Exploring Gene Expression and Epigenetic Data in Gastric Cancer

Débora Antunes[1], Rúben Rodrigues[1], Tânia Barata[1], and Miguel Rocha[1]
debiaantunes@gmail.com

Informatic Department, University of Minho, 4710-057 Braga, Portugal

Abstract. Nowadays, diseases like gastric cancer are objects of studies in several investigations. The data acquired is used to identify features useful to understand and characterize a disease. To analyze the amount of data produced every day, new bioinformatic tools and pipelines are needed. Diverse types of data can be produced by various techniques. Next Generation Sequencing (NGS) techniques like Chromatin Immunoprecipitation Sequencing (ChIP-Seq) and RNA Sequencing(RNA-Seq) can use various pipelines with different tools. The Sarkar *et al.* study uses RNA-seq and ChIP-seq raw data from two conditions, wild-type and Sox2 knock-out, to understand the role of Sox2 in gastric cancer of *Mus musculus*. This project aims to recreate the pipeline and consequently the results obtained in that paper using tools from the Docker platform.

Keywords: Gastric Cancer · Next Generation Sequencing (NGS) · RNA Sequencing (RNA-Seq) · Chromatin Immunoprecipitation Sequencing (ChIP-Seq) · Docker

1 Introduction

1.1 Context and Motivation

Worldwide, gastric cancer (GC), also known as stomach cancer it is the fifth most commonly diagnosed cancer and the third leading cause of cancer death with over 1,000,000 new cases in the last year and 783,000 deaths. The incidence shows wide geographical variation, with high-risk areas like Eastern Asia, Eastern Europe, and South America and low-risk areas like the African continent [9]. The 5-year relative survival rates for stomach cancer varies with how far cancer has spread, which can be classified as localized with 68%, as regional with 31% or as distant with 5% [27]. Given these factors, it is imperative to find more information about this type of cancer and its molecular basis.

At the moment, with all the Next Generation Sequencing (NGS) technology available, there is a large amount of data generated every day and it is necessary to use bioinformatics tools to analyze it. Some of this data is used to identify somatic mutations and gene expression profiles, to better understand a disease and its characteristics, but there are diverse approaches and it is important to

understand how they work and their limitations. For this reason, we will use and compare various tools for analysis of NGS gastric cancer cell data. Thereunto, it is required to identify scientific studies that describe gastric cancer data and corresponding analysis pipelines to replicate their results.

1.2 Goals

The purpose of this project is to compare and use diverse tools with docker platform, replicating the article "Sox2 Suppresses Gastric Tumorigenesis in Mice" [30]. In detail, the objectives are:

- Using the PRJNA327571 dataset of RNA-seq data from the article, perform an alignment with STAR, do read counts using HTSeq and do a differential expression analysis with the R package edgeR.
- With the PRJNA327572 dataset of ChIP-seq data from the study, do a peak-calling with MACs2 and a motif search with DAVID.
- Compare all results with the ones obtained by the authors of the article.

1.3 Document outline

2 State of art

2.1 Gastric Cancer

Generally, most GCs are sporadic but familial clustering is observed in 5%-10% of cases, of which 1%-3% are hereditary [31]. The principal genetic alterations that lead to GC are mutations, like small insertions or deletions, amplifications, and rearrangements.

According to the World Health Organization (WHO), GC includes gastric carcinoma, neuroendocrine neoplasms, lymphoma, and mesenchymal tumors [8]. Gastric carcinomas can sill be divided into two groups, intestinal or diffuse, according to microscopic and macroscopic differences [21].

Nowadays, the most sensitive and specific diagnostic screening method is an upper gastrointestinal endoscopy, but is an invasive procedure that can cause

hemorrhage and perforation [19]. Some other treatment options include surgery, chemotherapy, radiotherapy, and immunotherapy [31]. There are a few known risk factors for GC. Age has a positive correlation with the incidence rate of GC [19] and men have rates 2-fold higher compared with women [9]. Some infectious agents like Helicobacter pylori and Epstein–Barr virus (EBV) have been associated with a higher risk of developing GC [31]. In addition, there is some evidence that environment or occupational exposures, the diet or habits like smoking may play a role [19][31].

The American Joint Committee on Cancer (AJCC) developed the TNM classification for GC, that is based in size and localization of the tumor and if it spreads to other tissues or organs [32]. As shown in the Table 1, the T category describes the primary tumor, the N category indicates how many lymph nodes are affected and the M category shows if it is classified as metastatic disease.

Table 1: TNM classification for GC.

Stage grouping	Stage description
T0	No evidence of primary tumor
T1	Tumor confined to the mucosa or submucosa
T2	Tumor invades muscularis propria
T3	Tumor penetrates subserosa
T4a	Tumor invades serosa
T4b	Tumor invades adjacent structures
N0	No regional lymph node metastasis
N1	Metastasis in 1-2 regional lymph nodes
N2	Metastasis in 3-6 regional lymph nodes
N3	Metastasis in 7 or more regional lymph nodes
M0	No distant metastasis
M1	Distant metastasis

2.2 Gastric Cancer Biomarkers

To understand if GC is present in an individual or to determine its severity, the use of biomarkers is important. Biomarkers are measurable and reproducible indicators of the condition and are divided into four types according to their function: diagnostic, predictive, prognostic and therapeutic. [17][22].

One of the GC biomarkers is CDH1, a gene that codifies a transmembrane glycoprotein (E-Cadherin) involved in the adhesion and differentiation of the epithelial gastric cells. This gene is an important tumor suppressor gene in GC and its inactivation enables the cancer progression. Mutations on this gene are associated most frequently with hereditary diffuse gastric cancer (HDGC) syndrome [6][10]. It is weak prognosis biomarker but can also be considered a predictive biomarker [10].

HER2, a tyrosine receptor kinase (RTK), codified by the protooncogene ERBB2 and plays an important role in the regulation of the cell. The amplification of ERBB2 and consequent overexpression of HER2 triggers the PI3K-AKT and MAPK pathways, leading to the survival, growth, and proliferation of the cancer cell. For this reason, this protein is a prognostic and predictive biomarker [6][10][12].

Other biomarkers used are EGFR, FGFR2, PIK3CA, and MET. EGFR is a receptor that belongs to the tyrosine kinase receptors family and its overexpression is a prognostic indicator of a bad outcome in GC [10][12]. The FGFR2 is a fibroblast growth factor that when amplified works as a prognostic biomarker but also as a potential candidate for targeted therapy [12]. Mutations in one gene involved with the phosphatidylinositol-3-kinase (PIK3)/mTOR pathway, PIK3CA, are linked with a prognosis of decreased survival and increased lymph nodes metastasis in GC [10]. MET is a transmembrane tyrosine kinase receptor and an important predictive biomarker. Its amplification is related to a bad prognosis [10].

2.3 Omics Data

There are different types of omics data: genomics, epigenomics, transcriptomics, proteomics, metabolomics, and microbiomics. In this study will be approached epigenomics and transcriptomics. [13].

- Epigenomics: This field studies chemical modifications of the DNA that are associated with the regulation of gene transcription detected with NGS techniques like Chromatin Immunoprecipitation Sequencing (ChIP-Seq) [13][16].
- Transcriptomics: Focuses on detecting, quantifying and examining RNA transcripts. Some techniques used include microarrays and RNA Sequencing (RNA-Seq) [13][18].

3 Methodology

3.1 Overview

In this study, we explore pipelines for data analysis of raw data from RNA-seq and ChIP-seq. All the tools used in this project operate inside Docker.

RNA-seq uses Next-Generation Sequencing (NGS) to study the presence, quantity and quality of an RNA sample. The first step is the isolation of the RNA of a tissue followed by ribosome removal and then by its conversion to cDNA. The cDNA is fragmented and, after sequencing, the raw RNA reads obtained pass through quality control, using FASTQC. The quality control makes an analysis of the reads and informs about any problems associated with the data. Some of these problems can be corrected by Trimmomatic that trim and crop data as well as remove adapters. The filtered RNA sequences are then aligned with STAR software obtaining a SAM file. This alignment is necessary to determine the location from which the reads originated and to characterize them. It can

be performed against a reference genome or *de novo* and it is important to consider mutations like mismatches, insertions and deletions, and mechanisms like the alternative splicing. After, there is a sequence counting of the number of alignments per gene or transcript by HTSeq. The data pursue to normalization and to differential expression analysis using R packages (DESEq2/edgeR).[1] (Fig. 1).

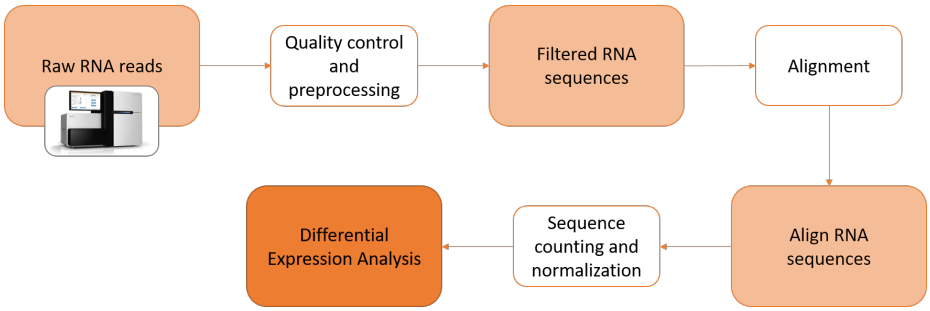


Fig. 1: Schematic pipeline to analyze raw RNA reads from RNA-seq.

ChIP-seq uses NGS to determine the sites of the genome where transcription factors, DNA-binding enzymes, histones, and other proteins bind. ChIP-seq cross-links bound proteins to DNA, shear DNA strands by sonicating and add specific antibodies to immunoprecipitate target protein. After precipitation, the DNA is purified and sequenced producing raw DNA reads. The quality control and preprocessing use the same tools as the RNA-seq but the alignment is done by Bowtie2. The align DNA sequence pursue to quality control and peak-calling done by MACS2. Peak-calling is a method that identifies the areas of the genome where the proteins bound by finding the regions that have been enriched with aligned reads. After identifying the ChIP-seq peaks it is done a final downstream analysis with motif finding. [4][26] (Fig. 2).

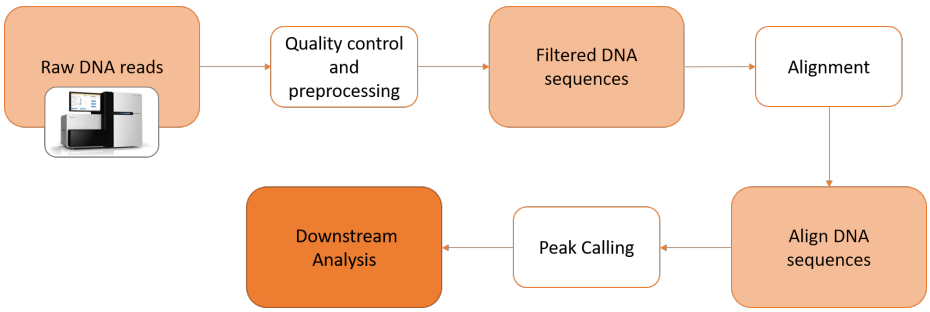


Fig. 2: Schematic pipeline to analyze raw DNA reads from ChIP-seq.

3.2 RNA-seq Pipeline

To realize the RNA-seq pipeline several bash scripts and an R script were created (available in https://github.com/Daantunes/Projeto/tree/master/rna_seq). All bash scripts can indicate what arguments are needed if "-h" is written. The first lines of each script have an if condition that allows this function.

The PRJNA327571 dataset from the article had six files ".fastq.gz".

Quality control

The first step of this pipeline was to perform quality control of this raw RNA reads. The "do_fastqc.sh" script uses the fastqc biocontainer (version v0.11.5_cv3) to access the quality of the raw RNA reads. Docker[25] uses the command run on this container with various options. These options are the same in every bash script:

`--rm` - Automatically remove the container when it exits.
`--user` - Sets the username. Is needed because some containers do not give write permissions.
`$(id -u):$(id -g)` - Gives user and group permissions.
`-v` - Allows to indicate a folder from the operating system that corresponds to the container working directory. That folder will be used to take input and output data.
`&` - Runs process in background of container.

FASTQC tool [3] uses several options and a file ".fastq.gz" for input. The only option used was "-o" that creates all output files in the specified output directory. This directory must exist, that's why it was created before running the container. This container writes text in the command window that sometimes are errors, so the text was redirected to a ".log" file and the option "2>&1" forces the stderr (2) into stdout (1). This approach was made in other scripts. The output from this tool is a FASTQC report in HTML format.

Preprocessing

After obtaining these reports, the biocontainer trimmomatic from Bioconda (version 0.39–1) was used to filter the raw data through "do_trimmomatic_SE.sh" script. The arguments for Trimmomatic [7] consists of multiple options, input and output files in the format ".fastq.gz" and trimming steps:

SE - Means that the trim is going to be performed in single-end reads.
ILLUMINACLIP:TruSeqALL-SE.fa:2:0:10 - Removes the adapters and overrepresented sequences in the TruSeqALL-SE.fa file. The first number specifies that a maximum of 2 mismatches will be allowed. The second is a parameter used in paired-end reads, so it is set to 0. The last indicates that the accuracy of the

match between the adapter sequence to trim and the read must be 10.
LEADING:0 - Stipulates the number of bases that are trimmed at the start of the read if they are below a threshold quality, in this case is 0.
TRAILING:0 - Stipulates the number of bases that are trimmed at the end of the read if they are below a threshold quality, in this case is 0.
SLIDINGWINDOW:0:O - Trims the sequence using a sliding window with a determined number of bases and an average threshold quality. It is set to 0:0 because we do not want to trim based on base quality.
MINLEN:36 - Specifies a minimum length, in this case 36, and excludes reads below that value.

Although no adapters were found by FASTQC, they were removed because there might be some contamination in the sequences that was missed by FASTQC. The Trimmomatic tool has some common adapter files, but because no adapter type was found associated with the sequencing library used, a new file was created with "create_adaptor.sh" script that merges all SE adapter files. After, the overrepresented sequences found by the FASTQC were added to this file.

Alignment

The next step was the alignment of the reads with the biocontainer star (version 2.7.0f–0) from Bioconda. The "do_STAR_index.sh" uses STAR [11] to generate the genome index and it needs two input files, the genome references sequences in the format ".fa" and the annotated transcripts in the format ".gtf". The files "Mus_musculus.GRCm38.75.dna.primary_assembly.fa" and "Mus_musculus.GRCm38.75.gtf" were downloaded from Ensembl and unzipped. The options used in STAR were:

--runThreadN - Sets the number of threads that will be used for genome generation.
--runMode - Directs STAR to generate the genome index.
--genomeDir - Identify the directory where the index is stored. This directory needs to be created before running the container.
--genomeFastaFiles - Specifies the path to the genome references sequences.
--sjdbGTFfile - Specifies the path to the annotated transcripts.
--outFileNamePrefix - Sets the output directory and the output file name prefix.
--sjdbOverhang - Indicates the length of the sequence around the annotated junction. In this case is 50 because this length should be equal to the maximum length of the read minus 1.

The output files generated are in an internal STAR format and contain binary genome sequence, text chromosome names/lengths, transcripts/genes information, suffix arrays and splice junctions coordinates.

The "do_STAR_map.sh" scrip realizes the alignment of the reads against the genome index. The only input needed is the ".fastq.gz" file with the reads. To run the mapping job were used several options:

--runThreadN - Sets the number of threads that will be used for genome generation.
--genomeDir - Identify the directory where the index were generated.
--readFilesIn - Path to the ".fastq.gz" files.
--readFilesCommand - Indicates that the ".fastq.gz" files need to be uncompressed.
--outFileNamePrefix - Sets the output directory and the output file name prefix. This directory needs to be created before running the container.

This command creates several output files:

- Log.out - Reports information about the run.
- Log.progress.out - File with the job progress statistics.
- Log.final.out - Contains a summary mapping statistics after the job has finalized.
- Aligned.out.sam - The alignment file in ".sam" format.

Sequence counting

The "do_htseq.sh" script uses the biocontiner htseq (version 0.11.2–py27h637b 7d7_1) from Bioconda to count the number of aligned reads per gene. The HTSeq tool [2] uses several options and needs a SAM file and a GTF file as input.

-m=intersection-strict - Specifies the mode used for handling overlaps of reads, in this case only the complete match between reads and reference gene are counted.
--stranded=no - Consider an overlapping whether the read and feature are in the same or opposite strands.
-t=gene - Sets the feature type.

Two output files were created. The "htseq_counts.readcounts" has the counts obtained by the HTSeq and the "htseq_counts.log" reports information about the run. These files were compiled together by the commands in "compilation_counts.sh" resulting in a "COUNTS.tab" file.

Differential expression analysis

The R packages DESEq2[23] and edgeR[29] were used to discover which genes were differentially expressed ("dif_analysis.R"). First, it was created a variable "counts" from the "COUNTS.tab" file created previously, a factor "condition" with the attributes of the read samples (three were from a WT stomach and the other three were from a Sox2 KO stomach) and a data frame "sampleTable"

that associates the names of the samples with the corresponding condition. The table "counts" was filtrated removing the lines with no counts.

The analysis with DESeq2 started with the construction of the DESeq object "dds" from the table "counts". After, the function "DESeq" did a normalization to this object and realized some tests. With the "results" function, a table with diverse statistical values (log2 fold changes, p-values, adjusted p-values, and others) was generated. The adjusted p-value cutoff (alpha) was set to 0.05. To discover what genes were down-regulated and up-regulated we set the log2FoldChange to less and greater than 1, respectively. To initialize the analysis with edgeR, an edgeR object was created by the function "DGEList" using the table "counts" and the factor "condition". As recommended in the manual, normalization was done by the function "calcNormFactors". The next step was to test for differentially expressed genes with the function "exactTest" that needs an estimated dispersion calculated by the function "estimateDisp". Once more we used the adjusted p-value (FDR) of 0.05 and the logFC of 1. Using the Database for Annotation, Visualization and Integrated Discovery (DAVID version 6.8)[15][14], a function annotation was done in the differential express genes.

3.3 ChIP-seq Pipeline

To realize the ChIP-seq pipeline a few bash scripts were created (available in https://github.com/Daantunes/Projeto/tree/master/chip_seq). Similar to the RNA-seq, all bash scripts can indicate what arguments are needed if "-h" is written. The PRJNA327572 dataset from the article had three files ".fastq.gz". The first was a wild-type stomach sample (WT), the second was a Sox2 knock-out stomach sample (KO) and the last was an input sample of wild-type stomach. As the quality control and preprocessing uses the same tools of RNA-seq, and for that reason, the bash scripts are the same.

Alignment

The Docker container bowtie2 (version 2.3.5–py27he860b03_0) from Bioconda was used to create a Bowtie index and aligned the sequences.

The script "do_bowtie2_index.sh" uses the function "bowtie2-build". This function needs the "fastq.gz" file and the basename for the index files as inputs and creates six output files. To generate the alignments was used the Bowtie2 tool[20] inside "do_bowtie2_align.sh" script. This tool uses several options:

-p - Sets the number of threads that will be used.
-q - Specifies that the reads are in the FASTQ format.
--no-unal - If a read failed to align, suppress its SAM record.
-x - Indicates the basename of the index.
-U - Indicates the "fastq.gz" that will be align.
-S - Name of the output file in the SAM format.

Peak calling

The MACS2 tool[34] was used through the container macs2 (version 2.1.2–py27r351h14c3975_1) from Bioconda with the "do_macs2.sh" script. To call peaks from alignment results, the function "callpeak" was used with some options:

-t - Indicates the file of the treatment.
-c - Indicates the control file.
-n - Basename of the output files.
--outdir - Specifies the output directory.

In this case, we used this tool to call peaks of the Sox2 knock-out and the wild-type samples with the input sample as control, and of the input sample without any control. This tool creates multiple output files:

"_model.r" - R scripts that create a PDF image about the model based on the data.
"_peaks.narrowPeak" - A file that contains the peak locations together with peak summit, p-value and q-value and is the the BED6+4 format.
"_peaks.xls" - A file with various types of information about the called peaks.
"_summits.bed" - A file with the peak summits locations for each peak in BED format.

Downstream analysis

In the downstream analysis, we used the webtool MEME to find motifs. First our files ".narrowPeak" had to be converted to FASTA format. The "convert2fa.sh" script converts the ".narrowPeak" file to a ".bed" file and uses the function getfasta from the container bedtools (version 2.23.0–hdbcaa40_3)[28] to create a ".fa" file with the options:

-fi - Indicates the FASTA file used in the alignment.
-bed - Indicates the sample file.
-fo - Specifies the name of the output file.

These files were uploaded on MEME[5] using the KO sample as a primary sequence, the input sample as the control sequence and the input motifs were from *Mus musculus*. The WT samples were also uploaded in GREAT[24] to assign biological meaning to the peaks found. Annotation of the WT peaks was made using two distinct R packages: the ChIPseeker [33] and the ChIPpeakAnno [35]. The first used the TxDb for *Mus musculus* to find the genomic features of the peaks and the second was used to filtrate the genes with Sox2 binding sites within 10kb of the transcription start site (TSS).

4 Results and Discussion

The FASTQC report has several modules that help the characterization of raw and filtered data.

The Basic statistics module generates the composition statistics for the chosen file like file type, encoding, total sequences, filtered sequences, sequence length, and %GC. The raw data from RNA-seq has a total of sequences that varies from 26614137 to 44718612, a sequence length of 51 in all files and a %GC from 48 to 50, while the raw data from ChIP-seq as a total of sequences that varies from 21270194 to 27079936, a sequence length of 76 in all files and a %GC from 42 to 53.

The Per base sequence quality module shows a plot of quality values across all bases at each position. In our reports, all files from RNA-seq produced similar plots with high scores which means a good base call with no need of a trim. The same happens with the reports from ChIP-seq (Fig. S1).

The Per sequence quality scores module indicates if a subset of the sequences has generally low-quality values. In all files, this module produced good results because the distribution of average read quality is in the upper range of the plot (Fig. S2).

The Per base sequence content module creates a plot that shows the proportion of the four DNA bases for each base position. The expected is little to no difference as seen in the ChIP-seq reports, however, the RNA-seq reports have shown a biased sequence composition at the start of the read that cannot be fixed by processing. Although this module fails in nearly all RNA-seq libraries, redoing the RNA-seq using a different library preparation may be a solution (Fig. S3).

Per sequence GC content is a module that allows seeing the distribution of GC content across the whole length of each sequence and compares it to the normal distribution of GC content. The majority of the RNA-seq reports raised warnings because the sum of the deviations from the normal distribution represented more than 15% of the reads. In RNA-seq, if the observed distribution is narrower than the theoretical one the warning is raised but the reads may have high quality. The ChIP-seq reports gave no warnings (Fig. S4).

The Adapter content module tries to find adapters sill present in the reads. None of the reports found any adapter.

The Overrepresented sequences module creates a list of sequences that appear more than expected. FASTQC attempts to identify the sequence. In the RNA-seq reports were found sequences that might be TruSeq Adapters (indexes 7, 20, 21, 22 and 25), removed later in the pre-processing, and in the ChIP-seq, against expectations, no overrepresented sequences were found. If any sequence was found in the ChIP-seq report, they should not be removed because they may be the transcription factors that the study wants to analyze.

The results from the FASTQC analysis after processing showed improvement in the Per sequence GC content of four files and in the Overrepresented sequences of all files. The FASTQC reports from Chip-seq after processing were very similar to others.

Table 2: Results from the article and from our analysis.

	Article	DESeq2	edgeR
Genes Deregulated	59	63	59
Genes Up-regulated	42	36	42
Genes Down-regulated	17	27	17

The authors of the article performed an RNA-seq of the gastric glands to determine if the loss of Sox2 in the stomach epithelium created differences in the gene expression. Their analysis was done by the edgeR package and their results showed 42 up-regulated and 17 down-regulated genes, making a total of 59 deregulated genes. They found out that these differentially expressed genes were involved in extracellular functions, signaling, and secretion, after a GO analysis. In our analysis, the DESeq2 package revealed less up-regulated genes (36) and more down-regulated genes (27), while the edgeR had the same results as the article with 42 up-regulated genes and 17 down-regulated genes (Table 2). The functional annotation from DAVID revealed that these genes were categorized by the keywords: signal, cell membrane, secreted, disulfide bond, transport and others (Table S1).

In the article, the analysis of the results from the ChIP-seq resulted in the discovery of more than 7000 binding sites and of the motifs Sox2, Klf4 and Gata6 (Fig. S5). These binding sites were mostly distal intergenic (40.80%), introns (39.40%) and promoters (14.40%) (Fig. S6). Our analysis found a total of 4638 peaks in the WT sample and 132 peaks in the KO sample. MEME, with the Differential Enrichment mode, found various motifs, but the only one similar to the results from the articles was the Sox2 (Fig. 3). GREAT produced results that show the importance of these binding sites for the organism. GO molecular functions are related to studies on gastric cancer and the mouse phenotype reveal abnormalities in the intestine and epidermal layer morphology (Fig. S7). The genomic features resulting from our analysis with ChIPseeker were mostly introns (35.70%), distal intergenic (35.63%) and promoters (27.45%) (Fig. 4).

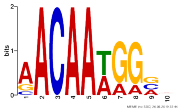
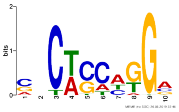
(a)  (b)  (c) 

Fig. 3: Motifs found with MEME: (a) Sox3, Sox9, Sox4, NANOG and SOX2; (b) SNAI1, BCL6, ZIC3, TEAD1 and CTCFL; and (c) ZIC3, LYL1, SNAI1, SOX10 and TFE2. Adapted from MEME.

The results from the article presented an overlap of 10 genes between their results from RNA-seq and ChIP-seq (Fig. S8). Using the genes that resulted

from the ChIPpeakAnno analysis and the differentially expressed genes from the edgeR analysis an overlap of 5 genes was found (Fig. 5).

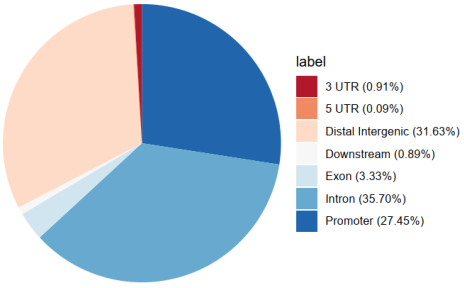


Fig. 4: Genomic features of genes with Sox2 binding sites.

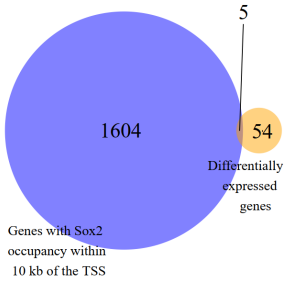


Fig. 5: Overlap between the differentially expressed genes and the genes with Sox2 binding sites $<$ 10kb.

The results obtained at the end of the RNA-seq pipeline with the edgeR package were equal to the ones from the article but the package DESeq2 produced different results. That was expected because some defaults of the two packages, like the normalization and filtering, are distinct. The majority of the motifs found at the end of the ChIP-seq pipeline were different from those of the article showing that the motif finding was not correctly replicated, but the fact that we found Sox2 motif shows that our analysis is correct. In the other end, the biological mean of the peaks and the genomic features were very similar. In the article, the methodology for ChIP-seq pipeline does not refer to some parameters and tools used, so the differences in results may be caused by the use of different options. In our analysis, the number of genes that overlap is minor than the ones obtained by the authors of the article, but that was expected because minus peaks were found and, consequently, minus genes with Sox2 binding sites within 10kb of the TSS. All of the results indicate that the RNA-seq pipeline has been successfully replicated but the ChIP-seq pipeline needs a different choice of parameters and tools to correctly replicate the results.

5 Conclusion

The RNA-seq and the ChIP-seq are sequencing methods very useful to explore the gene expression and epigenetic data in GC studies, but there are various tools to realize the different steps of the pipelines. The methodology used, by the authors of the study, in the RNA raw reads produced by the RNA-seq was very detailed which allowed a correct realization of the pipeline, using the same tools. Consequently, the results obtained were the same showing that the pipeline was correctly replicated. The methodology used for the ChIP-seq data

was not so clear and the tools used were different, so the results obtained were not the same, although the final analysis gave a very similar biological mean and genomic features. As there is currently no consensus in the scientific community in what is the best pipeline to use in this type of sequencing, in the future other tools and options may be used in this dataset to compare the differences and limitations between each tool. Overall, we can conclude that both pipelines were successfully implemented and can be used to process and analyze RNA-seq and ChIP-seq raw data.

References

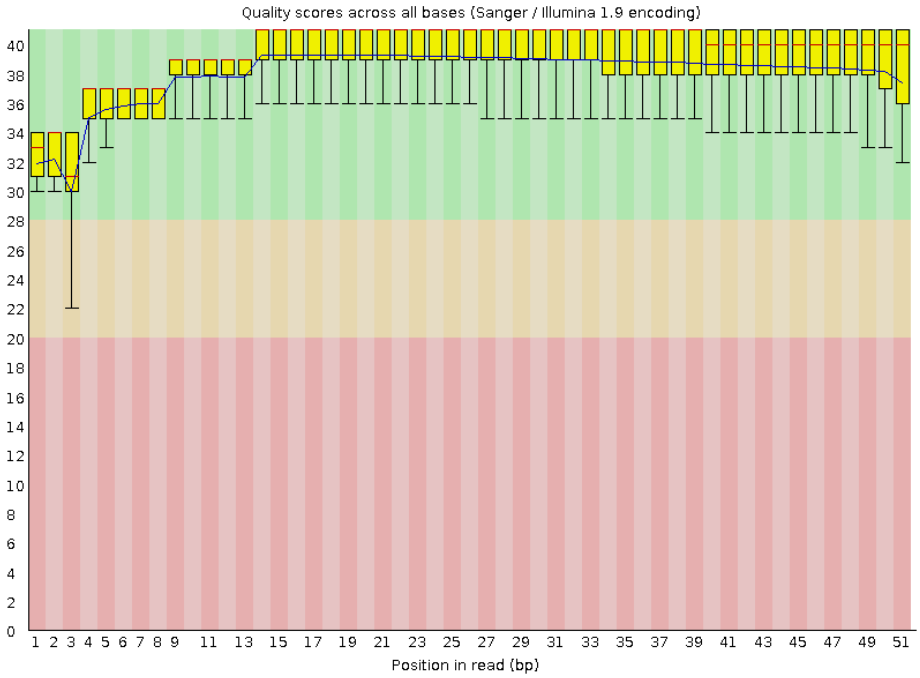
1. Rna-seqlopedia. https://rnaseq.uoregon.edu/, (Accessed on 15/04/2019)
2. Anders, S., Pyl, P.T., Huber, W.: Htseq—a python framework to work with high-throughput sequencing data. Bioinformatics **31**(2), 166–169 (2015)
3. Andrews, S., et al.: Fastqc: a quality control tool for high throughput sequence data (2010)
4. Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., Zhang, J.: Practical guidelines for the comprehensive analysis of chip-seq data. PLoS computational biology **9**(11), e1003326 (2013)
5. Bailey, T.L., Elkan, C., et al.: Fitting a mixture model by expectation maximization to discover motifs in bipolymers (1994)
6. Baniak, N., Senger, J.L., Ahmed, S., Kanthan, S., Kanthan, R.: Gastric biomarkers: a global review. World journal of surgical oncology **14**(1), 212 (2016)
7. Bolger, A.M., Lohse, M., Usadel, B.: Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics **30**(15), 2114–2120 (2014)
8. Bosman, F.T., Carneiro, F., Hruban, R.H., Theise, N.D., et al.: WHO classification of tumours of the digestive system. World Health Organization (2010), ed. 4
9. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **68**(6), 394–424 (2018)
10. Carlomagno, N., Incollingo, P., Tammaro, V., Peluso, G., Rupealta, N., Chiacchio, G., Sandoval Sotelo, M.L., Minieri, G., Pisani, A., Riccio, E., et al.: Diagnostic, predictive, prognostic, and therapeutic molecular biomarkers in third millennium: a breakthrough in gastric cancer. BioMed research international **2017** (2017)
11. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R.: Star: ultrafast universal rna-seq aligner. Bioinformatics **29**(1), 15–21 (2013)
12. Durães, C., Almeida, G.M., Seruca, R., Oliveira, C., Carneiro, F.: Biomarkers for gastric cancer: prognostic, predictive or targets of therapy? Virchows Archiv **464**(3), 367–378 (2014)
13. Hasin, Y., Seldin, M., Lusis, A.: Multi-omics approaches to disease. Genome biology **18**(1), 83 (2017)
14. Huang, D.W., Sherman, B.T., Lempicki, R.A.: Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research **37**(1), 1–13 (2008)
15. Huang, D.W., Sherman, B.T., Lempicki, R.A.: Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nature protocols **4**(1), 44 (2009)

16. Illumina, I.: An introduction to next-generation sequencing technology (2015)
17. Italiano, A.: Prognostic or predictive? it's time to get back to definitions. J Clin Oncol **29**(35), 4718 (2011)
18. Joyce, A.R., Palsson, B.Ø.: The model organism as a system: integrating'omics' data sets. Nature reviews Molecular cell biology **7**(3), 198 (2006)
19. Karimi, P., Islami, F., Anandasabapathy, S., Freedman, N.D., Kamangar, F.: Gastric cancer: descriptive epidemiology, risk factors, screening, and prevention. Cancer Epidemiology and Prevention Biomarkers **23**(5), 700–713 (2014)
20. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with bowtie 2. Nature methods **9**(4), 357 (2012)
21. Lauren, P.: The two histological main types of gastric carcinoma: diffuse and so-called intestinal-type carcinoma: an attempt at a histo-clinical classification. Acta Pathologica Microbiologica Scandinavica **64**(1), 31–49 (1965)
22. Lin, L.L., Huang, H.C., Juan, H.F.: Discovery of biomarkers for gastric cancer: a proteomics approach. Journal of Proteomics **75**(11), 3081–3097 (2012)
23. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome biology **15**(12), 550 (2014)
24. McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., Bejerano, G.: Great improves functional interpretation of cis-regulatory regions. Nature biotechnology **28**(5), 495 (2010)
25. Merkel, D.: Docker: lightweight linux containers for consistent development and deployment. Linux Journal **2014**(239), 2 (2014)
26. Nakato, R., Shirahige, K.: Recent advances in chip-seq analysis: from quality management to whole-genome annotation. Briefings in bioinformatics **18**(2), 279–290 (2016)
27. Noone, A., Howlader, N., Krapcho, M., Miller, D., Brest, A., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A., Lewis, D., et al.: Seer cancer statistics review, 1975–2015. Bethesda, MD: National Cancer Institute (2018)
28. Quinlan, A.R., Hall, I.M.: Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**(6), 841–842 (2010)
29. Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**(1), 139–140 (2010)
30. Sarkar, A., Huebner, A.J., Sulahian, R., Anselmo, A., Xu, X., Flattery, K., Desai, N., Sebastian, C., Yram, M.A., Arnold, K., et al.: Sox2 suppresses gastric tumorigenesis in mice. Cell reports **16**(7), 1929–1941 (2016)
31. Sitarz, R., Skierucha, M., Mielko, J., Offerhaus, G.J.A., Maciejewski, R., Polkowski, W.P.: Gastric cancer: epidemiology, prevention, classification, and treatment. Cancer management and research **10**, 239 (2018)
32. Washington, K.: of the ajcc cancer staging manual: stomach. Annals of surgical oncology **17**(12), 3077–3079 (2010)
33. Yu, G., Wang, L.G., He, Q.Y.: Chipseeker: an r/bioconductor package for chip peak annotation, comparison and visualization. Bioinformatics **31**(14), 2382–2383 (2015)
34. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al.: Model-based analysis of chip-seq (macs). Genome biology **9**(9), R137 (2008)
35. Zhu, L.J., Gazin, C., Lawson, N.D., Pagès, H., Lin, S.M., Lapointe, D.S., Green, M.R.: Chippeakanno: a bioconductor package to annotate chip-seq and chip-chip data. BMC bioinformatics **11**(1), 237 (2010)

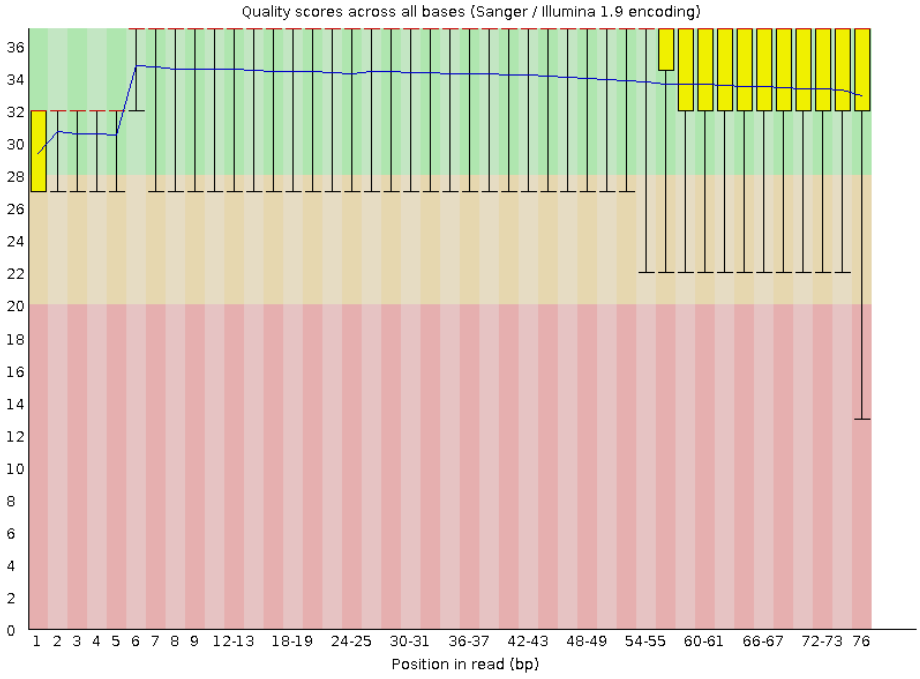
Supplementary Material

Table S1: Results from our analysis of RNA-seq deferentially expressed genes. Adapted from DAVID.

Keyword	Genes
Signal	16
Cell Membrane	15
Secreted	12
Disulfide bond	12
Transport	9
Ion Transport	6
Cell junction	6
Calcium	6

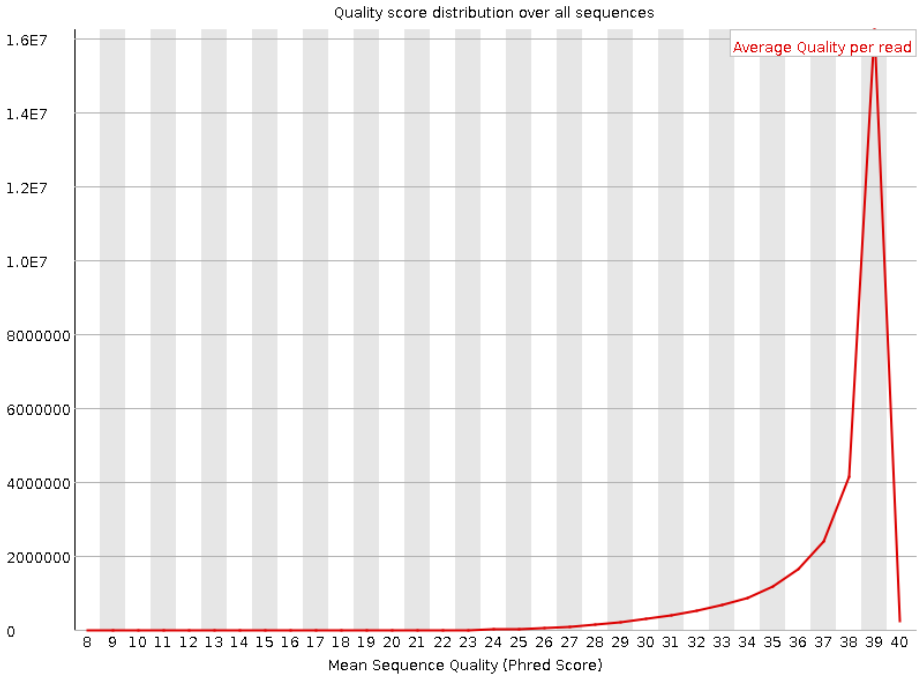


(a)

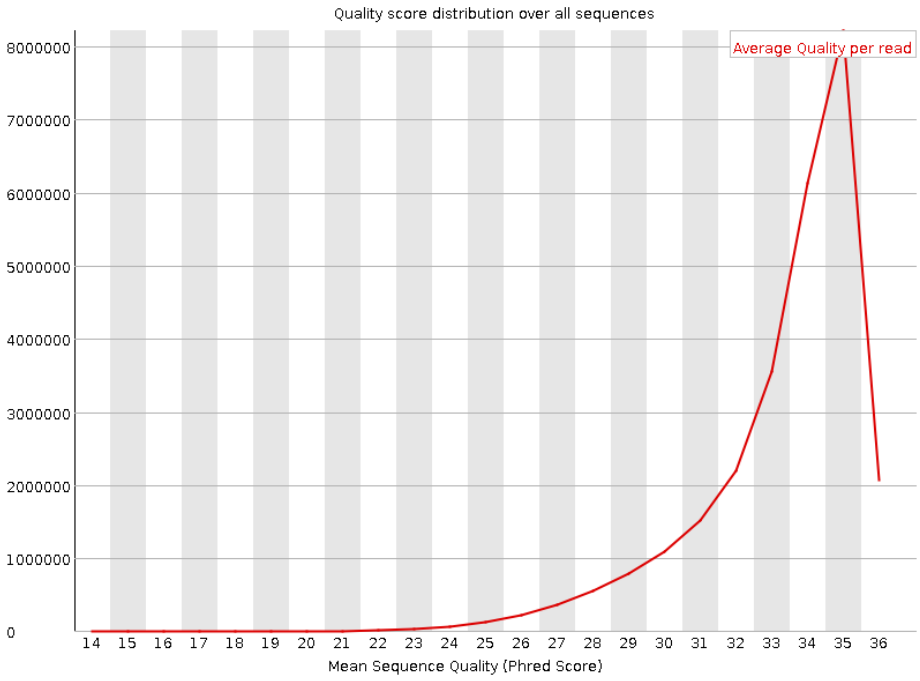


(b)

Fig. S1: Results from the FASTQC report module Per base sequence quality of (a) RNA-seq raw data and (b) Chip-seq raw data.

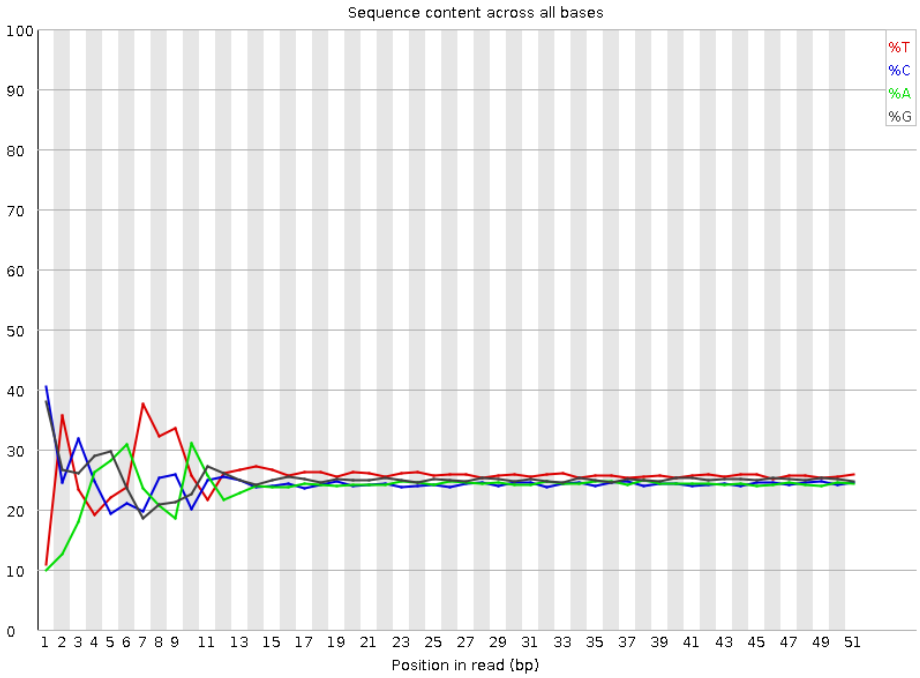


(a)

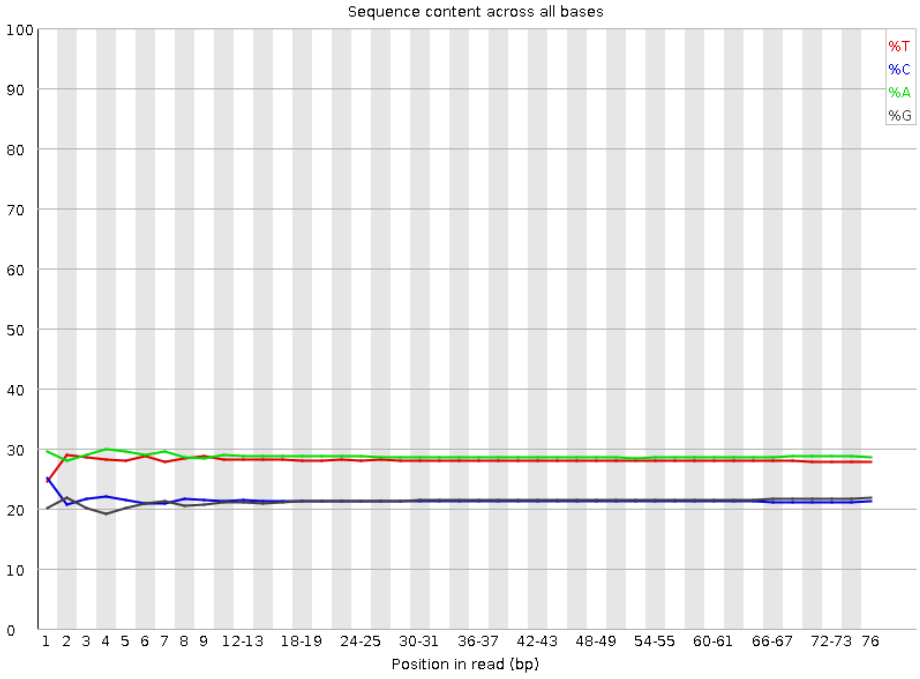


(b)

Fig. S2: Results from the FASTQC report module Per sequence quality scores of (a) RNA-seq raw data and (b) Chip-seq raw data.

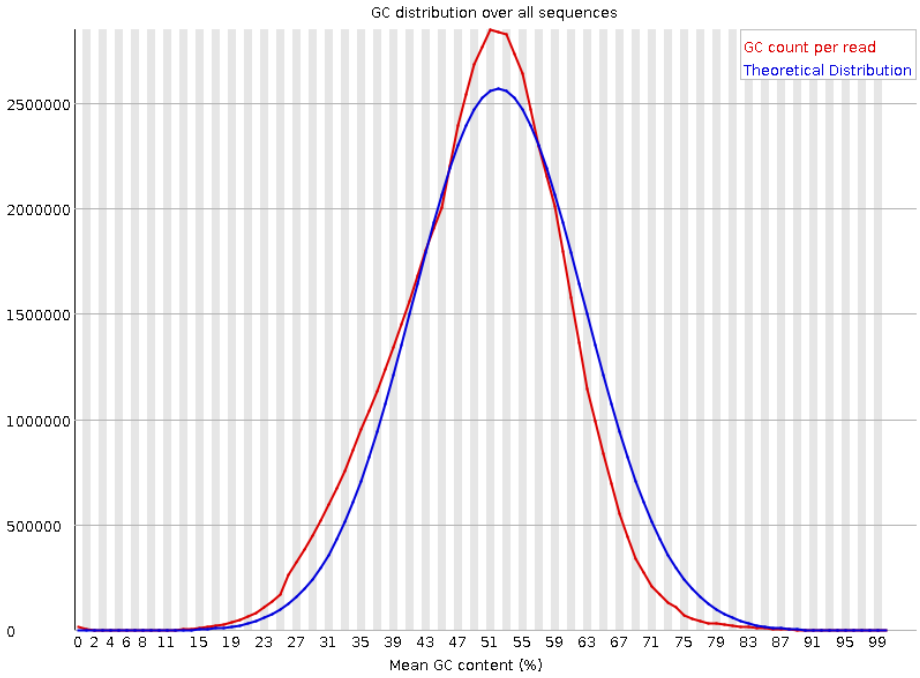


(a)

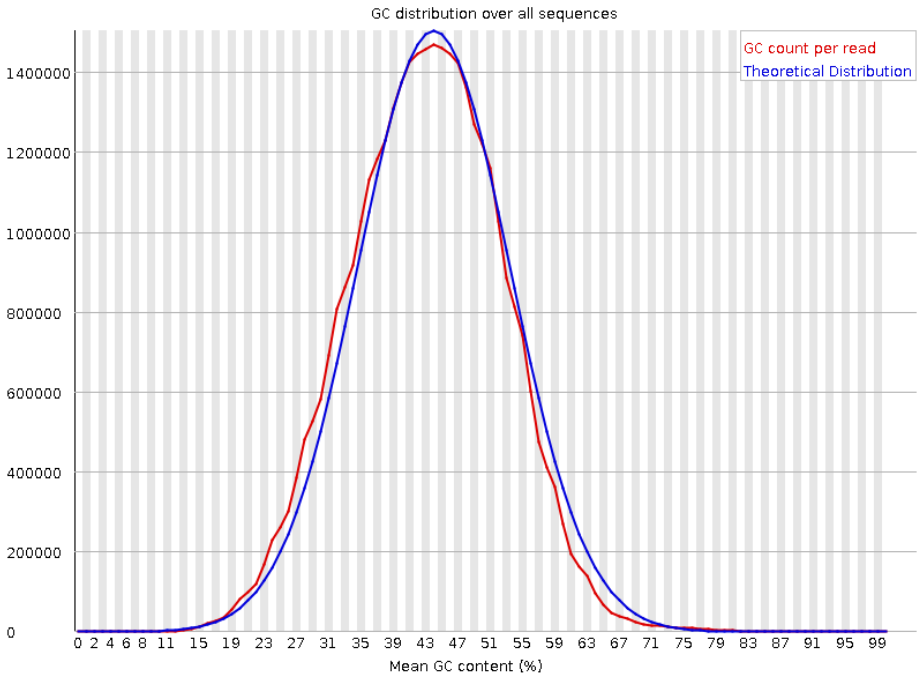


(b)

Fig. S3: Results from the FASTQC report module Per base sequence content of (a) RNA-seq raw data and (b) Chip-seq raw data.



(a)



(b)

Fig. S4: Results from the FASTQC report module Per sequence GC content of (a) RNA-seq raw data and (b) Chip-seq raw data.

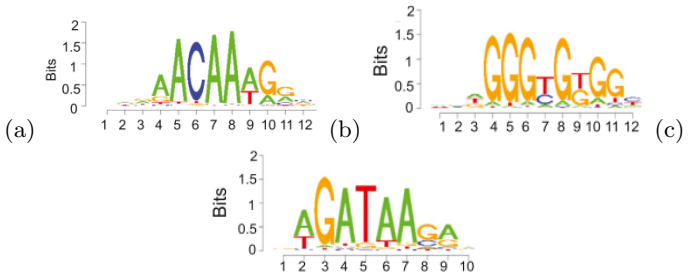


Fig. S5: Motifs found in the article: (a) Sox2; (b) Klf4; and (c) Gata6. Adapted from Sarkar *et al.*.

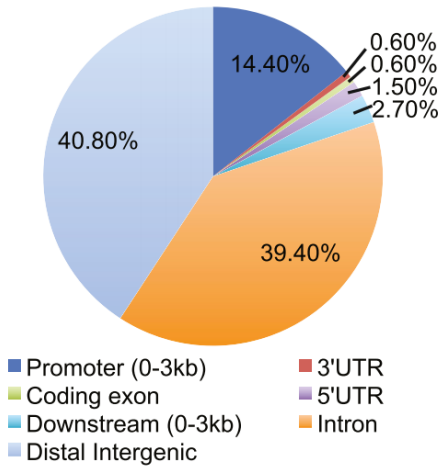
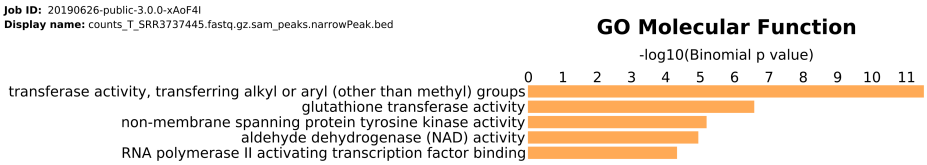
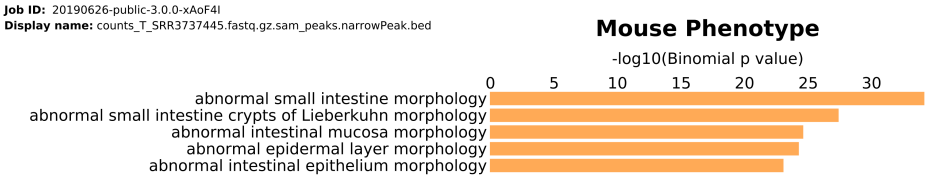


Fig. S6: Results from the article showing the distribution of Sox2 binding sites across the genome.



(a)



(b)

Fig. S7: Result from the GREAT analysis of the ChIP-seq wild-type peaks: (a) GO Molecular Function and (b) Mouse Phenotype.

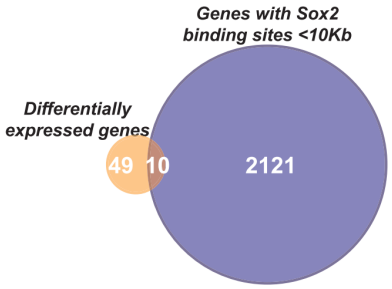


Fig. S8: Results from the article that shows the overlap between the differentially expressed genes from RNA-seq analysis and the genes with Sox2 binding sites within 10kb of the TSS from ChIP-seq analysis.