[3]. The pathway towards gastric cancer is a sequence of events known as Correa's model of gastric carcinogenesis, a stepwise inflammatory process from chronic-active gastritis (CAG), atrophy (AT), intestinal metaplasia (IM) and gastric adenocarcinoma[4].

This study examines gastric clinical specimens representing different steps of the Correa pathway with the aim of identifying the expression profiles of coding- and non-coding RNAs (microRNAs and small RNAs) that may have a role in Correa's model of gastric carcinogenesis and, potentially, to develop novel clinical biomarkers.

RNAseq (for microRNAs and non-coding RNAs) and microarrays (for coding RNAs) were used to screen for differentially expressed genes in gastric biopsies (antrum/corpus). The expression of a selection of genes was confirmed in a validation cohort of patients using quantitative real-time PCR (RT-qPCR). The general study design is illustrated in Fig. 1. Here we provide a detailed description of the experiments conducted, methods used and results generated. The datasets may help other scientists and clinicians to find new clues to the pathogenesis of

[1]Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Instituto de Salud Carlos III, Madrid, Spain. [2]Digestive Diseases Service, Hospital Universitari Parc Taulí, Institut d'Investigació i Innovació Parc Taulí I3PT, Universitat Autònoma de Barcelona, Sabadell, Spain. [3]Genome Analysis Platform, CIC bioGUNE, Bizkaia Technology Park, Derio, Bizkaia, Spain. [4]Bioinformatics Platform, CIBEREHD, Barcelona, Spain. [5]Oncology Service, Hospital Universitari Parc Taulí, Institut d'Investigació i Innovació Parc Taulí I3PT, Universitat Autònoma de Barcelona, Sabadell, Spain. [6]Departament de Medicina, UAB, Sabadell, Spain. [7]These authors contributed equally: Sergio Lario, María J. Ramírez-Lázaro, Anna Brunet-Vega, Juan J. Lozano, Ana M. Aransay. ✉e-mail: SLario@tauli.cat

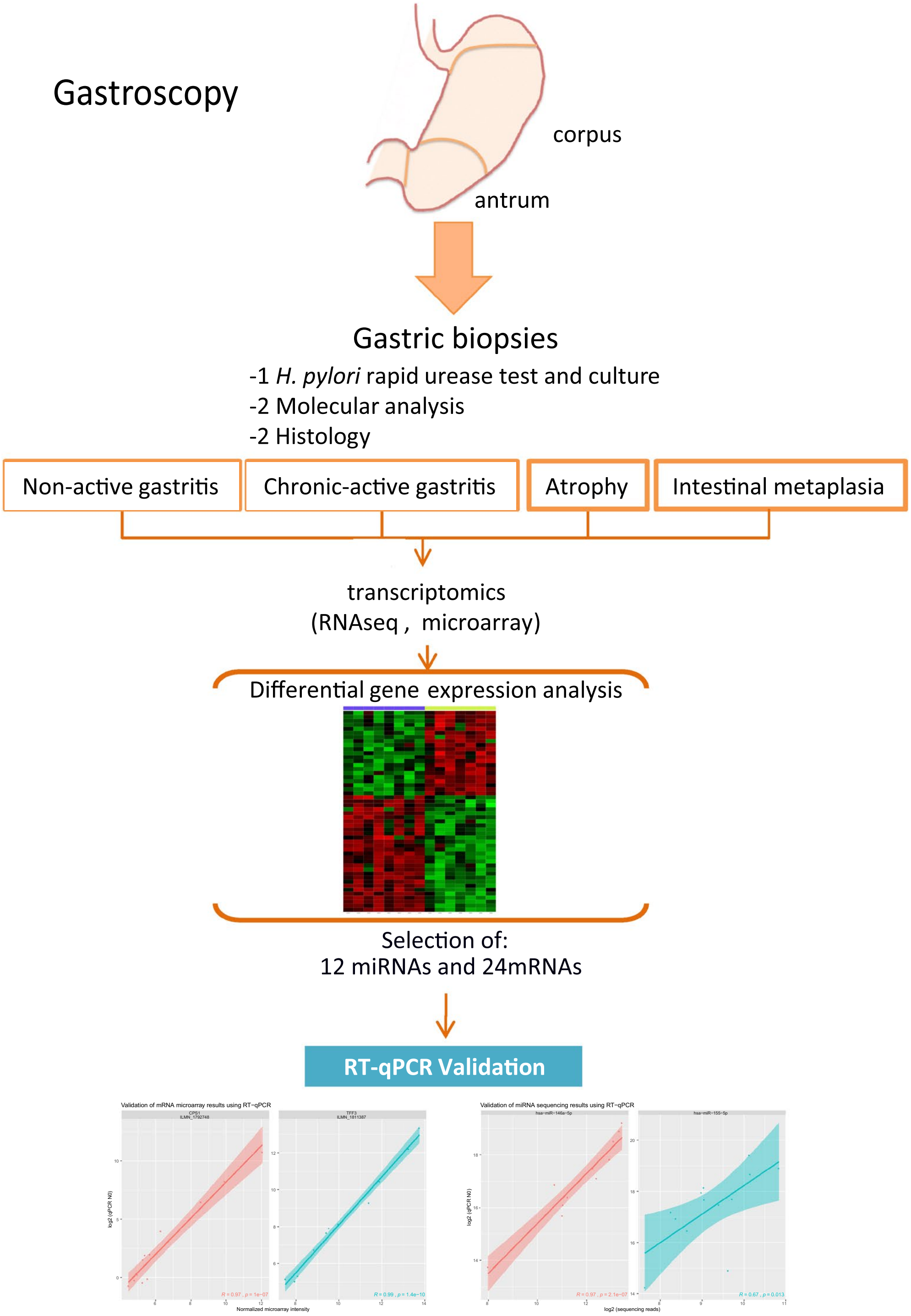

Fig. 1 Outline of the experimental design and workflow for this study from biopsy collection to data analysis. During the endoscopy procedure, antrum and corpus biopsies were collected for molecular analysis, rapid urease test and histology. Each specimen was analyzed for *H. pylori*, chronic-active gastritis, atrophy and intestinal metaplasia. Patients with *H. pylori* and neutrophil infiltrate (activity) were classified as chronic-active gastritis (CAG). Patients without activity and negative for *H. pylori* were classified as non-active gastritis (NAG).

H. pylori and the mechanisms of progression to severe disease states. The transcriptomics data is available in the ArrayExpress database[5].

Methods

Patient selection. The Digestive Service has assembled a collection of samples from dyspeptic patients. The study was undertaken in accordance with the Declaration of Helsinki, with the approval of the ethics committee at our institution (code: 2005511; approval date: 2006/1/11).

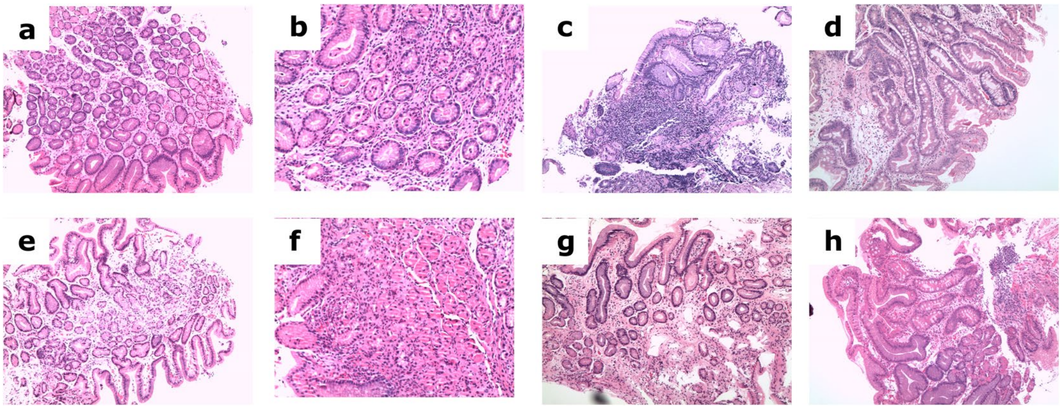

Fig. 2 Microscopic images from representative 10x hematoxylin-eosin–stained cases of non-active gastritis (**a, e**), chronic-active gastritis (**b**,**f**), atrophy (**c**,**g**) and intestinal metaplasia (**d**,**h**) are shown for antrum (**a**–**d**) and corpus (**e**–**h**).

At the time the transcriptomic experiments were performed, the collection included samples from 439 patients, enrolled from 2007–2012. The enrolment process was as follows: Dyspeptic patients referred for upper gastrointestinal endoscopy because of dyspepsia were contacted by phone and invited to participate. Those who agreed were instructed not to take antisecretory drugs for two weeks before undergoing the procedure. Exclusion criteria were: patients who were not able to stop antisecretory drugs, those who had received antibiotics in the four weeks before the endoscopy and those with a history of prior treatment for *H. pylori*. Before the endoscopy, a ^{13}[C]-urea breath test (UBT) (Cat.No. 654057, UBiTest 100 mg, Otsuka Pharmaceutical Europe Ltd, UK) was administered. During the endoscopy procedure, biopsies were taken for histology, rapid urease testing (RUT, Cat. No. 1100090, JATROX HP test CHR Heim Arzneimittel GmbH, Germany) and molecular analysis (RNAlater, Cat.No AM7021, ThermoFisher, MA, USA). After a positive RUT test, biopsies were plated on Pylori Agar (Cat. No. 413193, bioMérieux SA, Spain) in microaerophilic jars (Jar Gassing System, Don Whitley Scientific Limited, UK). After a maximum of a week, grown *H. pylori* isolates were subcultured on Columbia plates (Cat.No. 43041, bioMérieux) and identified by colony morphology, Gram-negativity and positivity for urease, catalase, and oxidase tests. *VacAs*, *VacAm* and *cagA* virulence factor genes of *H. pylori* were determined by PCR on isolated strains or biopsy samples by using custom locked nucleic acids primers (LNA, Exiqon, Denmark) and SensiMix SYBR Low-ROX Kit (Cat. No. QT625-05, Bioline, UK). Details on *VacAs*, *VacAm* and *cagA* amplification are described in detail elsewhere[6]. For histopathological evaluation, sections were stained with haematoxylin-eosin (Fig. 2) and evaluated for *H. pylori*, CAG, atrophy, intestinal metaplasia, and presence of lymphoid follicles by a pathologist specializing in digestive diseases.

Patients were considered to be *H. pylori* positive when two or more diagnostic tests (RUT, UBT, histology, culture) were positive and /or two or more *H. pylori* virulence PCR assays were positive. Patients with less than two positive diagnostic tests and with less than two positive PCR assays were considered uninfected.

Seventy antral and 26 corpus biopsies from 76 patients were selected. Due to chip restrictions, 2 antral biopsies (B294, B311) were not included in microarray analysis. In 18 cases, antrum and corpus biopsies were paired. The biopsy samples were classified into different groups based on histology: non-active gastritis (NAG, n = 16), chronic-active gastritis (CAG, n = 28), atrophic gastritis (AT, n = 15) and intestinal metaplasia (IM, n = 37). Demographic and clinical characteristics of patients can be found in Online only Table 1.

RNA extraction and quality control. Two antrum and two corpus biopsies were used to isolate total RNA. Total RNA was extracted using the mirVana miRNA isolation kit (ThermoFisher, MA, USA) as per the manufacturer's protocol and stored at −80 °C for downstream analysis. DNase treatment was performed as described in the DNA-free Kit protocol (Cat. No. AM1906, ThermoFisher, MA, USA). Total RNA was quantified with the Qubit® RNA Assay Kit (ThermoFisher, MA, USA). Quality was assessed using Agilent RNA 6000 Nano chips (Cat.No. 5067-1511) on an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA), including calculation of the RNA integrity number (RIN). The RIN score was 7.72 ± 0.6 (Fig. 3).

mRNA microarrays. Biotin-labeled cRNA samples for hybridization were prepared from 200 ng total RNA using Epicentre TargetAmp Nano-g Biotin-aRNA Labeling Kit for Illumina system (Cat. No. TAN091096; Epicentre, WI, USA). Labeled cRNA was hybridized to the HumanHT-12_V4.0 expression arrays (Cat. No. BD-103-0204; Illumina Inc., San Diego, CA) as described in the protocol/instructions. HumanHT-12 v. 4 Expression arrays were scanned with the iScan system (Illumina Inc., San Diego, CA, USA) and raw data were decoded using GenomeStudio Gene Expression Module (Illumina Inc., San Diego, CA, USA). Intensities were quantile-normalised and differentially detected transcripts were calculated using the Bioconductor *limma* package[7].

miRNA and small RNA sequencing. *TruSeq miRNA and small RNA library preparation.* Briefly, 3′ adapter ligation was performed by incubating 1 µg of total RNA of each sample with the adapter for 2 minutes at

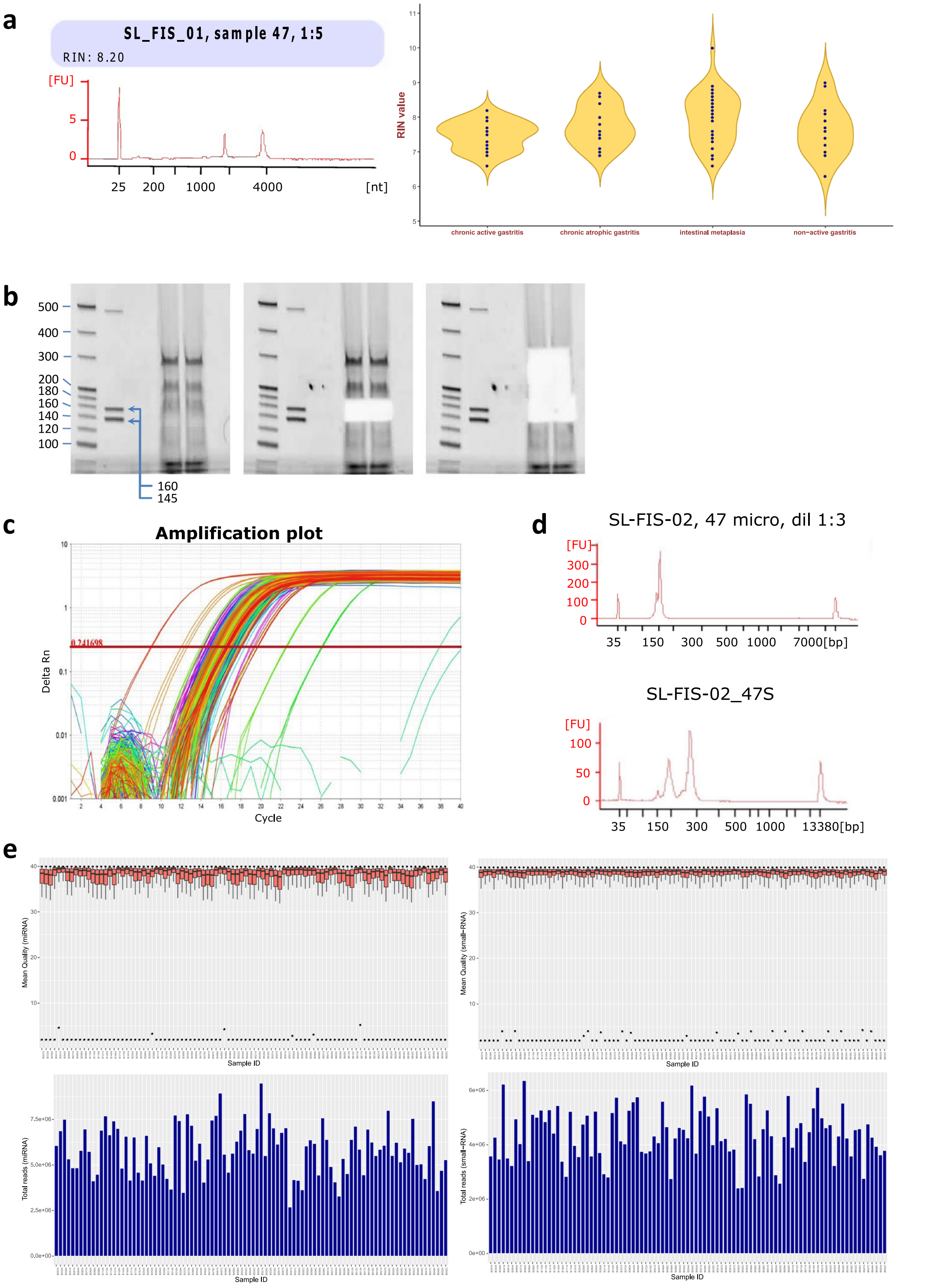

Fig. 3 Quality control of the RNA samples, sequencing libraries and sequencing reads. (**a**) Agilent Bioanalyzer electropherogram showing total RNA from sample #47 (*left*) and violin plot of RIN values according to disease state (*right*). (**b**) Library Size Selection by resolution of total RNA on 6% Novex TBE PAGE Acrylamide gels. Original acrylamide gel for sample #47 (*left*), after 145–160 bp miRNA bands (*middle*) and small-RNA fragments (200–300 bp) (*right*) were excised. (**c,d**) Final library QC. Precise library quantification was performed using real-time PCR and size distribution was assessed with Agilent BioAnalyzer High Sensitivity DNA Chips. Upper and lower panels show sample #47 miRNA and small-RNA libraries, respectively. (**e**) Sequencing of the 192 libraries generated over 9.8 × 10^8 raw reads. Mean library counts were 5.9 × 10^6 and 4.4 × 10^6 for miRNAs and small RNAs, respectively. The panels show the distribution of quality scores per base (*upper panels*) and the read count per library (*lower panels*) for both miRNA (*left panels*) and small-RNAs (*right panels*).

70 °C. Then 5′-adapter was added alongside using a truncated T4-RNA ligase 2 (Cat. No. M0351S, New England Biolabs, MA, USA) in an incubation at 28 °C for 1 hour. Half of the ligation product was used for the reverse transcription performed with SuperScript II reverse transcriptase (Cat. No. 18064-014, ThermoFisher, MA, USA) in a thermocycler for 1 hour at 50 °C. Next, enrichment of the cDNA was performed using PCR cycling: 98 °C for 30 secs; 11 cycles of 98 °C for 10 secs, 60 °C for 30 secs and 72 °C for 15 secs; a final elongation of 72 °C for 10 mins, and pause at 4 °C. PCR products were resolved on 6% Novex TBE PAGE gels (Cat. No. EC6265BOX, ThermoFisher, MA, USA). microRNA and Small_Non-coding-RNA fragments between 145–160 and 200–300 bp respectively, were cut from the gel. microRNA and Small_Non-coding-RNA libraries were extracted from polyacrylamide gel with the MinElute gel extraction kit (Cat. No. 28604, Qiagen, Germany) using an adapted protocol, in which gel slices were dissolved in a diffusion buffer (0.5 M ammonium acetate; 10 mM magnesium acetate; 1 mM EDTA, pH 8.0; 0.1% SDS) overnight at room temperature plus 3 hours and 30 min at 50 °C. The libraries were visualized on an Agilent 2100 Bioanalyzer with the Agilent High Sensitivity DNA kit (Cat. No. G2938-90320, Agilent Technologies, Santa Clara, CA) and quantified using quantitative PCR with the Kappa Library Quantification Kit (Master Mix and DNA Standards, Cat.No. KK4824, Roche-Kappa, Basel, Switzerland).

Next-generation sequencing (NGS). The libraries were pooled, and 12pM 12xmicroRNA-libraries and 14pM 12xSmall_Non-codingRNA-library pools were sequenced. Multiplexed libraries were hybridized to flow cells on a cBot Cluster Generation System (Illumina, San Diego, CA, USA) using TruSeq SR Cluster Kit v3-cBot-HS (Cat. No. GD-401-3001; Illumina, San Diego, CA, USA). The clustered flow cells were loaded onto a HiScanSQ sequencer. The sequencing was performed using the TruSeq SBS Kit v3-HS (Cat. No. FC-401-3002; Illumina, San Diego, CA, USA) for 50 cycles.

NGS Data analysis. Base calling was performed with the Illumina Real Time Analysis software (RTA, version 1.13.48) and the FASTQ files were generated with CASAVA (version: 1.8.1). Secondary data analysis was done using the sRNAbench package[8]. Briefly, reads were aligned to the human genome (UCSC hg19) using Bowtie 1.1.2[9]. miRNA annotations were obtained from miRBase[10] (version 21). Sequencing analysis was done by using the sRNAbench package[11]. Briefly, after adapter trimming and unique read grouping, reads were aligned to the human genome (UCSC hg19) using Bowtie[9] allowing for one mismatch. To provide annotations for RNA elements that mapped to the human genome, miRBase (version 21) for mature and pre-miRNA sequences was used and a matrix of counts were created. To process count and to identify differentially expressed miRNAs we use *edgeR* package[12].Transcripts were considered differentially expressed provided their edgeR FDR-adjusted P value was < 0.05.

Quantitative PCR validation. Twenty-five RNAs were reanalyzed to validate 24 messenger RNAs and 12 miRNAS. The RNAs used were a subset ($n = 25$) of the aliquots of the same RNA samples we used for sequencing and microarray analysis. Studied genes are summarized in Table 1.

cDNA synthesis. miRNA validation was performed using the miRCURY LNA Universal RT microRNA PCR system (Exiqon, Denmark). miRNAs were reverse transcribed according to the manufacturer's protocol using 10 ng of total RNA (Cat. No. 203301; miRCURY LNA™ Universal RT microRNA PCR, Polyadenylation and cDNA synthesis kit II). For coding RNAs, 1.0 µg of total RNA was converted into cDNA using PrimeScript RT Reagent Kit (Cat. No. RR037A, Takara, Japan).

Quantitative PCR. Coding RNAs were amplified using predesigned PrimeTime 5' Nuclease Assays (IDT, Iowa, USA) (assay catalog numbers are in Table 1) and PremixExTaq Probe qPCR mastermix (Cat. No. RR390W; Takara, Japan). miRNAs were quantified using predesigned microRNA LNA PCR Primer sets (Exiqon, Denmark) and SensiMix SYBR Low-ROX Kit (Cat. No. QT625-05, Bioline, UK). Amplification was performed in duplicate on a QuantStudio 7 Flex Real-Time PCR System (Applied Biosystems, Foster City, CA, USA) using 384-well plates.

qPCR data analysis. The raw PCR data was exported from QuantStudio Real-Time PCR Software v1.2 (Applied Biosystems) onto a RDML[13] file and imported into LinRegPCR (v2016.1)[14]. LinRegPCR was used to determine PCR efficiencies (E) and to calculate the starting concentration per sample (N_0). First, the program determines the baseline fluorescence and performs baseline subtraction. Then a Window-of-Linearity for all PCR samples per amplicon is set and then the algorithm determines: the mean PCR efficiency per amplicon (E_{mean}), the quantification cycle (C_q) value per sample and the fluorescence threshold set to determine the C_q (N_q). With these data, N_0 is calculated using $N_0 = N_q / (E_{mean})^{Cq}$.

Data Records

Individual miRNA and small-RNA FASTQ files and a tab-delimited file for the processed microarray data have been deposited in the ArrayExpress public repository[5]. The accession numbers are: E-MTAB-8890[15] for miRNAs, E-MTAB-8896[16] for small RNAs and E-MTAB-8889[17] for mRNAs. The sample metadata records are provided in Online only Table 1.

GENE Symbol	Refseq accession	Detects all variants(a)	Exon location(a)	Mean PCR efficiency(b)	Company and catalogue number
AGPAT2	NM_006412(1)	No	4–5	1,86	IDT Hs.PT.58.1470724
ACTB	NM_001101(1)	Yes	6–6	1,86	IDT hs.PT.56a.40703009.g
ANXA13	NM_004306(2)	Yes	10–11	1,90	IDT Hs.PT.56a.20938889.g
APOB	NM_000384(1)	Yes	8–9	1,87	IDT Hs.PT.56a.19389676
C3	NM_000064	Yes	27–28	1,90	IDT Hs.PT.56a.2840009
CDX1	NM_001804(1)	Yes	1–2	1,89	IDT Hs.PT.58.468499
CFH	NM_001014975(1)	No	9–10	1,89	IDT Hs.PT.58.41054235
CPS1	NM_001122633(3)	Yes	27–28	1,88	IDT Hs.PT.58.2708374
CREB1	NM_134442	No	3–5	1,89	IDT Hs.PT.58.4988504
CXCR5	NM_001716(1)	No	1–2	1,89	IDT Hs.PT.56a.1692541
EIF4G2	NM_001042559(3)	Yes	3–5	1,91	IDT Hs.PT.58.6917393
FUT9	NM_006581	Yes	2–3	1,89	IDT Hs.PT.58.22395619
HIPK3	NM_005734(2)	Yes	3–4	1,89	IDT Hs.PT.58.2927056
HNF4G	NM_004133(1)	Yes	2–3	1,86	IDT Hs.PT.58.26995600
IL8	NM_000584(1)	Yes	3–4	1,82	IDT Hs.PT.58.38869678.g
KRT20	NM_019010(1)	Yes	5–6	1,89	IDT Hs.PT.58.39027228
MEG3	NR_002766(8)	No	5–10	1,86	IDT Hs.PT.58.25426100
MMP9	NM_004994(1)	Yes	3–4	1,84	IDT Hs.PT.58.22814824.g
MTTP	NM_000253(1)	Yes	18–19	1,87	IDT Hs.PT.58.94887
MUC2	NM_002457(1)	Yes	28–30	1,89	IDT Hs.PT.58.4321237
POFUT1	NM_015352(1)	No	6–7	1,78	IDT Hs.PT.58.19361092
RUNX2	NM_001024630(3)	Yes	6–7	1,88	IDT Hs.PT.56a.19568141
SDHA	NM_004168(1)	Yes	3–4	1,88	IDT Hs.PT.58.41017719
TFF3	NM_003226(1)	Yes	1–2	1,89	IDT Hs.PT.58.1814807
WDR1	NM_017491(1)	No	4–5	1,88	IDT Hs.PT.58.40308614
MIR103A1	NR_029520.1	NA	NA	1,89	Exiqon 204063
MIR146A	NR_029701	NA	NA	1,86	Exiqon 204688
MIR153–1	NR_029563	NA	NA	1,81	Exiqon 204338
MIR155	NR_030784.1	NA	NA	1,90	Exiqon 204308
MIR182	NR_029614.1	NA	NA	1,89	Exiqon 206070
MIR191	NR_029690.1	NA	NA	1,86	Exiqon 204306
MIR192	NR_029578.1	NA	NA	1,91	Exiqon 204099
MIR196B	NR_029911.1	NA	NA	1,88	Exiqon 204555
MIR19B1	NR_029490.1	NA	NA	1,85	Exiqon 204450
MIR204	NR_029621.1	NA	NA	1,89	Exiqon 206072
MIR215	NR_029628.1	NA	NA	1,87	Exiqon 204598
MIR340	NR_029885.1	NA	NA	1,89	Exiqon 206068

Table 1. qPCR Primer assays used for mRNA and miRNA validation. [a]Primer assays targeting all splicing variants were chosen for validation purposes, and when possible, in the same exon where the Illumina probe was positioned. [b]PCR efficiency was calculated by LinRegPCR software. Using the raw qPCR data, the algorithm computes iteratively a Window-of-Linearity for a specific amplicon and calculates the C_q and PCR efficiency for each individual reaction and amplicon. NA: not applicable.

Technical Validation

Quality control. *Sample collection.* In order to ensure the collection of biopsy tissue samples would provide high-quality results for microbiology, molecular analysis and histology, a two round biopsy protocol was followed. During the endoscopy, a first set of biopsy samples was collected for microbiological (in sterile saline) and molecular analysis (in RNAlater) and a second set were fixed in formalin for histopathological examination. By doing this, we ensured that formalin contamination of biopsy forceps did not interfere with the RUT and *H. pylori* culture. Histological examination was performed by a pathologist specialized in digestive diseases. In order to increase the total RNA yield and because intestinal metaplasia is typically present as small mucosal patches, we isolated RNA from two gastric biopsies per anatomical location. The reason is that the biopsy cores examined by the pathologist are different from the biopsy specimens used for molecular analysis. By using two biopsies, we were more confident that if the pathologist reported intestinal metaplasia in the histology specimens, intestinal metaplasia would also be present in the molecular biology cores. Additionally, two biopsies are the minimum recommended by the Updated Sydney System[18].

RNA processing. Figure 2 shows the quality control procedures used in this study for RNA integrity, library preparation and sequencing.

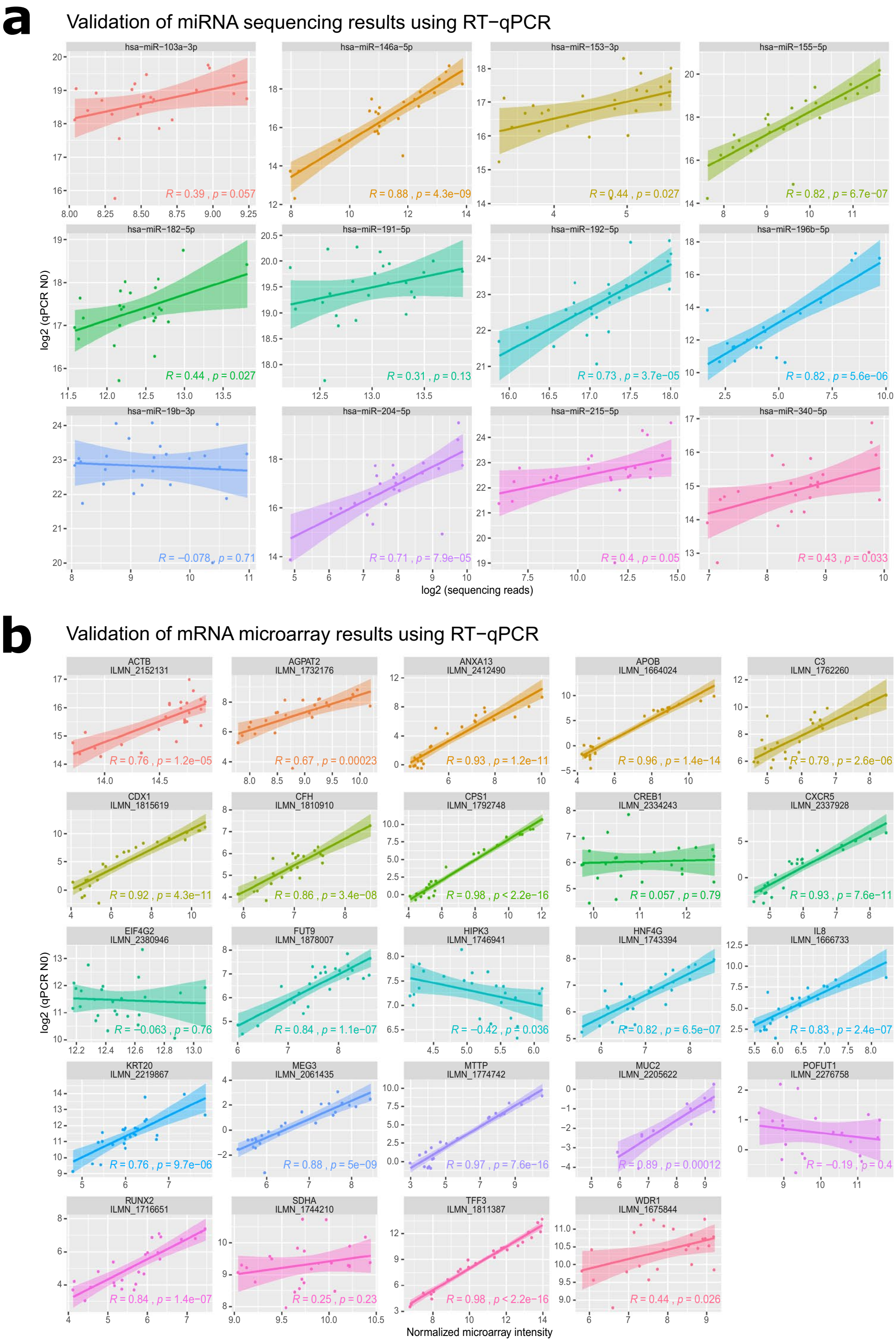

Fig. 4 Validation of miRNAs (**a**) and messenger RNAs (**b**) by RT-qPCR. A panel 12 of miRNAs and 24 mRNAs were selected for validation of 25 RNA samples. Aliquots of the same RNA samples were used for sequencing, microarray and qPCR measurements. Raw qPCR data was exported to LinRegPCR software. N_0 (an estimate of the target starting concentration per reaction) was calculated using the formula $N_0 = N_q/E^{Cq}$ where E is the amplicon PCR efficiency and N_q is the fluorescence threshold set to determine C_q. The Pearson correlation coefficient (*R*), the p-value and 95% confidence interval are indicated. Additional correlations to genes having multiple probes can be found in ref. [29].

Gene expression validation by qPCR. We used LinRegPCR[14] for calculating individual and mean PCR efficiencies. Amplicons showed high PCR efficiencies, ranging from 1.78 to 1.91. PCR inhibition can be detected using individual PCR efficiency values. Samples showing PCR efficiencies greater than 5% of the PCR mean efficiency per amplicon were excluded. The algorithm also calculates N_0. N_0 is the starting quantity of mRNA or miRNA (expressed in arbitrary fluorescence units). Quantitative N_0 values have been used in previous publications[19–24]. Determining N_0 has several advantages over relative quantification. First, the selection of a housekeeping gene is often controversial since the expression of all genes is regulated. Second, the expression of a housekeeping gene varies to a greater or lesser extent under experimental conditions[25]. Third, to solve this issue a quantitative PCR approach with a correction factor according to the starting amount of RNA used in the reverse transcription has been recommended (i.e. µg of RNA)[26] instead of relative quantification.

To evaluate the concordance in gene expression between microarray or RNA-seq and qPCR, we calculated the correlation between normalized microarray/RNA-seq and qPCR log transformed N_0 values (Fig. 4). Overall, high R and low p-values values ($R > 0.8$, $p < 0.001$) were observed between microarray and qPCR measurements. Some of them were probe dependent (i.e. C3 probe ILMN_1762260: $R = 0.79$, $p < 0.001$, but C3 ILMN_1662523 was not correlated). Five miRNA showed high correlation ($R > 0.7$, $p < 0.001$), 4 were poorly correlated ($R \sim 0.4$, $p < 0.05$) and 3 were not correlated.

Usage Notes

miRNA, small-RNA raw sequencing data (FASTQ) and normalized microarray data can be analysed by a variety of freely accessible packages and platforms, such as R/Bioconductor[27]. Some R/Bioconductor packages can be used without prior programming knowledge by using the Galaxy platform[28].

The authors encourage proper citation of data sources for any work based on this dataset.

Received: 6 April 2020; Accepted: 19 June 2020;
Published online: 08 September 2020

References

1. Hooi, J. K. Y. *et al*. Global Prevalence of Helicobacter pylori Infection: Systematic Review and Meta-Analysis. *Gastroenterology* **153**, 420–429 (2017).
2. Kusters, J. G., van Vliet, A. H. M. & Kuipers, E. J. Pathogenesis of Helicobacter pylori infection. *Clin. Microbiol. Rev.* **19**, 449–490 (2006).
3. Chmiela, M., Karwowska, Z., Gonciarz, W., Allushi, B. & Stączek, P. Host pathogen interactions in Helicobacter pylori related gastric cancer. *World J. Gastroenterol.* **23**, 1521–1540 (2017).
4. Amieva, M. & Peek, R. M. Pathobiology of Helicobacter pylori-Induced Gastric Cancer. *Gastroenterology* **150**, 64–78 (2016).
5. Athar, A. *et al*. ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res.* **47**, D711–D715 (2019).
6. Companioni, O. *et al*. Genetic variation analysis in a follow-up study of gastric cancer precursor lesions confirms the association of MUC2 variants with the evolution of the lesions and identifies a significant association with NFKB1 and CD14. *Int. J. Cancer* **143**, 2777–2786 (2018).
7. Ritchie, M. E. *et al*. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
8. Aparicio-Puerta, E. *et al*. sRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential expression. *Nucleic Acids Res.* **47**, W530–W535 (2019).
9. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
10. Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* **39**, D152–7 (2011).
11. Barturen, G. *et al*. sRNAbench: profiling of small RNAs and its sequence variants in single or multi-species high-throughput experiments. *Methods Gener. Seq.* **1**, (2014).
12. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinforma. Oxf. Engl.* **26**, 139–140 (2010).
13. Lefever, S. *et al*. RDML: structured language and reporting guidelines for real-time quantitative PCR data. *Nucleic Acids Res* **37**, 2065–9 (2009).
14. Ramakers, C., Ruijter, J. M., Deprez, R. H. L. & Moorman, A. F. M. Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neurosci. Lett.* **339**, 62–66 (2003).
15. Lario, S. *et al*. miRNA sequencing data from patients with non-active gastritis, chronic active gastritis and precursor lesions of gastric cancer. *ArrayExpress* https://identifiers.org/arrayexpress:E-MTAB-8890 (2020).
16. Lario, S. *et al*. small-RNA sequencing data from patients with non-active gastritis, chronic active gastritis and precursor lesions of gastric cancer. *ArrayExpress* https://identifiers.org/arrayexpress:E-MTAB-8896 (2020).
17. Lario, S. *et al*. mRNA microarray data from patients with non-active gastritis, chronic active gastritis and precursor lesions of gastric cancer. *ArrayExpress* https://identifiers.org/arrayexpress:E-MTAB-8889 (2020).
18. Dixon, M. F., Genta, R. M., Yardley, J. H. & Correa, P. Classification and grading of gastritis. The updated Sydney System. International Workshop on the Histopathology of Gastritis, Houston 1994. *Am J Surg Pathol* **20**, 1161–81 (1996).
19. Aggarwal, S. D. *et al*. Function of BriC peptide in the pneumococcal competence and virulence portfolio. *PLoS Pathog.* **14**, e1007328 (2018).
20. Bhavsar, S. P., Løkke, C., Flægstad, T. & Einvik, C. Hsa-miR-376c-3p targets Cyclin D1 and induces G1-cell cycle arrest in neuroblastoma cells. *Oncol. Lett.* **16**, 6786–6794 (2018).
21. Boissière, A. *et al*. Application of a qPCR assay in the investigation of susceptibility to malaria infection of the M and S molecular forms of An. gambiae s.s. in Cameroon. *PloS One* **8**, e54820 (2013).
22. Koenis, D. S. *et al*. Nuclear Receptor Nur77 Limits the Macrophage Inflammatory Response through Transcriptional Reprogramming of Mitochondrial Metabolism. *Cell Rep.* **24**, 2127–2140.e7 (2018).
23. van Wijk, B. *et al*. Epicardium and myocardium separate from a common precursor pool by crosstalk between bone morphogenetic protein- and fibroblast growth factor-signaling pathways. *Circ Res* **105**, 431–41 (2009).
24. Lario, S. *et al*. Expression profile of circulating microRNAs in the Correa pathway of progression to gastric cancer. *United Eur. Gastroenterol. J.* **6**, 691–701 (2018).
25. Lion, T. Current recommendations for positive controls in RT-PCR assays. *Leukemia* **15**, 1033–7 (2001).
26. Bustin, S. A. Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems. *J Mol Endocrinol* **29**, 23–39 (2002).
27. Huber, W. *et al*. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12**, 115–121 (2015).
28. Afgan, E. *et al*. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **46**, W537–W544 (2018).
29. Lario, S. *et al*. Validation of microarray expression by RT-qPCR. *figshare* https://doi.org/10.6084/m9.figshare.11932854 (2020).

Acknowledgements

This study was funded by the Fondo de Investigación en Salud (FIS), Instituto de Salud Carlos III, ISCiii (PI12/01802). We are also indebted to A. Salvador for her help with the English.

Author contributions

A.M.A., A.GH and J.L.L processed the microarras, performed the RNA sequencing and discussed and contributed to writing. J.J.L. and M.V.C. performed omics data analyses and discussed and contributed to writing. A.B.V. and M.E.Q. performed RNA extraction, QC and qPCR experiments. X.C. designed the study and discussed and contributed to writing. S.L. and M.J.R.L. designed the study, contributed to the patient consent, collection of samples, qPCR data analysis, wrote and edited the manuscript. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.