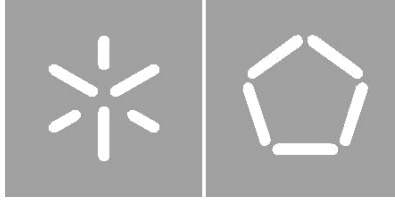


Universidade do Minho
Escola de Engenharia

João António Moreira da Silva

**Pattern analysis in multi-platform
genomic data available by TCGA**



Universidade do Minho
Escola de Engenharia
Departamento de Informática

João António Moreira da Silva

**Pattern analysis in multi-platform genomic
data available by TCGA**

Dissertação de Mestrado
Mestrado em Bioinformática

Trabalho realizado sob orientação de

Doutor Pedro Gabriel Dias Ferreira
Professor Doutor Miguel Francisco de Almeida
Pereira da Rocha

Outubro de 2016

Agradecimentos

Em primeiro lugar gostaria de agradecer ao meu orientador Pedro Ferreira, por tudo o que me ensinou, pelo apoio, pela disponibilidade e incentivo. Um enorme obrigado à líder do grupo Carla Oliveira por me ter acolhido e por toda a ajuda e apoio dados durante este tempo. Um agradecimento especial à Patrícia Oliveira pelas lições dadas que vão ser essenciais para o trabalho e para a vida. Fica também um muito obrigado aos restantes elementos do grupo Expression Regulation in Cancer, pelo ambiente acolhedor que me foi proporcionado durante esta estadia. Uma palavra de agradecimento ao meu co-orientador Professor Miguel Rocha, pelo auxílio que me prestou sempre que assim o solicitei.

Aos meus companheiros diários de trabalho, Abel Ernesto, Diana Lemos e Joana Ferreira, um muito obrigado pelos bons momentos passados. Por todas as conversas, desabafos e coisas estúpidas que fizemos, um especial obrigado irmã Lúcia, és a maior. A todos os meus amigos que me ajudaram fora do trabalho, Christophe Sousa, Luís Teixeira, Fábio Costa, Luís Arménio, Rui Ferreira, Celso Ferreira, Jorge Nogueira, Joaquim Santos e Sandro Mota, obrigado por tudo. Aos meus amigos da faculdade Catarina Lemos, Raphael Morais, Tiago Carvalho, Rita Silva, Adriana Nogueira, Daniela Cunha, Tiago Alves um muito obrigado, amigos da faculdade são para a vida. Um enorme obrigado à Sandra Oliveira pelo apoio incondicional e por me ter ajudado a ser melhor pessoa, nunca esquecerei.

O ultimo agradecimento, mas o mais importante à minha família, em especial aos meus pais e à minha irmã que representam tudo para mim, são o meu maior e melhor exemplo. Sem vocês nada disto era possível. Obrigada!

ABSTRACT

Gastric cancer is considered one of the most complex diseases in the world. There are more than 200 cancer subtypes characterised. Efforts such as TCGA and ICGC aim to catalogue and discover major cancer-causing genome alterations in large cohorts of human samples.

The accumulation of enormous quantities of multidimensional data requires new methods for integrative analysis of multiple data sources. We divided our work into two major topics: i) Recapitulation of genomic and epigenomic patterns of an independent study on gastric cancer conducted in the host group, and ii) Gender differential expression in gastric cancer. For these analyses, we used the TCGA as an independent cohort. Multiple genomic platforms were analysed such as DNA Methylation, copy number variation (CNV) and messenger RNA (mRNA).

In an internal previous study, a gene amplified in 20% of gastric cancer cases showed regions of co-amplification. Our goal was to assess if we could recapitulate these results using the TCGA cohort. These amplifications can offer the tumour mechanisms of resistance to therapies and may have potential as therapeutic targets.

For the same previous study, epigenomic patterns were also characterised. where a set of genes hypermethylated in 80% of the samples were found. Our goal was to investigate if these results could be recovered in the TCGA cohort. Eight genes were found constantly hypermethylated in the TCGA tumour samples. In the future, they can be important for early detection and classification of the tumour stages.

Differences between males and females go beyond anatomy and include differential susceptibility to a variety of diseases. Cancer has an important and considerable gender differential susceptibility that has confirmation in several epidemiological studies. A second major goal of this dissertation was to evaluate gender differential expression in gastric cancer. We analysed RNA sequencing data from TCGA, revealing specific differentially expressed genes (DEGs) in males and females. There are a limited number of previous studies that integrate gender into their design. Our study shows that there are significant transcriptomic differences between genders, therefore, this subdivision should be taken into account in the design of cancer studies since the resulting heterogeneity may affect their results. Overall, our results shed light on the genomic mechanisms that may drive in a gender differential susceptibility to cancer.

In conclusion, we were able to recapitulate the results from the internal group study of genomic and epigenomic data using an independent cohort of TCGA using. With gender differential expression analysis, we provide novel insights into the differential risk underlying gastric cancer. The fast growth of high-throughput biology will further expand our knowledge of molecular dimensions of cancer.

Keywords: Gastric Cancer; TCGA; DNA Methylation; RNA Sequencing; copy number variation (CNV)

RESUMO

O cancro gástrico é considerado uma das doenças mais complexas do mundo. Existem mais de 200 subtipos deste cancro. Esforços como o TCGA e o ICGC visam catalogar e descobrir as principais alterações e causadores de alterações em cancro, em grandes grupos de amostras humanas.

A acumulação de enormes quantidades de dados multidimensionais requer novos métodos de análise integrativa de múltiplas fontes de dados. Dividimos o nosso trabalho em dois grandes tópicos: i) Recapitulação de padrões genómicos e epigenómicos de um estudo independente em cancro gástrico realizado no grupo; e ii) Expressão diferencial por género em cancro gástrico. Para estas análises, usamos o coorte independente do TCGA. Múltiplas plataformas genómicas foram analisadas, tais como metilação, a variação do número de cópias (CNV) e dados de RNA mensageiro (mRNA).

Num estudo realizado internamente, um gene amplificado em 20% dos casos de cancro gástrico apresentou regiões de co-amplificação. O nosso objetivo era avaliar se conseguiríamos recapitular esses resultados, usando o coorte do TCGA. Estas amplificações podem conferir ao tumor mecanismos de resistência a terapias e podem ter potencial como alvos terapêuticos.

No mesmo estudo interno, dados epigenómicos foram também caracterizados, onde foi encontrado um conjunto de genes hipermetilados em 80% das amostras. O nosso objetivo foi investigar se estes resultados poderiam ser recuperados no coorte do TCGA. Oito genes foram encontrados constantemente hipermetilados em amostras de cancro do TCGA. No futuro, estes podem ser importantes na deteção precoce e a classificação das fases do tumor.

As diferenças entre homens e mulheres vão para além da anatomia, incluindo diferentes suscetibilidades a uma variedade de doenças. O cancro possui importante e considerável suscetibilidade diferencial entre géneros, confirmada em vários estudos epidemiológicos. O segundo grande objetivo desta dissertação foi avaliar a expressão diferencial dos genes nos dois sexos em cancro gástrico. Analisámos dados de sequenciação de RNA a partir do TCGA, revelando genes específicos diferencialmente expressos em homens e mulheres. Há um número limitado de estudos anteriores que integram o sexo no seu desenho. O nosso estudo mostra que existem diferenças transcritómicas significativas entre sexos, e por isso esta subdivisão deve ser tida em conta no desenho de estudos em cancro, uma vez que a heterogeneidade resultante pode ter impacto nos seus

resultados. No geral, os nossos resultados podem clarear os mecanismos genómicos que podem conduzir a uma suscetibilidade diferencial entre género para o cancro.

Em conclusão, fomos capazes de recapitular os resultados do estudo interno de dados genómicos e epigenómicos, usando um coorte independente do TCGA. Com a análise de expressão diferencial entre géneros, fornecemos novas pistas sobre o risco diferencial que está subjacente ao cancro gástrico. O rápido crescimento da biologia de alto rendimento vai expandir ainda mais o nosso conhecimento sobre as dimensões moleculares no cancro.

Palavras-Chave: Cancro gástrico; TCGA; metilação de DNA; sequenciação de RNA; variação no número e cópias (CNV)

INDEX

Agradecimientos.....	iii
Abstract.....	v
Resumo.....	vii
Index.....	ix
List of figures.....	xiii
List of tables.....	xv
List of acronyms.....	xvii
1. Introduction.....	1
1.1 Context and Motivation.....	1
1.2 Technologies used by the TCGA.....	3
1.3 Dissertation Goals.....	4
1.4 Organisation of the contents.....	5
2. TCGA data.....	7
2.1 Data types collected.....	7
2.2 Data Access.....	9
2.3 Data Level.....	9
2.4 TCGA Identifier.....	10
3. Genomic Data Types.....	11
3.1 Copy Number Variation.....	11
3.2 DNA Methylation.....	12
3.3 mRNA expression.....	12
3.4 Internal group study.....	14
4. Tools.....	15
4.1 Genome Annotation.....	15
4.2 Hierarchical Clustering.....	16
4.3 BEDTools.....	17

4.4	Wilcoxon signed rank sum test.....	18
4.5	Differential Gene Expression (DGE)	19
4.5.1	<i>NOISeq</i>	19
4.5.2	<i>DESeq2</i>	20
4.5.3	<i>EdgeR</i>	20
4.5.4	<i>TweeDEseq</i>	21
5.	Hypothesis.....	23
5.1	Validation of genomic and epigenomic patterns of GROUPSTUDY in the TCGA cohort	23
5.1.1	TCGA Copy Number Variation	23
5.1.2	TCGA DNA Methylation	23
5.1.2.1	Correlation between cohorts.....	23
5.1.2.2	Tumour vs Normal - Differential methylation	23
5.1.2.3	DNA methylation patterns	24
5.2	Gender differential expression in gastric cancer.....	24
6.	Methods	25
6.1	Validation of genomic and epigenomic patterns of GROUPSTUDY in the TCGA cohort	25
6.1.1	TCGA Copy Number Variation	25
6.1.2	TCGA DNA Methylation	25
6.1.2.1	Correlations between cohorts.....	26
6.1.2.2	Tumour vs Normal - Differential methylation	28
6.1.2.2	DNA Methylation patterns	29
6.2	Gender differential expression in gastric cancer.....	29
7.	Results	33
7.1	Validation of genomic and epigenomic patterns of GROUPSTUDY in the TCGA cohort	33
7.1.1	TCGA Copy Number Variation	33
7.1.2	TCGA DNA Methylation	34
7.1.2.1	Correlation between cohorts.....	34
7.1.2.2	Tumour vs Normal - Differential methylation	35
7.1.2.3	DNA Methylation patterns	37

7.2	Gender differential expression in gastric cancer.....	43
7.2.1	Functional enrichment analysis.....	52
8.	Discussion.....	59
	References.....	63
	APPENDIX I – <i>GroupGO</i> results.....	67

List of figures

Figure 1 - Hypothetical TCGA barcode.....	10
Figure 2 - Samples distribution for five TCGA samples in each data type analysed.....	13
Figure 3 - Example of a hypothetical clustering and hierarchical clustering.....	17
Figure 4 - General scheme with the basic processes realized in a differential expression analysis	22
Figure 5 - Scheme of consensus annotation for DNA methylation correlation tests.....	27
Figure 6 - First differential expression analysis workflow.....	30
Figure 7 - Second differential expression analysis workflow	31
Figure 8 - Plot with CNVs from stomach cancer TCGA samples.....	33
Figure 9 - Mean methylation values across samples (GROUPSTUDY vs. TCGA)	34
Figure 10 - Correlation between TCGA and GROUPSTUDY datasets	35
Figure 11 - Venn diagram with differentially methylated and significant genes for TCGA - paired, TCGA - unpaired and GROUPSTUDY.....	36
Figure 12 - Gene from the hypothesis tests in the correlation plots from TCGA and GROUPSTUDY	37
Figure 13 - Heat Maps for analysis of HM450 platform - Tumour tissue samples	38
Figure 14 - Heat Maps for analysis of HM27 platform - Tumour and Normal tissue samples.....	40
Figure 15 - Heat Maps for analysis of HM450 platform - Healthy tissue samples.....	42
Figure 16 - Heat Maps with autosomal DEGs for each method for analysis 1	49
Figure 17 - Heat Maps with autosomal DEGS for each method in analysis 2	50
Figure 18 - Box plot with DEGs found in edgeR.....	52
Figure 19 - GO and KEGG analysis for DEGs from tumour vs normal tissue samples	53
Figure 20 - GO and KEGG analysis for DEGs from male vs female – tumour tissue samples.....	54
Figure 21 - Folate biosynthesis pathway	55
Figure 22 - GO and KEGG analysis for DEGs from tumour vs normal tissue samples by gender	56
Figure 23 - Gastric acid secretion pathway	57
Figure 24 - Linoleic acid metabolism pathway	58
Figure 25 - GroupGO (Biological Processes) analysis for DEGs obtained by edgeR – Paired Tumour samples Male vs Female	67

Figure 26 - GroupGO (Molecular Functions) analysis for DEGs obtained by edgeR – Paired Tumour samples Male vs Female	68
Figure 27 - GroupGO (Cellular Components) analysis for DEGs obtained by edgeR – Paired Tumour samples Male vs Female	68
Figure 28 - GroupGO (Biological Processes) analysis for DEGs obtained by edgeR – Paired Normal samples Male vs Female	68
Figure 29 - GroupGO (Molecular Functions) analysis for DEGs obtained by edgeR – Paired Normal samples Male vs Female	68
Figure 30 - GroupGO (Cellular Components) analysis for DEGs obtained by edgeR – Paired Normal samples Male vs Female	68

List of tables

Table 1 - Types of assay used in each TCGA data type studied	4
Table 2 - Clinical data for patients studied in TCGA stomach cancer cohort.....	7
Table 3 - Tumour samples available by TCGA of data type studied	8
Table 4 - Paired samples available by TCGA of data type studied	8
Table 5 - TCGA Data Levels.....	9
Table 6 - Parameters used in each data type studied - TCGA	14
Table 7 - BED file constitution	18
Table 8 - DNA Methylation platforms from TCGA	26
Table 9 - Samples used in methylation hypothesis tests.....	28
Table 10 - Samples used for differential expression analysis	29
Table 11 - Selected genes and respective percentage of samples with hypermethylation – HM450 platform	39
Table 12 - Selected genes and respective percentage of samples with hypermethylation – HM27 platform	41
Table 13 - Selected genes and respective percentage of samples with hypermethylation – Healthy samples	42
Table 14 - Differential expression analysis between paired tumour and normal tissue samples	44
Table 15 - Intersection of autosomal DEGs between tissue types across methods – paired samples ...	44
Table 16 - Differential expression analysis between genders in non-paired tumour tissue samples	45
Table 17 - Intersection of autosomal DEGs between genders across all methods – non-paired tumour tissue samples	45
Table 18 - Differential expression analysis between genders in paired tumour tissue samples	46
Table 19 - Intersection of autosomal DEGs between genders across all methods – tumour paired samples	46
Table 20 - Differential expression analysis between genders in paired normal tissue samples.....	47
Table 21 - Intersection of autosomal DEGs between genders across all methods – normal paired samples	47
Table 22 - Intersection between paired tumour, normal and GTEx (male vs female)	48

Table 23 - Autosomal DEGs by gender across methods	50
Table 24 - Specific autosomal DEGs by gender across methods.....	50
Table 25 - Intersections between analysis 1 and analysis 2	51

List of acronyms

BED – Browser extensible data

BCR – Biospecimen core resource

BP – Biological process

CC – Cellular component

cDNA – complementary DNA

CIN – chromosomal instability

CLA - Conjugated linoleic acid

CNV – Copy number variation

DNA - Deoxyribonucleic acid

DGE – Differential gene expression

DEG – Differentially expressed gene

EBV - Epstein - Barr virus

EGFR - Epidermal growth factor receptor

ERIC – Expression regulation in cancer

FDR – False discovery rate

GO – Gene ontology

GS - Genomically stable

HC – Hierarchical clustering

HM27 – Illumina HumanMethylation27

HM450 - Illumina HumanMethylation450

ICGC - International cancer genome consortium

IPATIMUP - Instituto de Patologia e Imunologia Molecular da Universidade do Porto

KEGG - Kyoto Encyclopedia of genes

LR – Log-ratio

NCI – National cancer institute

NHGRI - National Human Genome Research Institute

NGS – Next-Generation sequencing

MF – Molecular function

MMR - Mismatch repair

mRNA – Messenger RNA
MSI - Microsatellite instable
Probeid – Probe identifier
PT – Poisson-Tweedie
 r – Pearson correlation coefficient
Refid – Reference identifier
RefSeq – NCBI Reference sequence database
RNA – Ribonucleic acid
RNA-Seq - RNA sequencing
RHOA - Ras homolog gene family, member A
RPKM - Reads per kilobase per million mapped reads
RRBS - Reduced representation bisulfite sequencing
RTK - Receptor tyrosine kinase
SNP – Single nucleotide polymorphism
STAD - Stomach adenocarcinoma
TCGA – The cancer genome atlas
TMM - Trimmed mean of M values
TSS - Tissue source site
UCSC - University of California Santa Cruz
UUID - Universally unique identifiers
WHO – World health organisation

1. INTRODUCTION

1.1 Context and Motivation

Cancer is a disease determined by genetic and epigenetic alterations, inherited or acquired (Wang, 2013). These alterations can happen through several mechanisms: copy number changes, chromosomal rearrangements, DNA methylation of CpG islands and DNA sequence changes (McLendon et al., 2008).

Cancer genome alterations can be divided into two classes: driver mutations, which cause tumour growth, and passenger mutations, which do not necessarily offer a growth advantage (Wang, 2013).

Tumours are characterised by the presence of cell populations that suffer uncontrolled division and show the potential to invade other tissues (Soper and Rasooly, 2016). Each tumour is unique and typically has a large number of genetic alterations. Nevertheless, not all these lesions drive proliferation and metastasis (Akavia et al., 2010). The identification of driver mutations will offer perceptions into cancer biology and highlight new drug targets and diagnostic tests (Hudson (Chairperson) et al., 2010).

Gastric cancer is currently the fifth most diagnosed cancer and the third principal cause of cancer-related death in the world (GloboCan, 2012). Reaching a detailed molecular understanding of gastric cancer pathogenesis is crucial to improving patient outcome, but this goal has been hampered due to the histological and etiologic heterogeneity of these tumours (Tan and Yeoh, 2015).

The majority of the gastric cancers are adenocarcinomas (Bass et al., 2014). These have been subdivided into intestinal and diffuse types according to the Lauren classification (Zhang, 2014). An alternative classification, proposed by the World Health Organisation (WHO), divides gastric cancer into papillary, tubular, mucinous and signet-ring cell carcinomas (Brambilla et al., 2001). Nonetheless, these classifications have shown limited clinical utility (Zhang, 2014).

The Cancer Genome Atlas (TCGA) network, is a project funded by the United States National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) (Zhang, 2014), and aims to catalogue and discover key cancer-causing genome alterations in large cohorts of human tumours using high-resolution microarrays and Next-Generation Sequencing (NGS) platforms (McLendon et al., 2008). Stomach adenocarcinoma is one of the 25 main cancer types to be widely characterised by these platforms (Zhang, 2014).

With the large variety of rare mutations found in cancer cells, genome-wide studies for identifying cancer driver mutations demand to sequence large cohorts of patients (Vandin et al., 2012). Through the use of large cohorts, DNA sequencing allowed uncovering a list of frequent genomic alterations (e.g. amplifications, deletions, translocations) (Chang et al., 2013).

The recent landmark TCGA study (Bass et al., 2014) revealed data from 295 gastric cancer that were simultaneously profiled on multiple molecular platforms. Six molecular platforms were characterised: array-based somatic copy number analysis, microRNA (miRNA), array-based DNA methylation, whole-exome sequencing, messenger RNA sequencing and reverse-phase protein array (RPPA). About 77% of the tumours for all platforms were tested. This study proposed a classification of gastric cancer into four genomic subtypes: Epstein-Barr virus (EBV), genomically stable (GS), microsatellite unstable (MSI) and chromosomal instability (CIN).

Tumours positive for EBV pinpoints the viral etiology of gastric cancer (Zhang, 2014), which shows common *PIK3CA* mutations, higher occurrence of DNA hypermethylation than other subtypes and amplifications of *JAK2*, *CD274* and *PDCD1LG2* (Bass et al., 2014).

GS tumours are best represented by the diffuse type of gastric cancer, with a lower number of mutations when compared to other subtypes (Choi, 2015). Mutations in *CDH1* and in the *Ras homolog gene family, member A* (RHOA) gene or fusions which involve *RHO-family GTPase-activating* proteins are characteristics of GS (Bass et al., 2014).

MSI tumours show elevated mutation rates (Bass et al., 2014). Mutations in *PIK3CA*, *ERBB2*, *ERBB3* and *epidermal growth factor receptor* (EGFR) were identified (Zhang, 2014). These mutations are normally related to the loss of function of mismatch repair (MMR) genes and associated with the intestinal type, women, and older age (Choi, 2015).

In tumours with chromosomal instability, genomic amplifications of receptor tyrosine kinases (RTKs) were identified (Zhang, 2014). Frequent mutations in TP53 are also related with CIN subtype (Choi, 2015).

In fact, these subtypes appear to be different at the molecular level, suggesting not only that molecular processes driving tumorigenesis can differ between patients but also that treatment may need to be tailored for each subtype of tumour (Shmulevich, 2014).

Further goals of TCGA include: the development and application of new technologies, identification of cancer-specific molecular alterations, production of data and results freely available to the scientific community (Chang et al., 2013).

Each project assays a large number of cases in different data types: the sequencing of the exome, copy number variation (CNV), DNA methylation, mRNA expression, microRNA expression and protein expression (Shmulevich, 2014).

1.2 Technologies used by the TCGA

Two types of assays to measure the six data types from the TCGA project were used: array-based or sequenced-based.

Within array-based methods, DNA microarrays are a collection of DNA probes arrayed on a solid support. These probes (radioactively labelled) can be a small fragment of a gene and are used to hybridise a complementary DNA (cDNA). Microarray probe intensity is assumed to be proportional to the concentration of the transcript (Gresham et al., 2008). High throughput, speed and automation are advantages of microarray techniques (Schrenzel et al., 2009).

Within sequence based methods, RNA-Sequencing (RNA-Seq) is a technique for whole-genome sequencing transcriptome profiling (Li et al., 2010). RNA-Seq is the direct sequencing of transcripts by sequencing technologies (Zhao et al., 2014).

Microarrays are less efficient in detecting genes with low expression level due to cross-hybridization while RNA-Seq is more sensitive, detecting genes with low expression and being very accurate on extremely abundant genes. The detection of new transcripts, isoforms, splice variants and allele-specific expression are some advantages of RNA-Seq avoiding some issues like the cross-hybridization (Zhao et al., 2014). The assays used in each data type by TCGA are listed in Table 1.

Table 1 - Types of assay used in each TCGA data type studied
CNV and DNA methylation were measured by microarray assays while mRNA expression through RNA-Seq.

Data type	Platform	Assay
CNV	Affymetrix Genome-Wide Human SNP Array 6.0	Microarray - measured by Single Nucleotide Polymorphisms (SNP) arrays
mRNA expression	IlluminaHiSeq_RNASeqV2	RNA-Seq
DNA Methylation	Illumina Infinium HumanMethylation27 Illumina Infinium HumanMethylation450	Microarray

1.3 Dissertation Goals

In this dissertation, we perform two major types of analyses:

a) **Recapitulate genomic and epigenomic results previously found in an internal study with data from a TCGA independent cohort.** This internal project used 50 Tumour/Normal tissue samples of individuals from Portuguese origin. It made available for each sample three data types: CNV, DNA methylation and mRNA expression. However, only genomic and epigenomic platforms were analysed in this dissertation. There is currently a study underway with the results of genomic and epigenomic data analysis made in this dissertation and they will be published in the near future. For reasons of confidentiality, full details of the results cannot be provided here. Therefore, all the results obtained in this dissertation from this study will be masked;

b) **Understand gender differential expression in gastric cancer.** It is well established that gender influences the response to cancer treatment. Nevertheless, the molecular causes are still undiscovered and therapies to cancer are usually carried out without taking into account the gender of the patient. In another research line in the group, we are trying to understand the molecular differences between genders that could explain a predisposition to cancer. Data from a TCGA cohort were used for this analyses.

In order to achieve the previous goals, the following general tasks were performed: i) TCGA data was accessed; ii) this data was processed in order to allow an efficient analysis and interrogation; iii) genetic patterns found in an independent study were validated, and iv) differential expression analysis was performed between two males and females.

1.4 Organisation of the contents

Chapter 2. TCGA data

Comprehensive review of the TCGA data with the description of the data collected, how the TCGA data were divided (levels and access) and how the TCGA samples were identified.

Chapter 3. Genomic data types

The description of the genomic data types from the TCGA cohort and the internal group study cohort, used in this dissertation. Descriptions of how the data were obtained, thresholds and range of the values for each platform.

Chapter 4. Tools

The description of the tools used in this dissertation. Several types of bioinformatics tools were described. Two different programming languages were used: R system and batch (command-line) processing.

Chapter 5. Hypothesis

The hypothesis that stood at the core of this work and the specific questions for each data type studied.

Chapter 6. Methods

The developed methods used in this dissertation were described.

Chapter 7. Results

The results obtained for each analysis performed in this dissertation.

Chapter 8. Discussion

A global analysis of this work was described, along with possible improvements and future work.

2. TCGA DATA

TCGA is an open data resource that provides publicly available cancer genomic data classified by data type (e.g. gene expression, methylation, somatic mutations) and data levels that allow a structured and easy access to this resource with suitable patient confidentiality protection (Bass et al., 2014).

2.1 Data types collected

TCGA examines a very large number of samples, aiming for 500 samples for each tumour type. This sample size offers the statistical power needed to produce comprehensive genomic profiles. The same samples are studied by different TCGA research teams, which can result in a most complete and consistent view of cancer genomes since every sample is analysed several times and for each platform (National Institute of Health, 2016a).

A Tissue Source Site (TSS) collects samples (tissue, cell or blood) and clinical metadata from eligible patients. Next, these samples are sent to a Biospecimen Core Resource (BCR), which is a TCGA centre where samples are carefully catalogued, processed, quality-checked and stored along with participant clinical information (Tomczak et al., 2015).

Table 2 shows the general clinical data for 445 patient studied in the TCGA cohort.

Table 2 - Clinical data for patients studied in TCGA stomach cancer cohort
This table consists of the most important clinical data from the patients of the TCGA cohort.

Clinical Data		Absolute value
Race	White	278
	Black or African American	13
	Asian	89
	NA	64
	Other	1
Gender	Male	285
	Female	158
	NA	2
Ethnicity	Not Hispanic or Latino	318
	Hispanic or Latino	5
	NA	112

Family history of gastric cancer	Yes	18
	No	324
	NA	103
Age at initial diagnosis	<50	35
	51-60	107
	61-70	143
	>70	153
	NA	5
Tumour status	Tumour free	291
	With tumour	92
	NA	62

For most cases of TCGA, only tumour tissue samples were available (Table 3). However, for some TCGA cancer samples studied, normal adjacent tissue samples and samples of cancerous tissue were collected (Table 4). Pairs of matched normal and tumour tissue allow researchers to identify the genomic changes that might play a role in the development of studied cancer (National Institute of Health, 2016a).

Table 3 - Tumour samples available by TCGA of data type studied

Tumour samples quantity present in each platform and existing samples between two platforms.

Tumour Samples	CNV	DNA Methylation	mRNA expression
CNV	442		
DNA Methylation	441	443	
mRNA expression	412	415	416

Table 4 - Paired samples available by TCGA of data type studied

Paired samples quantity present in each platform and existing samples between two platforms.

Paired Samples	CNV	DNA Methylation	mRNA expression
CNV	411		
DNA Methylation	26	27	
mRNA expression	33	0	35

2.2 Data Access

There are two data access tiers for TCGA data: Open Access data tier and Controlled Access data tier. The Open Access tier contains public data, which is not unique to an individual such as gene expression data, genotype frequencies, copy number alterations by regions, and epigenetic data.

Controlled Access tier requires user authorization in order to access them and includes data that impacts the confidentiality and privacy of the individuals like primary sequencing data, processed Exon array data, clinical text fields and individual germline variant data (National Institute of Health, 2016b). In this dissertation, we only focus on the Open Access data tier.

2.3 Data Level

Data level is a data classification method used in the TCGA network to assist researchers in communicating and locating their data of interest. There are four data levels (Table 5): Level 1 (raw data), Level 2 (processed data), Level 3 (segmented or interpreted data) and Level 4 (region of interest data) (Klinger, 2014).

While the access to most of the level 1 and 2 data is restricted, the entire level 3 data are publicly available (Zhu et al., 2014). Some data types can only be accessed in some levels like CNV and mRNA expression. However, data types like DNA methylation allows access to all levels. The data used in this project will normally be the most processed and normalised data (Ayala, 2014a).

Table 5 - TCGA Data Levels

The first three levels apply to individual sample while level four analyses across sample sets. Level 3 is the most used level. Based on (Klinger, 2014).

Data Level	Level type	Description
1	Raw, not normalised data	Not normalised low-level data for single sample
2	Processed data	Normalised data for single sample; verification of existing molecular abnormalities
3	Segmented/Interpreted data	Set of processed data from single sample grouped by probed loci
4	Summarised Data	Quantified association between classes of samples based on molecular abnormalities, sample characteristics or clinical variables

3. GENOMIC DATA TYPES

The data generated from tumours in the TCGA project are very large in volume. For the molecular characterization of gastric cancer different genomic data types and molecular analysis different technologies were used (Shmulevich, 2014) that are reviewed in this chapter.

3.1 Copy Number Variation

CNVs are usually found in tumour tissues and involve losses (deletions) or gains (amplifications) of one or both copies of chromosomal regions. CNVs change the quantity and organisation of genomic material, which can cause an alteration of the transcriptional activity that may affect critical genes (Singh and Salnikova, 2015).

Deletions, duplications, inversions, or translocations, can be causal alterations, possibly changing the gene function or affecting large chromosomal regions (Singh and Salnikova, 2015).

Segments of copy number data (level 3) from the TCGA website with intensities for each array probe were downloaded. The obtained files were classified as *nocnv* or *cnv* using Hg18/Hg19. *Nocnv* means that a fixed set of probes that frequently contain germline CNVs are removed before segmentation, while *cnv* has both germline and somatic variations (Zhu and Ji, 2014). Hg18 or Hg19 specifies that in preparation for segmentation, the probes are organised based on the order of reference genome HG18 or Hg19 respectively (Human Genome) (Broad Institute of MIT and Harvard, 2016).

Segment means are measured by *log ratio* (LR) also called *log-2 ratio*. LR can be calculated by the standard formula: $\log_2(\text{observed intensity}/\text{reference intensity})$, which is the ratio of probes in the segment (Laddha et al., 2014).

The objective is to define regions in the genome where the sample mean LR value is different from the reference. A mean LR of zero means that a sample has the same number of copies as the reference. When the LR segment mean is higher than zero represents a copy number gain while an LR segment mean lower than zero indicates copy number loss (Bozeman, 2014).

3.2 DNA Methylation

DNA methylation is an epigenetic mechanism (Huang et al., 2015) that plays an important role in cancer and occurs when methyl groups are attached to cytosine bases at the C-5 position, usually in a CpG sequence context. Genes with high levels of methylation in their promoter region tend to be transcriptionally silent (Jin et al., 2011).

TCGA used two platforms to obtain DNA methylation profiles: Illumina HumanMethylation27 (HM27) and Illumina HumanMethylation450 (HM450) (Zhu and Ji, 2014).

The HM27 and HM450 arrays target 27,578 and 482,421 CpG sites, respectively. Level 2 contains a summary of intensities for methylated (M) and unmethylated (U) data. Detection probabilities (p -values) were also calculated. Level 3 data contains beta values for each locus with annotations for the gene symbol, chromosome and CpG coordinate (Bass et al., 2014).

For each locus, the score for DNA methylation is represented as a beta (β) value ($\beta = M / (M+U)$), in which M represents the mean methylated signal intensities and U indicates the same but for unmethylated signal (Bass et al., 2014).

Beta-values range from 0 and 1 (or 0 and 100%) and under ideal conditions a value of zero indicates no DNA methylation, while a value of one indicates complete DNA methylation (Du et al., 2010).

In methylation arrays, detection p -values reflect the strength of DNA hybridization. It can be evaluated through the comparison between the CpG-intensity and the intensities of negative control probes. Non-significant detection p -values (greater than 0.05) typically means bad hybridization, bad probe design or chromosome abnormalities and is masked as "NA" on the array (Du et al., 2014).

3.3 mRNA expression

Messenger RNA carries the genetic information derived from DNA in the form of code (in the form of a sequence of three base pairs) which specifies a particular amino acid (Lodish et al., 2000).

RNA-Seq is a technology for transcriptome profiling. RNA-Seq can rapidly identify and quantify rare and common transcripts and non-coding RNAs, among a large range of samples. NGS quantifies discrete, digital sequencing read counts offering consistent and complete results (Tomczak et al., 2015).

The quantification of RNA-Seq data is made by a high-throughput sequencer which generates millions of reads from the cDNA fragments (Li et al., 2010). Next, there is the quantification of expression

by *Reads per kilobase per million mapped reads* (RPKM). RPKM is a method of quantifying gene expression from RNA-Seq data by normalising for total read length and the number of sequencing reads (Zhao et al., 2014).

For this data type, RNASeqV2 data (instead of RNA-Seq, which uses a different set of algorithms to determine the expression levels) was downloaded. Level 3 data was available and files with *normalised_results* extension were used (files containing gene IDs and normalised counts for gene expression) (Pihl, 2013).

To better understand the range and the underlying distribution of the values in these technologies, we have analysed five randomly selected samples for each data type (Figure 2).

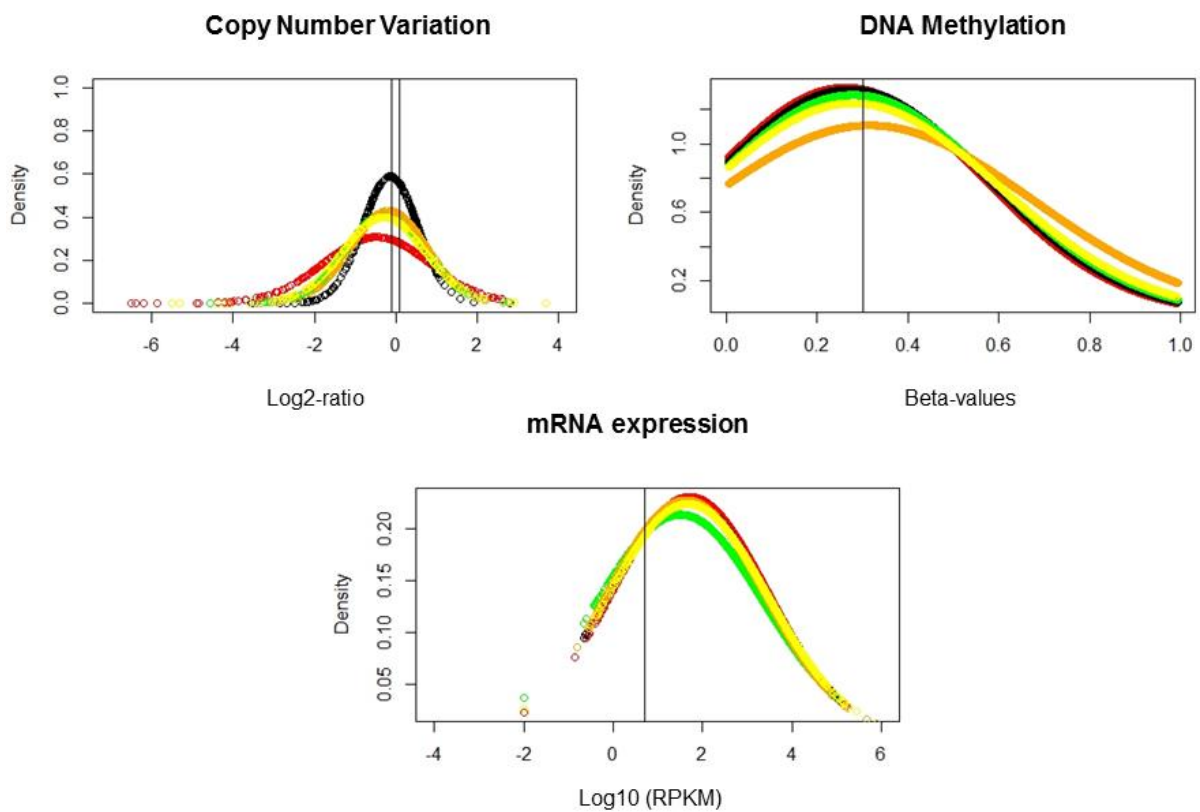


Figure 2 - Samples distribution for five TCGA samples in each data type analysed

a) CNV- log2-ratio range from [-9:7], follow a normal distribution c) mRNA expression – RPKM range from [-2:7], follow a log normal distribution; c) DNA Methylation – beta values range from [0:1], follow a log normal distribution.

Table 6 summarizes the parameters used in each data type analysis for the TCGA data.

Table 6 - Parameters used in each data type studied - TCGA

These are the parameters used by TCGA in each data type. Bass et al used the described thresholds (Bass et al., 2014) in the study of gastric adenocarcinoma and will be the considered values.

Data type	Parameters	Description (Thresholds)
CNV	Log-2-ratio (LR)	LR \geq 0.1 : copy number gain LR \leq - 0.1 : copy number loss - 0.1 < LR < 0.1 : no copy number variation
DNA Methylation	Beta-value (β)	$\beta \geq$ 0.3 : Hypermethylated $\beta <$ 0.3 : Hypomethylated
mRNA expression	RPKM	RPKM > 5 : eligible for expression analysis

3.4 Internal group study

The internal group study (GROUPSTUDY) is a study developed by elements within the Expression Regulation in Cancer (ERIC) group from IPATIMUP-I3S. The study consists of 50 Tumour/Normal samples of gastric cancer of individuals from Portuguese origin. For each of these individuals, three types of data such as CNV, DNA methylation, and mRNA expression were analysed. CNVs were obtained through whole genome sequencing while DNA methylation data used *Reduced Representation Bisulfite Sequencing* (RRBS), different technologies from the ones used on TCGA. These two platforms are the ones studied in this dissertation. The analyses (genomic and epigenomic data like CNV and DNA methylation, respectively) performed in this dissertation were conducted to validate and better understand the results of this study.

4. TOOLS

To achieve the goals that we set ourselves, we used several types of bioinformatics tools, of distinct complexity. The programs that will be described here were used in two different ways: R system and batch (command-line) processing. In R, we used R programming mostly to treat data and R packages to perform the actual analysis. Just one tool was used in batch processing, BEDTools.

4.1 Genome Annotation

After sequencing the genome, the next logical step is to perform the respective annotation (Boundless Microbiology, 2016). The genome must be annotated for genes, or described, in a way that all biologists can use them (Stein, 2001).

Genome annotation is the process which identifies the locations and structure of genes and every coding region in a genome (Boundless Microbiology, 2016). Available genomic annotations in the human genome are: Ensembl (Hubbard et al., 2002), RefSeq (Pruitt et al., 2007) and GENCODE (Harrow et al., 2012). Further annotation includes gene function being commonly represented by Gene Ontology (GO) (Gene Ontology Consortium, 2004) which includes molecular functions, biological processes and cellular locations (Beaver et al., 2010).

Functional Enrichment Analysis uses statistical methods to find enriched functional annotations among the provided list (e.g., metabolic pathways). The analysis depends on biological databases (e.g. GO, Kyoto Encyclopedia of genes (KEGG)) with the information of genes and associated functions (Tipney and Hunter, 2010). In this dissertation, functional enrichment analysis was done using the *clusterProfiler* package.

4.1.1 clusterProfiler:

In the postgenomic era, high-throughput experimental techniques such as microarray and RNA-Seq generate a huge volume of data, driving to the development of techniques to capture biological information. Searching for shared function between genes by incorporating biological knowledge provided by biological ontologies is a typical way of doing this. For example, GO (Gene Ontology Consortium, 2004)

annotates genes to biological processes (BP), molecular functions (MF), and cellular components (CC), and KEGG (Kanehisa et al., 2010) annotates genes to pathways, being widely used to achieve this goal.

The *clusterProfiler* package provides a gene classification method called *groupGO*, to classify genes based on GO distribution at a specific level. Functions such as *enrichGO* and *enrichKEGG*, to calculate enrichment test for GO terms and KEGG pathways based on hypergeometric distribution, are also possible to use (Yu et al., 2012).

This package adjusts the estimated significance level to account for multiple hypothesis testing. Therefore, *q*-values (Storey, 2002) are estimated to control false discovery rate (FDR). The FDR is the expected proportion of false positives tests declared statistically significant in which the null hypothesis is actually true (Glickman et al., 2014).

Several visualisation methods are supported, such as *barplot*, *cnetplot* and *enrichMap* (Yu et al., 2012).

4.2 Hierarchical Clustering

Cluster analysis is a technique for data reduction that divides data into groups (clusters) which are relatively homogeneous between them and heterogeneous with each other (Figure 3a) (Yim and Ramdeen, 2015).

The goal is that within the same group, objects will be more similar, while different from the objects in other groups (Hale, 1981). To make this division, it is necessary to choose the variables according to required similarities within a group (Norusis, 2009).

There are two key methods: hierarchical and non-hierarchical cluster analysis. To form a hierarchy of clusters is the goal of hierarchical clustering, combining cases into homogeneous clusters by merging them together sequentially (Figure 3b). Non-hierarchical clustering techniques (e.g. k-means clustering) establish a primary set of cluster means assigning each case to the closest cluster mean (Yim and Ramdeen, 2015).

Strategies for hierarchical clustering generally are conceptualised into two types, agglomerative or divisive. Agglomerative hierarchical clustering separates each observation into its own cluster and at successive steps, the similar clusters are merged, until every case is grouped into one single cluster. Divisive hierarchical clustering works in an opposite way with all observations starting in one large cluster

and being separated into groups of clusters (Yim and Ramdeen, 2015). The results of hierarchical clustering are typically presented in a dendrogram (Figure 3b) (Hale, 1981).

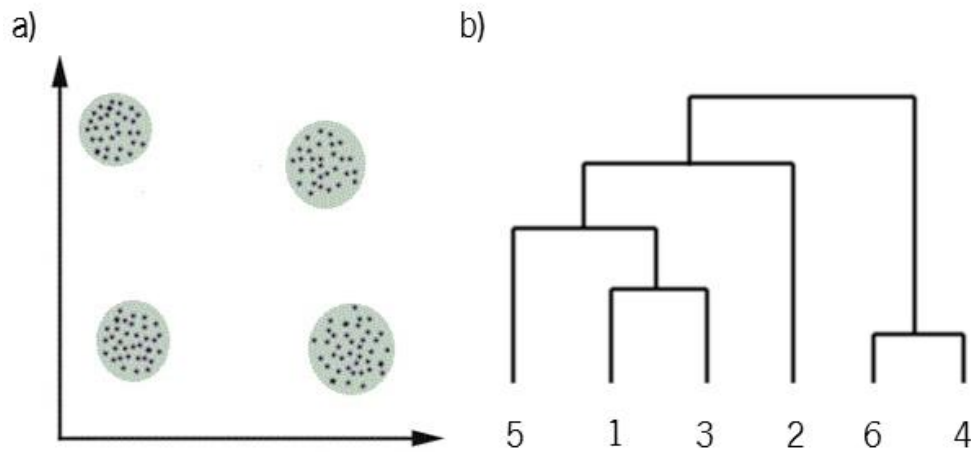


Figure 3 - Example of a hypothetical clustering and hierarchical clustering

a) Hypothetical clustering with four distinct groups. Based on (Hale, 1981) b) Hypothetical dendrogram of an agglomerative hierarchical clustering. Based on (Yim and Ramdeen, 2015).

4.3 BEDTools

Genomic features are normally represented by the Browser Extensible Data (BED) format and could be compared using the University of California Santa Cruz (UCSC) Genome Browser (Karolchik, 2003) 'Table Browser'.

BED format offers a flexible way to define the data lines that are shown in an annotation track. BED lines require three fields and nine could be used as additional fields (we describe here only three of nine because those are the fields that we used). The three required fields are *chrom*, *chromStart* and *chromEnd*, while the additional fields are the *name*, *score*, *strand*, *thickStart*, *thickEnd*, *itemRgb*, *blockCount*, *blockSizes*, and *blockStarts* (Table 7).

An example of a BED 6 file could be:

chr19	58864866	58866865	A1BG	10	+
--------------	-----------------	-----------------	-------------	-----------	----------

Table 7 - BED file constitution

BED files consists of one line per feature, each containing 3-12 columns of data separated by tab.

	Fields	Fields Description	Example
Required	chrom	The name of the chromosome	chr19
	chromStart	The starting position of the feature in the chromosome	58864866
	chromEnd	The ending position of the feature in the chromosome	58866865
Optional	name	The name of the BED line	A1BG
	score	The feature score	0-1000
	strand	The feature strand	'+' or '-'

This tool offers a consistent method for such analyses; however, it is difficult to work with large datasets. The necessity to work with high data volume produced by current DNA sequencing technologies drove to BEDTools (Quinlan and Hall, 2010) development.

BEDTools is a fast and flexible set of utilities for common operations on genomic features. This tool supports an extensive variety of operations. One of the most used is *intersectBed*, which returns the overlap between two BED files. Others like *mergeBed* (merges overlapping features into a single feature) or *sortBed* (sort BED files in useful ways) can be used.

The speed and wide functionality of BEDTools allow better flexibility in genomic comparisons between ever-larger genomic datasets.

4.4 Wilcoxon signed rank sum test

The Wilcoxon signed rank sum test is a nonparametric or distribution-free test. It is used to test the null hypothesis that the median of a distribution is equal to some value. There is the option to test for paired samples, depending on the samples to be used.

This method consists of three steps such as: 1) Ranks all observation in increasing order of magnitude, independently of the group; 2) if groups differ in size, the ranks of the smaller group are added, if their size is equal, either one can be chosen 3) Calculate the p -value.

4.5 Differential Gene Expression (DGE)

Gene expression is a biochemical process that determines which genes are actively transcribed into mRNA (and then translated into proteins) (Deraitus and Freeman, 2001).

Gene expression can be quantified by software packages such as *edgeR*, *DESeq2*, *NOISeq* and *tweeDEseq*. These methods consist in the determination of changes in gene expression (upregulation and downregulation) between a given set of samples (Houle, 2008).

NOISeq implements a non-parametric test while *DESeq2*, *edgeR* and *tweeDEseq* use parametric methods. While *DESeq2* and *edgeR* use a negative binomial distribution, *tweeDEseq* uses the Poisson-Tweedie (PT) distribution. These packages are publicly available at the Bioconductor repository (Gentleman et al., 2004).

4.5.1 *NOISeq*

NOISeq is a method for analysing count data coming from NGS technologies. This package is divided into three modules: 1) Quality control of count data; 2) Normalisation and low-count filtering, and 3) Differential expression analysis. The *NOISeq* method is nonparametric and data-adaptive. Thus, as there are no distributional assumptions to be done for the data, differential expression analysis can be made for raw counts or previously normalised datasets (Tarazona et al., 2014).

The package includes two robust non-parametric approaches for differential expression analysis: *NOISeq* and *NOISeqBIO*. The first is used when there are no biological replicates, while the second when there are biological replicates (Tarazona et al., 2015). When no replicates are available, the methods simulates technical replicates based on multinomial distribution (Tarazona et al., 2011).

NOISeq makes a null distribution of count changes through the comparison between the numbers of reads of each gene in samples in the same condition. This distribution is then used to evaluate if the change in count number represents or not a differential expression (Tarazona et al., 2011).

4.5.2 DESeq2

The package *DESeq2* provides methods to test for differential expression using a negative binomial distribution and a shrinkage estimator (raw estimate is improved by combining it with other information) for the distribution's variance (Anders and Huber, 2010).

It is divided into different steps as: 1) Input data - this package expects count data as obtained (raw counts of sequencing reads) because only the raw counts allow assessing the measurement precision correctly; 2) Normalisation - a geometric mean is calculated for each gene across all samples. The counts for a gene in each sample are then divided by this mean, allowing to the effective total number of reads estimation; 3) Variance Estimation – a dispersion value is estimated for each gene through a model fit procedure, using shrinkage estimation. It is necessary to have biological replicates of each test to estimate dispersion correctly. The inference depends on an estimation of the association between the data's variance and their mean, or vice versa, and 4) Inference – after calculating dispersion for each gene, it is possible to look for DEGs, according to some chosen threshold for the FDR, p value and fold-change (Anders and Huber, 2012).

4.5.3 EdgeR

EdgeR is a package that performs differential gene expression using count data under a negative binomial model.

This method can be divided into several steps. These steps are: 1) Reading data and creating a *DGEList* object – only accepts raw count data to create an object with the read counts and the associated metadata; 2) Filtering and Normalisation – many genes will not be expressed, or will not have enough reads, so removing these genes it is important. The trimmed mean of M values (TMM) is used for sample normalisation (method of averaging that removes a small designated percentage of the largest and smallest values before calculating the mean); 3) Estimating Dispersion - using an empirical Bayes method to shrink the genewise dispersion estimates, the dispersion is calculated; 4) Differential Expression – in this step, the DEGs are found using the FDR and fold-change thresholds (Pereira and Rueda, 2015).

4.5.4 *TweeDEseq*

TweeDEseq method offers statistical trials to test differential expression in RNA-Seq count data. It uses the Poisson-Tweedie (PT) family of distributions as the statistical model for count data. The PT distribution allows one to test for the goodness of fit to a particular count data distribution defined by a specific value of the PT shape parameter a (Puig and Valero, 2007).

The *tweeDEseq* package provides a function to normalising count data using other functions from the *edgeR* package – using TMM (it is necessary to have *edgeR* packaged installed). This package contains a function for testing differential expression among two different conditions using a score based test (Esnaola et al., 2013).

Figure 4 is a general scheme with basic steps for differential expression analysis using the methods described anteriorly for gene differential expression. In this scheme, is also represented the step of the functional enrichment analysis. These methods require read counts as input. As first steps, filtering (removing lowly expressed genes) and normalisation are performed on the input data. Prior to the differential expression analysis, thresholds are selected for the parameters (*p-value*, *q-value*, *FDR* and *fold-change*) that each method require. After obtaining the significant DEGs, the functional enrichment analyses for GO and KEGG is performed.

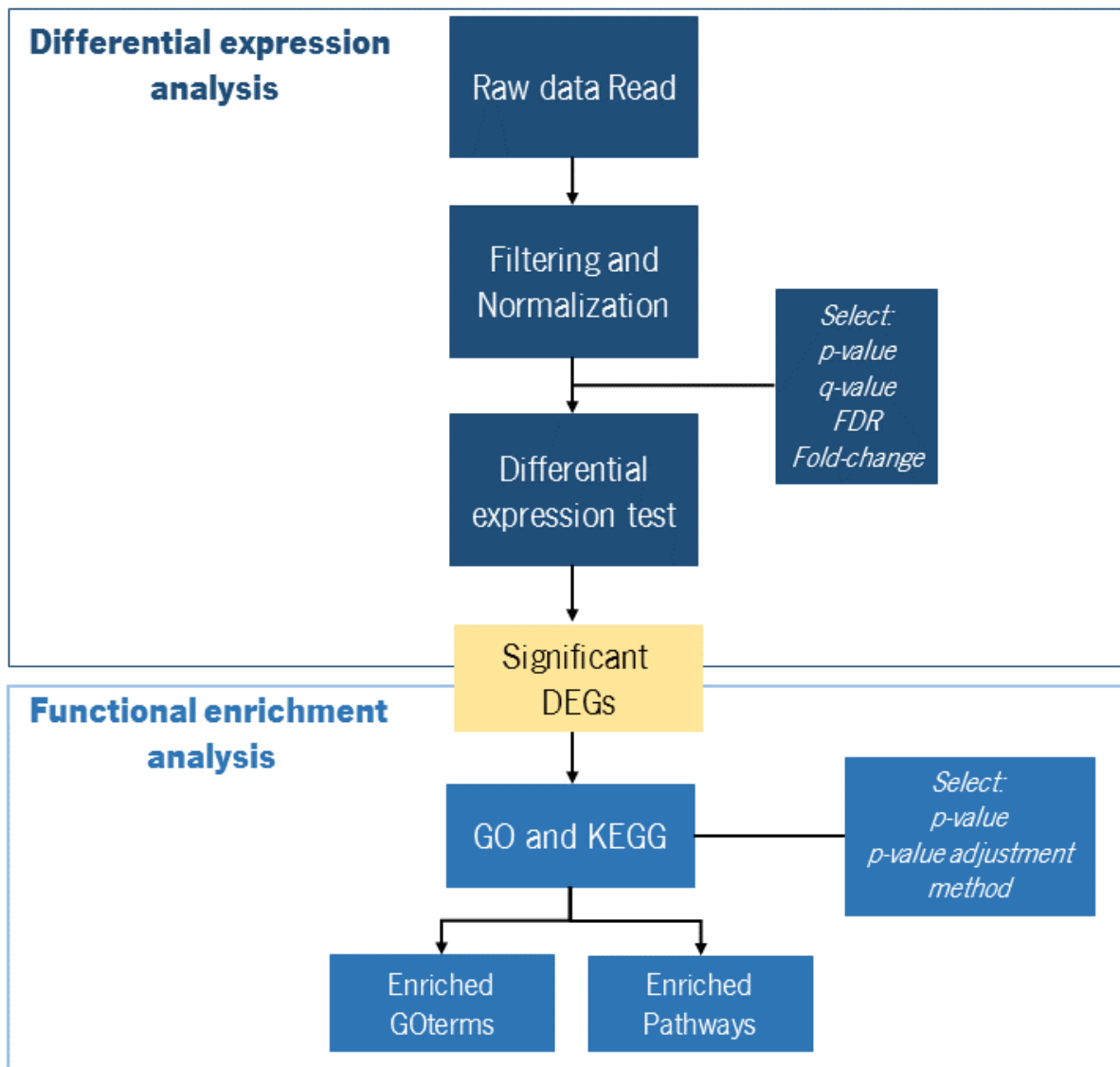


Figure 4 - General scheme with the basic processes realized in a differential expression analysis
 This scheme represents the basic step for a differential expression analysis. We divided the scheme in two different processes: i) differential expression analysis - significant DEGs are obtained; ii) Functional enrichment analysis – enriched GO terms and analysis.

5. HYPOTHESIS

In this section, we introduced the hypothesis that stands at the core of this work and the specific questions for each data type that we studied in this dissertation.

5.1 Validation of genomic and epigenomic patterns of GROUPSTUDY in the TCGA cohort

5.1.1 TCGA Copy Number Variation

The goal of this analysis is to test if the results found in the GROUPSTUDY can be recapitulated in a different cohort (in this case the TCGA). We hypothesise that when a gene X is amplified, there is a set of genes that are co-amplified with this gene (in GROUPSTUDY, we observed that gene X was amplified in ~20% of stomach cancer samples) and that a proportion of samples bearing amplification of gene X, concomitantly and recurrently displayed amplification of large regions of the same chromosome and other specific chromosomal regions.

5.1.2 TCGA DNA Methylation

5.1.2.1 Correlation between cohorts

We wanted to test if the methylation patterns in the GROUPSTUDY cohort can be recapitulated in the TCGA dataset. The DNA Methylation dataset from GROUPSTUDY and the TCGA DNA Methylation dataset were obtained with two different technologies, RRBS and microarrays, respectively. We hypothesise that despite the different technologies used and the difference in the size and nature of the cohorts, we can recover similar patterns of methylation.

5.1.2.2 Tumour vs Normal - Differential methylation

We have two different groups of samples from the TCGA cohort: paired samples and unpaired samples. We remind that the number of unpaired samples is almost ten times larger than the number of paired samples. Here, we wanted to investigate if the results obtained with the paired and the unpaired cohort provided similar results in terms of methylation patterns.

5.1.2.3 DNA methylation patterns

In a group study made by the ERIC group, a set of genes was found to be hypermethylated in 80% of samples. Our goal was to assess if the methylation of set of genes predominantly hypermethylated in the GROUPSTUDY cohort was also hypermethylated in the TCGA cohort.

5.2 Gender differential expression in gastric cancer

Gender is known as an important factor in the progression of cancer. Differences between males and females are not limited to anatomic differences. There are different susceptibilities to a variety of diseases and more importantly, in the response to treatment. Despite gender differential cancer incidence has been consistently described in epidemiologic studies, few cancer studies have considered the gender covariate in their design.

Typically, males are more disposed to develop cancers, in particular blood malignancies (Ma et al., 2016). Cancers such as Colorectal, Lung, Non-hodgkin Lymphoma, and Bladder or Kaposi sarcoma have the highest male to female ratio. On the other hand, Breast, Thyroid or Gall Bladder are examples of cancers with a higher incidence in females (Dorak and Karpuzoglu, 2012).

It is expected that environmental causes explain the large proportion of cancer risk. However, this alone is not enough to clarify the full gender differential cancer risk. Early onset cancers still occur in larger proportion in males (Pearce and Parker, 2001). Females commonly have more efficient immunological responses than males. This may be an important factor for the differentiated behaviour in cancer, but also in autoimmune and inflammatory disorders, infectious diseases and vaccines responses (Klein et al., 2012). The sex chromosome can be another aspect of these different effects. The X-chromosome contains the largest number of immune-related genes of the whole human genome. Given the gender differential expression of genes in the sex chromosomes, it is expected that they play an important role in the mechanisms underlying gender differential cancer susceptibility (Pinheiro et al., 2011).

In order to better understand gender differential genomic patterns and their relevance to human cancer biology, we set to investigate and identify transcriptomic differences between males and females that may be important to differential gastric cancer susceptibility.

6. METHODS

6.1 Validation of genomic and epigenomic patterns of GROUPSTUDY in the TCGA cohort

6.1.1 TCGA Copy Number Variation

To test our hypothesis that several genes and chromosomal regions are co-amplified with gene X, we downloaded *nocnv* tumour samples from TCGA for the Stomach adenocarcinoma in a total of 274 samples.

Of the 274 initial samples, we filtered samples with amplification for the gene X resulting in a set of 52 samples. We used as threshold for the segment mean, a \log_2 ratio of 0.25.

The files obtained from TCGA did not provide the genes in which there were amplifications (TCGA provides CNV information oriented for genomic regions). Thus, we had to impute amplification to genes. For this, we used the genomic coordinates of gene X. We knew the chromosome and the genomic coordinates of start and end of the gene and, thus, we were able to find the amplifications occurring in that range.

Having all required amplifications, we created a binary array with the genes and samples. Zero corresponds to no copy number amplification and one to a copy number amplification. Then, we performed a Hierarchical Clustering (Pearson correlation was the metric used) to find out the groups of samples more similar and their subsequent order.

The next step was to find amplifications on other chromosomes (we searched specifically for two chromosomes found in the GROUPSTUDY). After this step, we had all amplifications in three chromosomes (chromosome of gene X and the other two that we searched before).

The last step was to plot these amplifications using R package *gplots*.

6.1.2 TCGA DNA Methylation

Here our objective is to recapitulate results obtained in GROUPSTUDY with DNA methylation cohort using the TCGA cohort. In these analyses, we used DNA Methylation data downloaded from TCGA for the

Stomach adenocarcinoma (STAD) and the GROUPSTUDY DNA methylation samples. TCGA samples were divided into two platforms: HM450 and HM27 (Table 8).

Table 8 - DNA Methylation platforms from TCGA

The probes correspond to CpG sites throughout the genome. The HM450 platform was analysed for a greater number of genes and probes than HM27 platform. While HM450 has only tumour samples, HM27 has 25-paired samples.

DNA Methylation		
Platforms	HM450	HM27
Tumour Samples	248	25
Normal Samples		25
Genes	25,095	15,311
Probes	482,421	187,281

Two filters for both platforms were applied. First, we filtered for CpG sites that were in the promoter region. Using a BED file with the promoter annotation from the GROUPSTUDY, we intersected our probes with a file containing the promoter region (genomic coordinates of start and end of promoter region) for each gene. As a result, we obtained only probes within the promoter region. The genome browser Ensembl (Hubbard et al., 2002) was used to search for the regions in which the probes were located for our set of genes. Some of these probes were located in a region next to the promoter. If the probes were in the region before the initiation codon (untranslated region), they were considered. Genes without information in 50% or more of the samples were filtered.

After applying these filterers, we obtained our TCGA DNA Methylation dataset.

6.1.2.1 Correlations between cohorts

With the goal of testing data consistency between the GROUPSTUDY DNA methylation dataset and the TCGA DNA methylation dataset, we decided to do correlation tests between datasets.

First, we needed to process the data from both datasets. As they used different technologies to assess methylation status, different annotations for the genes were used. The GROUPSTUDY cohort was divided by gene regions (transcripts) represented by reference identifiers (*refids*). The TCGA cohort was divided by probes (fragment of DNA or RNA of variable length) and was represented by probe identifiers (*probeids*). Therefore, we needed to create a consensus annotation.

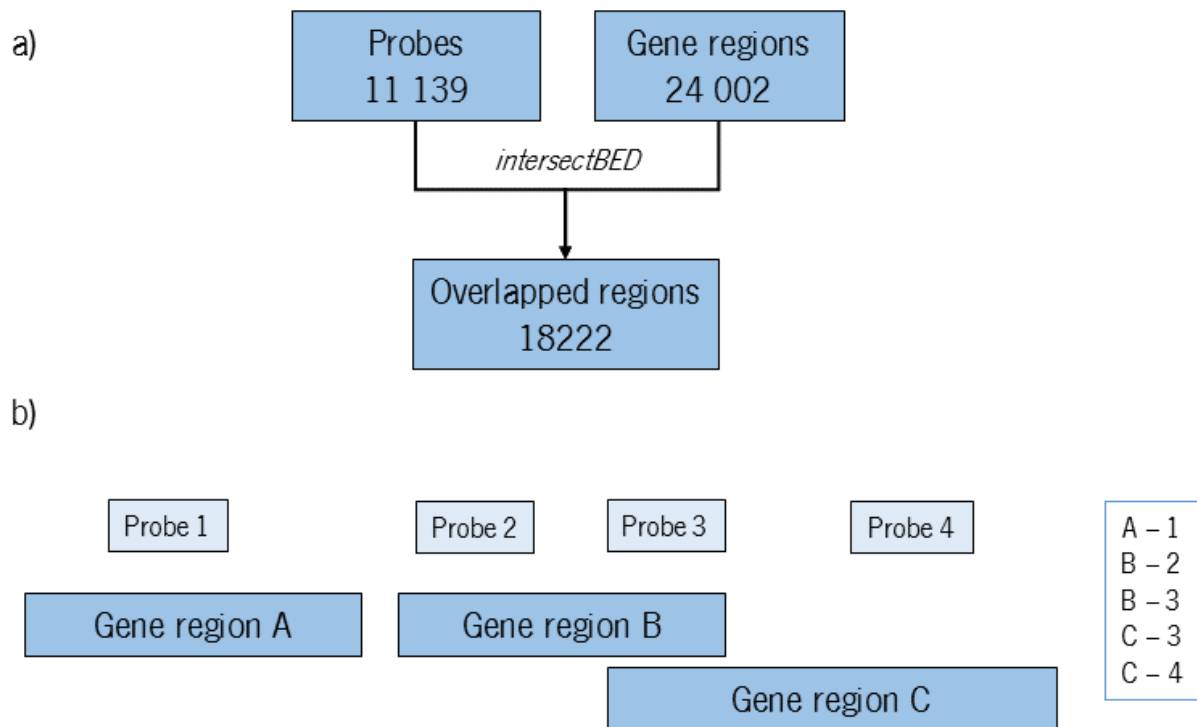


Figure 5 - Scheme of consensus annotation for DNA methylation correlation tests

a) We intersected two cohorts. TCGA with probeids (11,139) and GROUPSTUDY with gene regions (24,002). This intersection, performed with *intersectBED*, originated 18222 overlapped regions; b) in this figure it is possible to see that one probe can be in several gene regions and a gene region can have several probes.

We used the HM27 platform because we had the need to use paired samples as GROUPSTUDY did. After the filtering, we obtained a TCGA dataset with 11,139 probes. GROUPSTUDY had 24,002 transcripts. Then, creating BED files (BED 6) for both datasets and using the operation *intersectBed*, we were able to obtain the overlap between annotations with chromosome and genomic coordinates of each of these probes and transcripts (Figure 5a). We obtained 18,222 overlap regions. One probe could have several transcripts and vice-versa (Figure 5b).

After, the mean beta-value for each overlapped probe-transcript in each dataset (n=18222) was calculated (we tested with the median and the results were similar). The last step was to build the plot all

the probe-transcript between datasets in R, calculating the correlation between datasets using Pearson correlation test.

6.1.2.2 Tumour vs Normal - Differential methylation

Next, we tested if it would be the same to use paired samples or a dataset with unpaired samples, or if we would get a significant difference.

For that, we divided the datasets into three groups shown in Table 9. The groups consist of: i) 25-paired samples of gastric cancer from the TCGA cohort; ii) 248 tumour samples and 25 normal adjacent samples of gastric cancer from the TCGA cohort; iii) 50-paired samples of gastric cancer from the GROUPSTUDY cohort.

Table 9 - Samples used in methylation hypothesis tests

The three used groups were: TCGA – paired with 25 paired samples; TCGA – unpaired with 248 Tumour samples and 25 normal samples; and GROUPSTUDY with 50 paired samples.

Groups \ Samples	Tumour Samples	Normal Samples
	TCGA – paired	25
TCGA – unpaired	248	25
GROUPSTUDY	50	50

For each group, we calculated the p -value and the mean difference for each gene, to find significant genes and genes differentially methylated between Tumour and Normal, respectively. For evaluating significant genes, we used the Wilcoxon Rank-Sum Test (Tumour samples vs Normal Samples for each gene). These p -values were adjusted for multiple hypotheses testing using the Benjamini-Hochberg method. To calculate the mean difference for each gene, we calculated the absolute value of the difference between tumour and normal ($|Tumour - Normal|$). With these values calculated, we selected genes with a p -value lower than 0.05 and mean difference bigger than 0.25 for the TCGA groups. In the GROUPSTUDY, we changed the mean difference threshold to values below -1 and above 1 because of the scale of values in this group will be higher.

Finally, we obtained significant and differentially methylated genes between tumour and normal for each group.

6.1.2.2 DNA Methylation patterns

Our goal was to test the methylation status in a set of genes previously found hypermethylated in 80% of samples from the GROUPSTUDY.

Since each gene may contain multiple probes, we calculated a single gene value based on the mean of all probes beta-values. Next, we assessed methylation percentage for our set of genes in all platforms (HM450 – Tumour samples, HM27 – Tumour samples, HM27 – Normal samples). As threshold, we used a beta-value of 0.3. Genes with a beta-value < 0.3 were considered hypermethylated, while genes with beta-values > 0.3 were considered hypomethylated. After evaluating whether the gene was hypermethylated or hypomethylated for each sample, we calculated the percentage of hypermethylation for each gene per platform.

To perform the plots, we used the R package *gplots* more precisely the function *heatmap.2*. We applied the *manhattan* method to calculate the distance between genes.

6.2 Gender differential expression in gastric cancer

Our goal was to better understand gender differential gene expression and their relevance to human cancer biology. For that, we used the gene expression data (RNAseqV2) download from TCGA for the STAD. Table 10 shows the gene expression data collected.

Table 10 - Samples used for differential expression analysis

283 samples for 273 participants were downloaded from 171 males and 102 females for tumour samples (n=273). There are 10-paired samples, thus ten additional normal samples. In the paired samples, seven are males and three females.

Samples Gender	Tumour Samples	Normal Samples
Male	171	7
Female	102	3
Total	273	10

Here, our work consists of: i) processing the downloaded data; ii) identification of genes that show differential expression in cancer with respect to patient sex. For this study, two different methods were

used: non-parametric tests (*NOISeq*) and parametric tests (*DESeq2*, *tweeDEseq* and *edgeR*); and iii) Functional enrichment analysis for DEGs.

We performed four different types of analysis (Figure 6). The analyses for differential expression were: i) tumour vs normal samples, working as starting point; ii) male non-paired tumour tissue samples vs female non-paired tumour tissue samples; iii) male paired tumour tissue samples vs female paired tumour tissue samples; and iv) male paired normal tissue samples vs female paired normal tissue samples. In the three last analyses, we used gender to divide samples into two different groups.

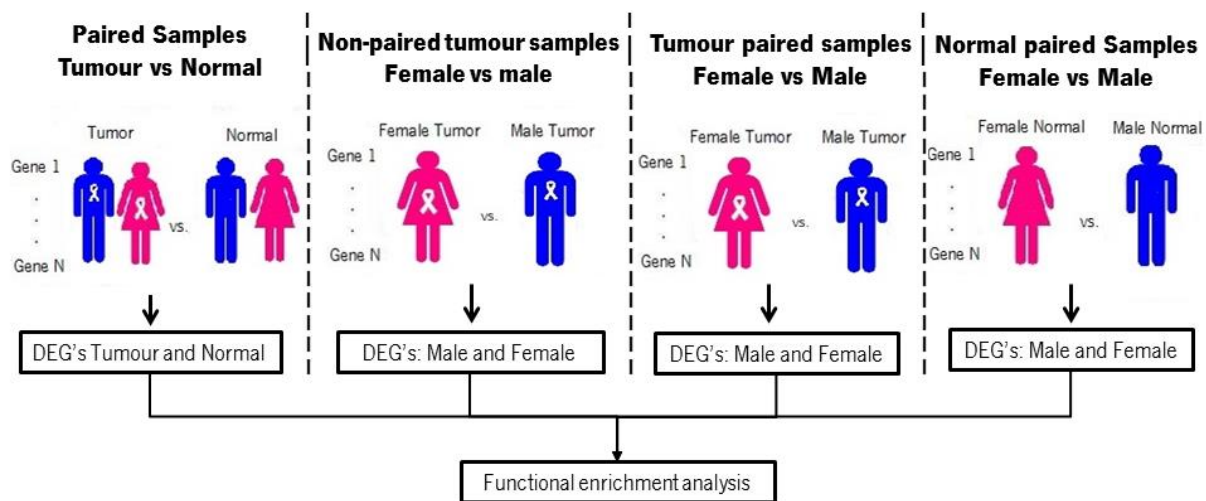


Figure 6 - First differential expression analysis workflow

This first analysis consists of four tests. One with non-paired tumour samples and three with paired samples. Two types of test were used: Male vs Female and Tumour vs Normal. For each test, the DEGs were obtained and their functional enrichment analysis.

In a second analysis (Figure 7), we identified DEGs in Tumour vs Normal samples by gender. Doing this process, we removed the normal component of samples. The input was all paired samples (n=10), divided into males (n=7) and females (n=3). The differential expression analysis was done individually (for each donor, we opposed his tumour to normal), obtaining then the DEGs between tumour and normal tissues from each donor.

In the next step, we made the intersection between all these genes for each gender, obtaining then all DEGs between tumour and normal by gender. Finally, we separated the genes that appear only in male and female, and these are the DEGs by gender.

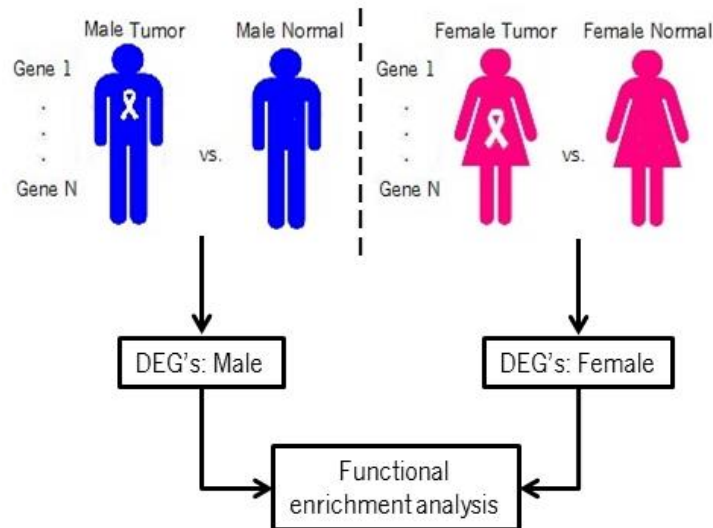


Figure 7 - Second differential expression analysis workflow

This second analysis consists of two tests. One with male samples and other with female samples, separately. The DEGs for each gender were obtained, being made next the functional enrichment analysis.

The four methods that we used required the choice of thresholds. For p -values, FDR, q -value and log2fold-change. For p -value, we used values below 0.05, considering them statistically significant. FDRs smaller than 0.05 and a q -value (only used for *NOISeq* and is equivalent to 1-FDR) of 0.95. We considered genes whose value is at least 1.5 times higher expression in male or female and tumour or normal tissues are differentially expressed, therefore we used a log2fold-change of 0.6 (a corresponding increase of 50% in the gene expression is already significant). Other tests were performed with different thresholds (for instance, log2fold-change ≥ 1 and p -value ≤ 0.1) (results not shown), but these values were too much restrictive.

We used gene expression levels calculated by TCGA consortium. This is based on UCSC annotation (Karolchik, 2003). Therefore, this was the version that we used to annotate the obtained genes. For some genes, we did not find their chromosome with any annotation that we tried to match them (GENCODE, Ensembl and RefSeq).

We did the intersection between genes obtained by the different methods in both studies. DEGs found by at least two methods were considered. As input, we have files with DEGs for each method across all types of analyses, while the output was a heat map with genes. These heat maps had the four methods in the rows and genes in the columns.

Next, we searched for enrichment of genes. A common way to do it is to integrate biological knowledge through biological ontologies as GO and KEGG.

7. RESULTS

7.1 Validation of genomic and epigenomic patterns of GROUPSTUDY in the TCGA cohort

7.1.1 TCGA Copy Number Variation

In this section, we focus on the analysis of copy number variation. Our goal was to recreate the results obtained by GROUPSTUDY and find regions of co-amplification for target gene X. We recall that we have analysed $n=52$ samples, obtained by microarrays technology. Figure 8 shows the plot with these amplifications and co-amplifications found on TCGA stomach tumour samples.

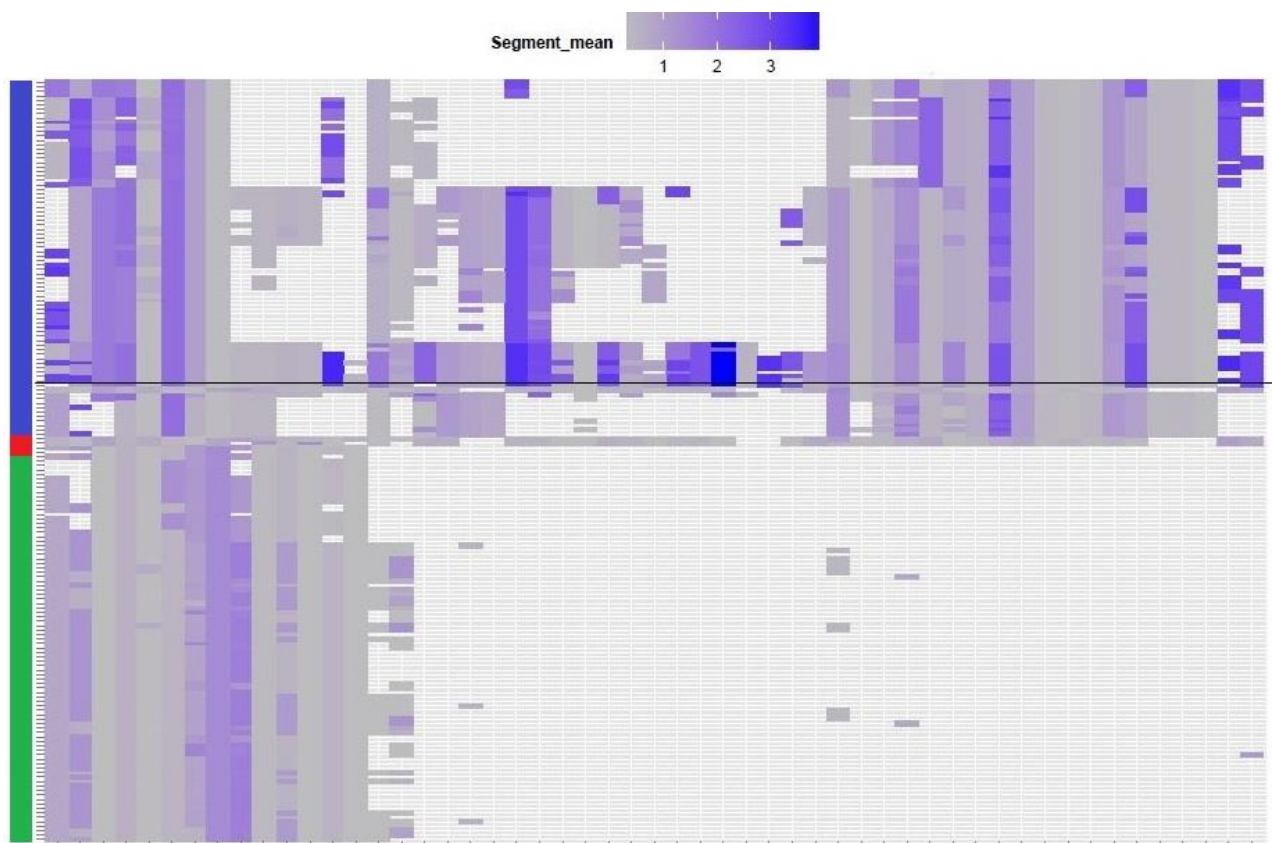


Figure 8 - Plot with CNVs from stomach cancer TCGA samples

The analysis was made for three chromosomal regions represented by colours (blue for genes on chromosomal region A, red for genes on chromosomal region B and green for genes on chromosomal region C). Each column represents the $n=52$ samples analysed, while rows represent the $n=171$ genes on the three chromosomal region. The black line on the middle of the plot represents gene X that is amplified in every sample. On top of figure, log₂-ratio represents the segment mean value and the more copies the gene has, darker is the colour.

Despite the use of different CNV technologies, we were able to recapitulate in TCGA cohort the results from the GROUPSTUDY. Large regions, not only on the same chromosome as well in the other two chromosomes, were found co-amplified.

In the chromosomal region A, genes with co-amplifications were found in almost all samples. The chromosomal region B, although with fewer genes than other chromosomal regions, it was found a homogeneous behaviour in all samples. For the third chromosomal region (C), we found a sub-set of samples in which the co-amplifications were well defined.

7.1.2 TCGA DNA Methylation

7.1.2.1 Correlation between cohorts

The goal here is to recapitulate the signal obtained in the GROUPSTUDY DNA methylation assay cohort that used a different technology from TCGA DNA methylation assay. To test this, we performed a correlation analysis between datasets, which can be visualised in Figure 9.

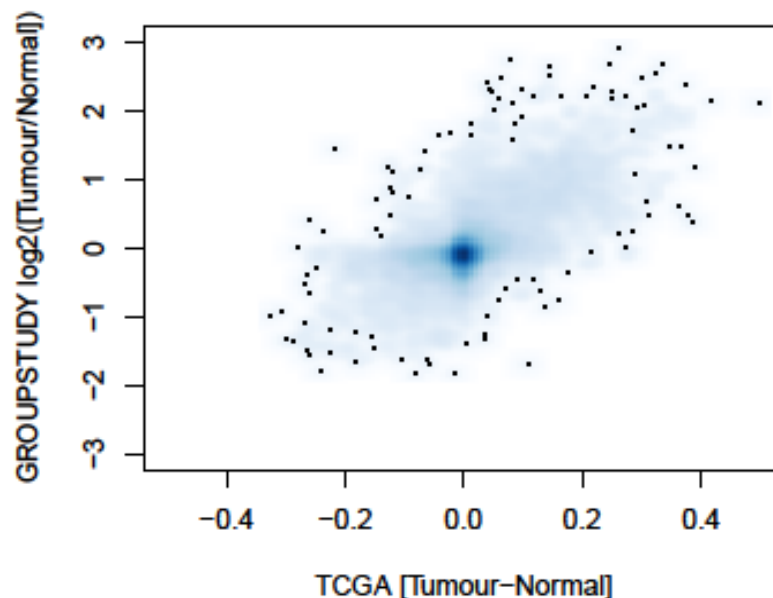


Figure 9 - Mean methylation values across samples (GROUPSTUDY vs. TCGA)

This scatterplot consists of 18222 probes corresponding to 7282 genes in each dataset. It is possible to verify that the scales were obtained in a different way (x and y-axis). This happens due to different technologies used to obtain each datasets and because of the different scale ranges, (TCGA scale is smaller and for that more sensitive to variation with low numbers). Moreover, in GROUPSTUDY we used the ratio between tumour and normal, and in TCGA we did the difference between tumour and normal (0 means that there are no differences between tumour and normal, below zero means that gene are more methylated in normal, and values above zero means that the tumour tissue is more methylated than normal tissue).

A high correlation of 0.67 was obtained between these two datasets, despite the fact that they were obtained with different technologies (Figure 10). This shows a good equivalency between datasets. We conclude that the TCGA cohort is a good proxy for the methylation status in the GROUPSTUDY.

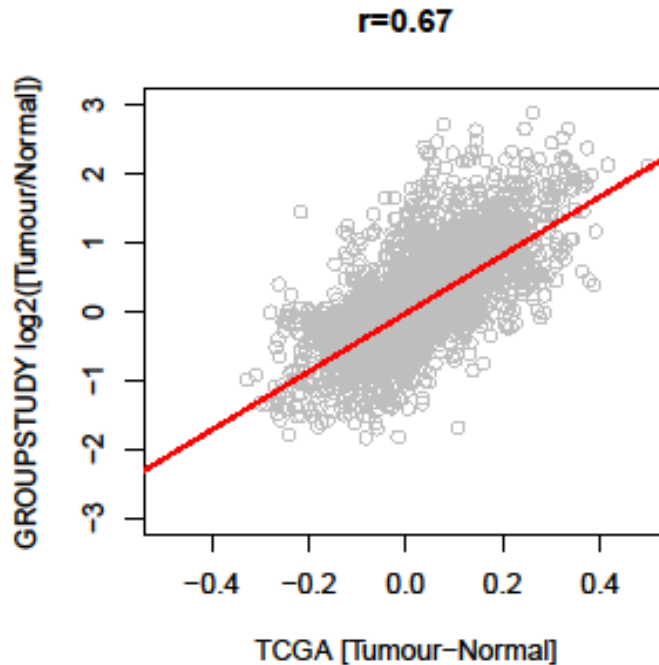


Figure 10 - Correlation between TCGA and GROUPSTUDY datasets

Datasets revealed a correlation of 0.67 what suggests a strong correlation. The red line represents the regression curve between TCGA and GROUPSTUDY. The same values from figure 9 were used. These two figures only differ in the type of plot used.

7.1.2.2 Tumour vs Normal - Differential methylation

In the second test, we investigated if using paired samples provide a better agreement between cohorts or if larger cohort size provides stronger correlation.

We recall that the datasets were divided into three different groups: TCGA - paired, TCGA - unpaired and GROUPSTUDY. For each group, genes differentially methylated and statistically significant were obtained (Figure 11).

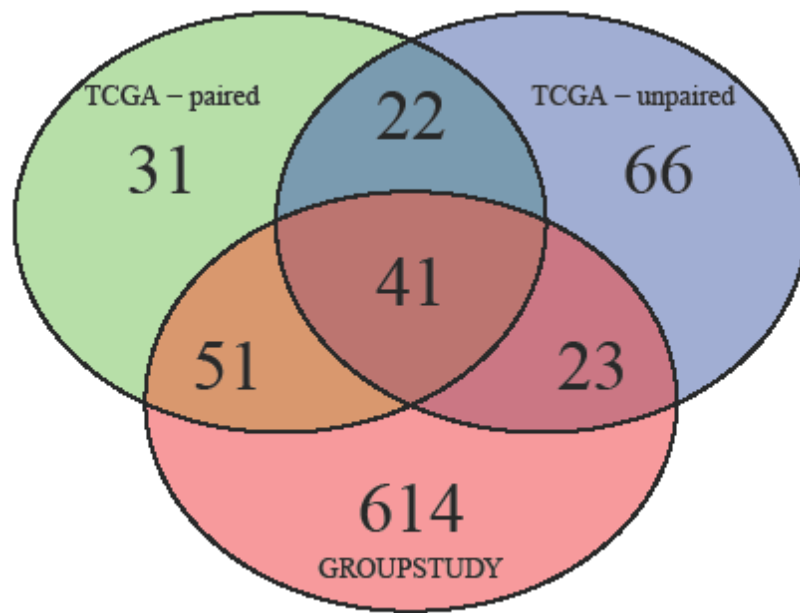


Figure 11 - Venn diagram with differentially methylated and significant genes for TCGA - paired, TCGA - unpaired and GROUPSTUDY.

Differentially methylated and significant genes (p -value<0.05 (all groups); $diffmean>0.25$ (TCGA); $diffmean>1$ (GROUPSTUDY)).

Differential methylation analysis resulted in 145, 152 and 729 differential methylated genes in TCGA - paired, TCGA - unpaired and GROUPSTUDY, respectively. The intersection between the TCGA groups was 63 genes. Of these genes, 22 are specific for TCGA groups (34.9%). TCGA - paired and GROUPSTUDY intersected 92 genes, while GROUPSTUDY intersected 64 genes with TCGA - unpaired. The group of paired samples intercepted more genes with GROUPSTUDY than the unpaired group; however, the difference is not significant. Among the three groups, 41 genes were shared.

Next, we calculated the correlation between both TCGA groups with all the probes (see Figure 10).

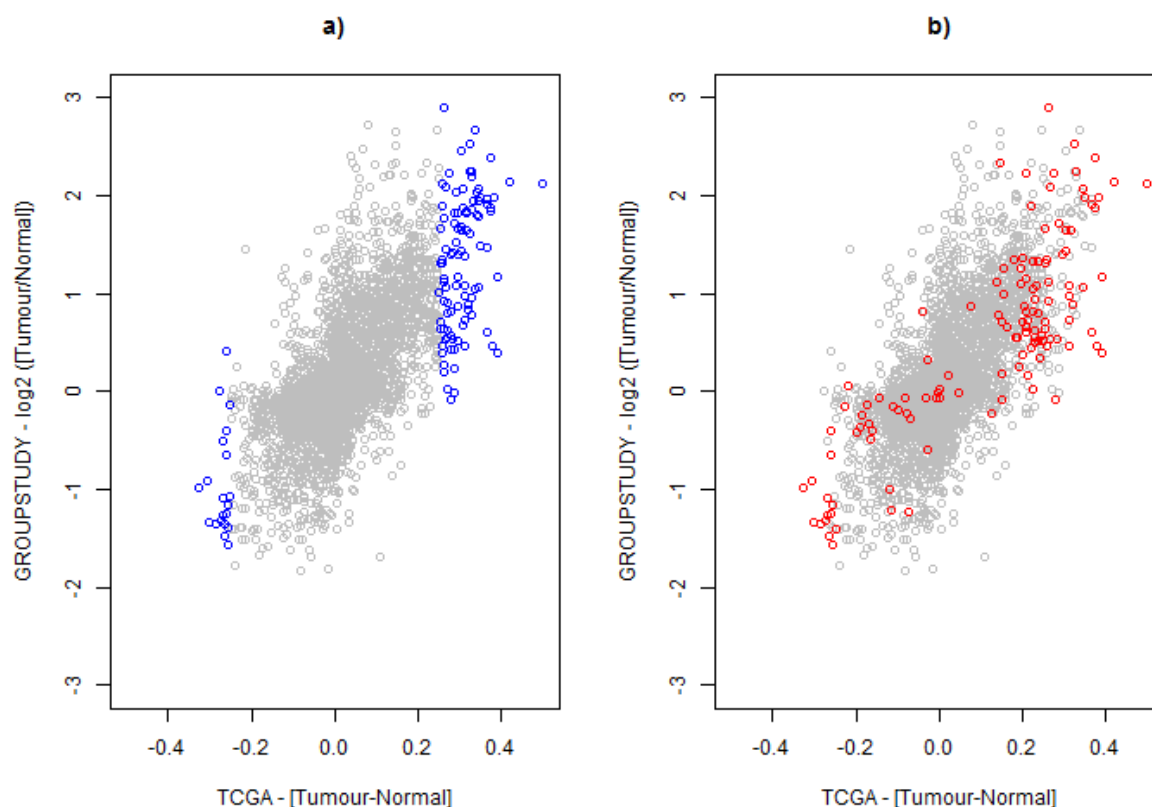


Figure 12 - Gene from the hypothesis tests in the correlation plots from TCGA and GROUPSTUDY
 This plot is the same from Figure 10 (only paired samples were used). a) Genes from TCGA - paired in blue; $r=0.81$ b) Genes from TCGA - unpaired in red; $r=0.83$. Although it appears that some of the genes were not differentially methylated, this plot was made with the paired samples. Then, we found where the differentially methylated genes from the TCGA - unpaired were positioned in the TCGA - paired vs GROUPSTUDY plot.

The results obtained suggests a strong correlation, with a correlation of 0.81 and 0.83 for TCGA - paired and unpaired, respectively (Figure 12). Thus, we can conclude that TCGA – paired and unpaired provide roughly similar results. Therefore, in future validations of GROUPSTUDY DNA Methylation results, both TCGA - paired and -unpaired samples could be used. The fact that TCGA-paired group gets similar results to a cohort with a much larger sample size is a surprising and interesting result.

7.1.2.3 DNA Methylation patterns

In this Section, DNA methylation data of stomach cancer from TCGA was analysed. The goal was to test the methylation status in our samples for both platforms for a set of genes found hypermethylated in 80% of samples in GROUPSTUDY (shown in Section 4.2.2). Several types of heat maps were built such as: heat maps for all genes (not taking into account hypermethylation or hypomethylation); heat maps for

genes hypermethylated in at least 80% of samples in tumour samples; and heat maps for genes hypomethylated in at least 80% of samples in normal samples.

Results were divided by platform, to be easier to interpret.

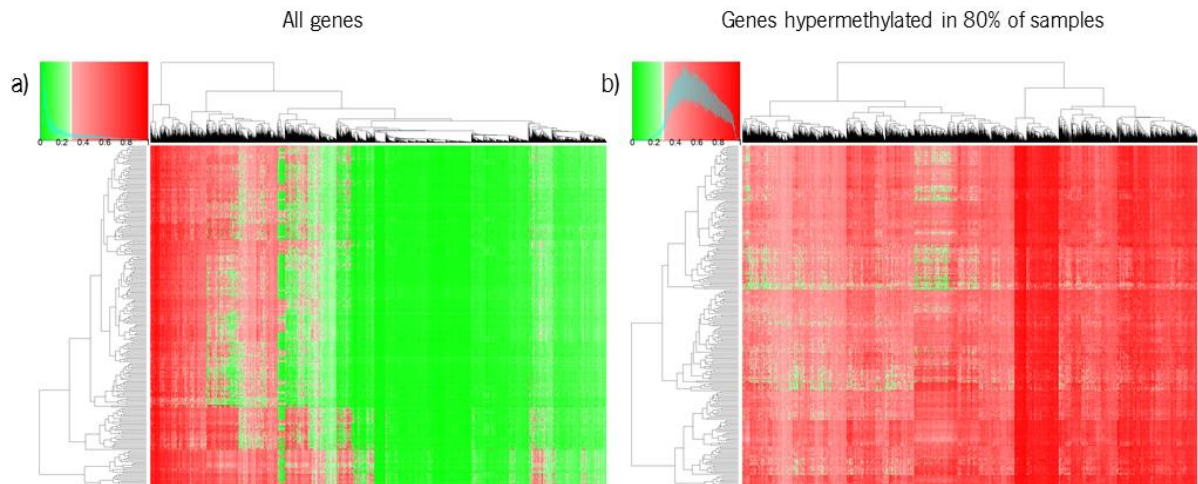


Figure 13 - Heat Maps for analysis of HM450 platform - Tumour tissue samples

a) Analysis for all genes in Tumour samples; after filtering: 14,274 genes in 248 samples; b) genes hypermethylated in 80% of the cases; after filtering: 2,876 genes in 248 samples.

Genes with a beta-value above 0.3 were considered hypermethylated while genes with beta-values below 0.3 were considered hypomethylated. Then, we calculated the percentage of hypermethylation for each gene.

In this platform, after the filtering step, 14,274 genes were analysed. In total, 2,876 genes were found hypermethylated in at least 80% of the samples in 248 tumour samples (Figure 13).

Table 11 shows that from our set of genes, eight genes (47%) are hypermethylated in more than 80% of samples. If we consider a minimum percentage of 70% of samples, we can see additional three genes, increasing to eleven genes (65%). However, we found three genes (17.6%) that have a low percentage of hypermethylation (lower than 50%). Despite that, about 83% of genes were found hypermethylated in more than 50% of samples.

Table 11 - Selected genes and respective percentage of samples with hypermethylation – HM450 platform
248 samples from HM450 platform were used.

Genes	% of hypermethylation ($\beta \geq 0.3$)
Gene A	88
Gene B	80
Gene C	81
Gene D	57
Gene E	78
Gene F	19
Gene G	35
Gene H	64
Gene I	83
Gene J	88
Gene K	95
Gene L	74
Gene M	85
Gene N	40
Gene O	79
Gene P	100
Gene Q	51

For HM27 platform, we performed more heat maps for being composed of paired samples (Figure 14).

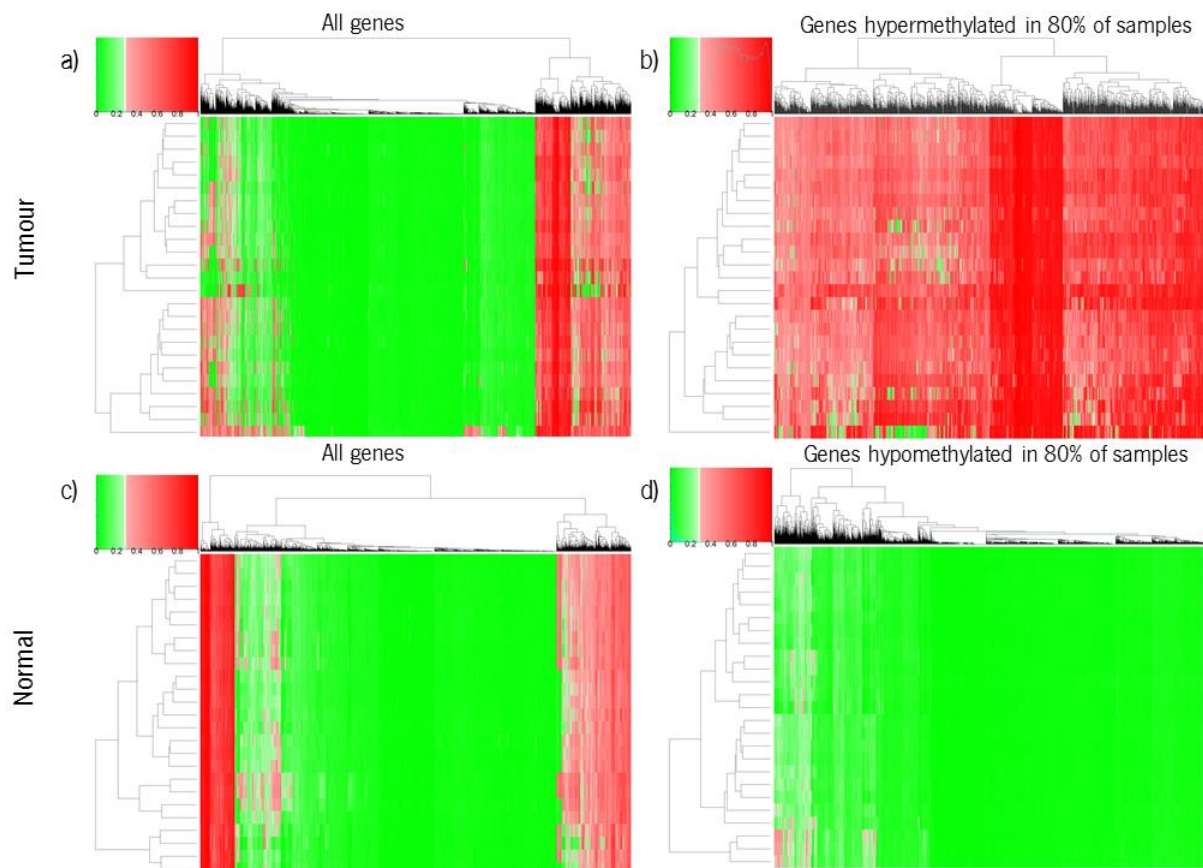


Figure 14 - Heat Maps for analysis of HM27 platform - Tumour and Normal tissue samples

a) Analysis for all genes in Tumour tissue samples; after filtering: 9,485 genes in 25 samples. b) genes hypermethylated in 80% of cases – Tumour tissue samples; after filtering: 1,463 genes in 25 samples. c) Analysis for all genes in normal tissue samples; after filtering: 9,485 genes in 25 samples. d) Genes hypomethylated in 80% of cases – normal tissue samples; after filtering: 5,998 genes in 25 samples.

After filtering, 9,485 genes were analysed for HM27 platform – Tumour and Normal. In 25 tumour samples, 1,463 genes were found hypermethylated in at least 80% of the samples. For normal tissue samples, 5,998 genes were found hypomethylated in at least 80% of samples.

Table 12 shows the hypermethylation status of tumour and normal tissues for our set of genes. In normal samples, the heat map was made for genes hypomethylated in at least 80% of samples. Thus, we considered as genes of interest those that have hypermethylation in tumour samples and a low percentage of hypermethylation in normal samples.

Five genes (31.3%) presented results of hypermethylation in at least 80% of tumour samples and were hypermethylated in less than 50% of normal samples. Six genes (37.5%) were found hypermethylated in more than 50% of normal samples.

Table 12 - Selected genes and respective percentage of samples with hypermethylation – HM27 platform
25 samples from HM27 platform were used; Gene G was not available in this platform.

Genes	% of hypermethylation Tumour tissue ($\beta \geq 0.3$)	% of hypermethylation Normal tissue ($\beta \geq 0.3$)
Gene A	100	100
Gene B	80	0
Gene C	100	32
Gene D	24	0
Gene E	84	24
Gene F	80	100
Gene H	60	0
Gene I	84	44
Gene J	92	20
Gene K	60	100
Gene L	92	60
Gene M	20	8
Gene N	92	100
Gene O	88	48
Gene P	100	100
Gene Q	12	0

TCGA samples from normal tissue are collected from adjacent tissue to the tumour. These samples may be contaminated by the tumour, possibly influencing methylation results of normal samples. For this reason, we searched for samples from stomachs of healthy donors (Lokk et al., 2014). We found 4 samples, that were obtained by one of the platforms that we analysed before (HM450). Therefore, it was possible to compare the hypermethylation percentages obtained from normal stomach samples of TCGA and stomach samples from healthy people. With this analysis, we tried to understand if the proximity of tumour and normal tissue had an impact on the methylation results.

The heat maps of Figure 15 were performed in the same way as before.

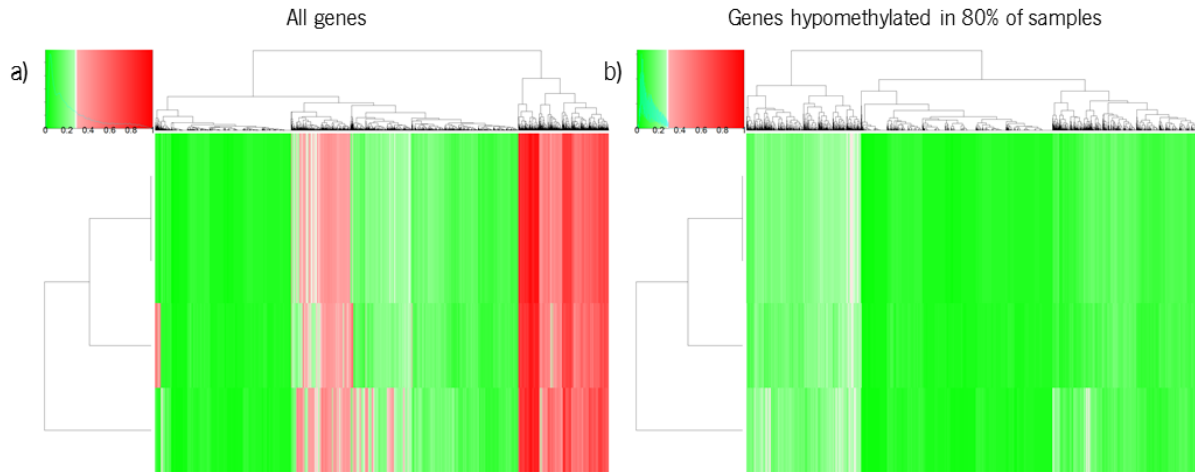


Figure 15 - Heat Maps for analysis of HM450 platform - Healthy tissue samples

a) Analysis for all genes in healthy tissue samples; after filtering: 14,274 genes in four samples. b) Genes hypomethylated in 80% of cases – healthy tissue samples; after filtering: 8,933 genes in four samples.

Of the 14,274 genes obtained after filtering, 8,933 genes were found hypermethylated in at least 80% of healthy tissue samples.

Table 13 shows the percentage of samples with hypermethylation in the healthy tissue samples. The percentages of samples with hypermethylation were high in two genes (Gene K and P). Despite the fact of having only four samples, in most genes, the percentage was zero (70.1%). If we take into account the samples hypermethylated in only 25% of cases, we increase the number of genes with a low percentage of hypermethylation (88%) to 15.

Table 13 - Selected genes and respective percentage of samples with hypermethylation – Healthy samples
Four samples from HM450 platform were used.

Genes	% of hypermethylation Healthy tissue ($\beta \geq 0.3$)
Gene A	25
Gene B	0
Gene C	0
Gene D	0
Gene E	0
Gene F	25
Gene G	0
Gene H	0

Gene I	0
Gene J	0
Gene K	100
Gene L	0
Gene M	0
Gene N	0
Gene O	25
Gene P	100
Gene Q	0

This result seems to indicate that normal samples from areas adjacent to the tumour may have contributed to the high hypermethylation values in some of our selected genes since in samples from healthy individuals most of the genes were not found hypermethylated.

7.2 Gender differential expression in gastric cancer

We set to investigate patterns of gender differential expression in both tumour and normal tissue samples. We used n=283 samples obtained from gastric cancer study by TCGA to test gene differential expression, with the goal of identifying transcriptomic differences between genders.

First, we divided the study into two different analyses. The first analysis, consisting of four different tests: i) tumour vs normal samples, worked as a starting point. ii) male non-paired tumour tissue samples vs female non-paired tumour tissue samples; iii) male paired tumour tissue samples vs female paired tumour tissue samples; and iv) male paired normal tissue samples vs female paired normal tissue samples. After the application of four different methods and evaluating their consistency (*NOISeq*, *tweeDEseq*, *DESeq2* and *edgeR*), we obtained results described in the following eight tables (Table 14-21). These tables represent the DEGs for each method (columns), the respective chromosome (X|Y or autosomal) and intersection results between methods and genes upregulated in male and female or tumour and normal (rows).

Table 14 - Differential expression analysis between paired tumour and normal tissue samples
 10 paired samples were used; DESeq2 and EdgeR obtained best and similar results; NOISeqBIO obtained the lowest number of genes;
 Intersection (2+) - Intersection in at least 2 methods.

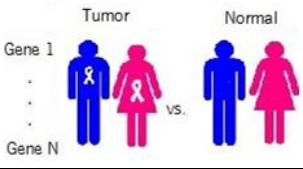
Paired Samples Tumour vs Normal 	<i>NOISeqBIO</i> (~ gender)	<i>tweeDEseq</i> (~ gender)	<i>DESeq2</i> (~ race + gender)	<i>EdgeR</i> (~ gender)
Total gene	89	954	2,178	1,525
X Y genes	2 0	20 1	73 3	47 2
Autosomal genes	81	848	1,971	1,346
Intersection (2+)	1,410			
Genes upregulated in tumour	87	505	1,213	908
Genes upregulated in normal	2	449	965	617
Unknown genes	6	85	131	130

Table 15 - Intersection of autosomal DEGs between tissue types across methods – paired samples
 For this test, all methods obtained a large number of DEGs (except the NOISeqBIO). EdgeR and DESeq2 intersected for almost every DEGs found with edgeR (89.6%). TweeDEseq and DESeq2 intersected 90.4 % of DEGs found with tweedeDEseq. NOISeqBIO and edgeR intersected 76 of 81 DEGs found with NOISeqBIO (93.8%)

Methods	<i>NOISeqBIO</i> (n=81)	<i>tweeDEseq</i> (n=848)	<i>DESeq2</i> (n=1,971)	<i>edgeR</i> (n=1,347)
<i>NOISeqBIO</i> (n=81)				
<i>tweeDEseq</i> (n=848)	39			
<i>DESeq2</i> (n=1,971)	71	767		
<i>edgeR</i> (n=1,347)	76	585	1,207	

Table 16 - Differential expression analysis between genders in non-paired tumour tissue samples
 505 non-paired tumour samples were used; the proportions between chromosomes of genes are quite similar. EdgeR stands out for number of DEGs, X and Y genes detected. Intersection (2+) - Intersection in at least 2 methods.

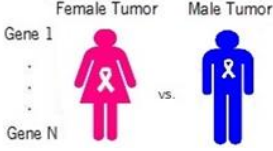
Non-paired tumour samples Female vs male 	<i>NOISeqBIO</i> (~ gender)	<i>tweeDEseq</i> (~ gender)	<i>DESeq2</i> (~ race + gender)	<i>EdgeR</i> (~ gender)
Total gene	20	23	65	453
X Y genes	2 12	5 12	4 12	10 14
Autosomal genes	4	3	45	401
Intersection (2+)	45			
Genes upregulated in male	2	19	52	301
Genes upregulated in female	18	4	13	152
Unknown genes	2	3	4	28

Table 17 - Intersection of autosomal DEGs between genders across all methods – non-paired tumour tissue samples
DESeq2 and *edgeR* intersected for almost every of *DESeq2* genes (91.1%) while *tweeDEseq* and *NOISeqBIO* intersected every of their genes with *edgeR*

Methods	<i>NOISeqBIO</i> (n=4)	<i>tweeDEseq</i> (n=3)	<i>DESeq2</i> (n=45)	<i>edgeR</i> (n=401)
<i>NOISeqBIO</i> (n=4)				
<i>tweeDEseq</i> (n=3)	0			
<i>DESeq2</i> (n=45)	1	2		
<i>edgeR</i> (n=401)	4	3	41	

Table 18 - Differential expression analysis between genders in paired tumour tissue samples

Ten tumour samples were used; EdgeR and DESeq2 were the more sensible methods for detecting X and Y genes; tweedEseq reported the biggest number of DEGs; Intersection (2+) - Intersection in at least 2 methods.

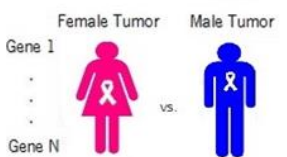
Tumour paired samples Female vs Male 	<i>NOISeqBIO</i> (~ gender)	<i>tweedEseq</i> (~ gender)	<i>DESeq2</i> (~ race + gender)	<i>EdgeR</i> (~ gender)
Total gene	10	161	45	80
X Y genes	1 8	5 8	4 10	3 11
Autosomal genes	0	129	29	61
Intersection (2+)	25			
Genes upregulated in male	1	112	38	38
Genes upregulated in female	9	49	7	42
Unknown genes	1	19	2	5

Table 19 - Intersection of autosomal DEGs between genders across all methods – tumour paired samples

DESeq2 intersected 51% of their DEGs with the DEGs found with *tweedEseq* while *edgeR* only intersected 11% of their genes with *tweedEseq* DEGs. *EdgeR* and *DESeq2* intersected 55% of DEGs found with *DESeq2*.

Methods	<i>NOISeqBIO</i> (n=0)	<i>tweedEseq</i> (n=129)	<i>DESeq2</i> (n=29)	<i>edgeR</i> (n=61)
<i>NOISeqBIO</i> (n=0)				
<i>tweedEseq</i> (n=129)	0			
<i>DESeq2</i> (n=29)	0	15		
<i>edgeR</i> (n=61)	0	6	16	

Table 20 - Differential expression analysis between genders in paired normal tissue samples

Ten normal samples were used; Here *NOISeqBIO* stands out. Intersection (2+) - Intersection in at least 2 methods


Normal paired Samples Female vs Male 	<i>NOISeqBIO</i> (~ gender)	<i>tweeDEseq</i> (~ gender)	<i>DESeq2</i> (~ race + gender)	<i>EdgeR</i> (~ gender)
Total gene	1,636	67	19	17
X Y genes	62 14	5 11	2 12	1 12
Autosomal genes	1,364	44	2	2
Intersection (2+)	7			
Genes upregulated in male	225	43	16	14
Genes upregulated in female	1,411	24	3	3
Unknown genes	196	7	3	2

Table 21 - Intersection of autosomal DEGs between genders across all methods – normal paired samples

For this test, only *NOISeqBIO* and *tweeDEseq* intersected their DEGs. *TweeDEseq* intersected 15.9% of their DEGs with the DEGs found with *NOISeqBIO*

Methods	<i>NOISeqBIO</i> (n=1,364)	<i>tweeDEseq</i> (n=44)	<i>DESeq2</i> (n=2)	<i>edgeR</i> (n=2)
<i>NOISeqBIO</i> (n=1,364)				
<i>tweeDEseq</i> (n=44)	7			
<i>DESeq2</i> (n=2)	0	0		
<i>edgeR</i> (n=2)	0	0	0	

Table 22 - Intersection between paired tumour, normal and GTEx (male vs female)

Intersection 1 – Intersection between autosomal DEGs

Genes \ Methods	<i>DESeq2</i> (~ race + age + gender)		<i>edgeR</i> (~ gender)		
	Tumour: Male vs Female	Normal: Male vs Female	Tumour: Male vs Female	Normal: Male vs Female M vs F	GTEx Normal: Male vs Female
Total genes	45	19	80	17	48
X Y genes	4 10	2 12	3 11	1 12	11 14
Autosomal genes	29	2	61	2	23
Intersection 1	0		0		
Genes upregulated in male	38	16	38	14	33
Genes upregulated in female	7	3	42	3	15
Unknown genes	2	3	5	2	0

The intersection between autosomal genes from the paired tumour tissue samples (male vs female) and paired normal tissue samples (male vs female) was done to evaluate if the genes found in the tumour could be specific for this tissue or if they are also in normal tissue. There were no genes found in tumour and normal tissue with *edgeR* (61 from tumour tissue vs 2 from normal tissue) or with *DESeq2* (29 from tumour tissue vs 2 from normal tissue) (Table 22). This suggests that the gender differential expression observed is tumour specific.

No genes were found in the intersection between the differential expression genes in normal tissue (with *edgeR*, n=2 vs n=23 in GTEx).

To find genes that were differentially expressed in at least two methods, we did the intersection between the genes and the methods for each type of analysis. With this information, we created heat maps for each analysis with the gene list (Figure 16).

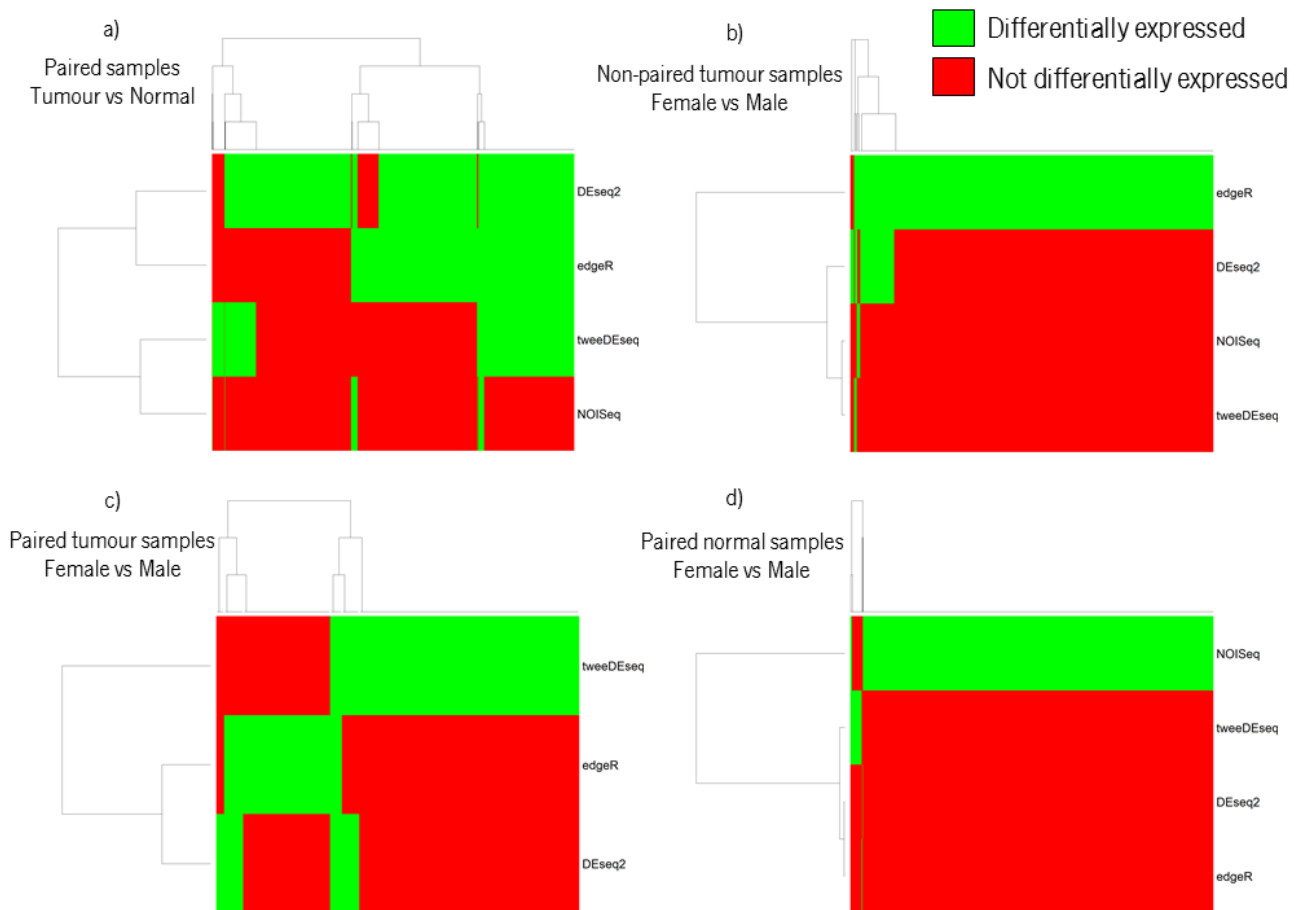


Figure 16 - Heat Maps with autosomal DEGs for each method for analysis 1

Each row contains the methods used in the analyses while columns represents all the genes found for each analysis.

a) Paired samples – Tumour tissue Vs Normal tissue - were found 1,410 genes in at least two methods for a total of 2,186 genes; b) Tumour tissue samples – Male x Female - were found 45 genes in at least two methods for a total of 405 genes; c) Paired samples - Tumour: Male Vs Female - were found 25 genes in at least two methods for a total of 188 genes; were found 45 genes in at least two methods for a total of 1,406 genes.

On the second analysis in which we test DEGs between tumour and normal samples by gender, we obtained DEGs for male and female for four different methods (Table 23). Considering genes in at least two methods, we found 290 autosomal DEGs for male and 283 autosomal DEGs for female. These genes may or may not be specific to each gender. Thus, we evaluated which would be specific for male and female (Table 24), finding 152 autosomal DEGs specific for male and 205 autosomal DEGs specific for female.

Table 23 - Autosomal DEGs by gender across methods

Rows represent the autosomal genes intersections across methods and the genes in at least two methods by gender. Gene 2+ - genes in at least two methods.

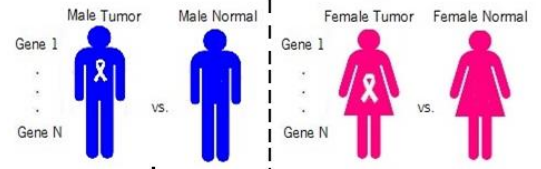
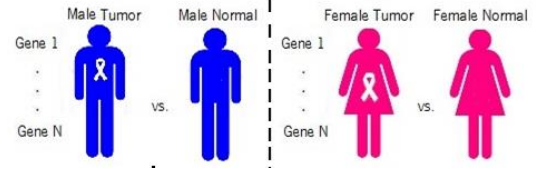
	NOISeq	tweeDEseq	DESeq2	EdgeR	Genes 2+
Autosomal genes - Male	370	1,692	62	119	290
Autosomal genes - Female	396	2,030	16	113	383

Table 24 - Specific autosomal DEGs by gender across methods

Genes from table 22 could be found differentially expressed on both genders. For this, we evaluated which DEGs were specific for each gender. Gene 2+ - genes in at least two methods.

	NOISeq	tweeDEseq	DESeq2	EdgeR	Genes 2+
Specific Male Genes	243	1,136	58	99	152
Specific Female Genes	269	1,474	12	93	205

To find genes that were differentially expressed in at least two methods in the second analysis, we did the intersection between the genes and the methods for each type of analyses (Figure 17).

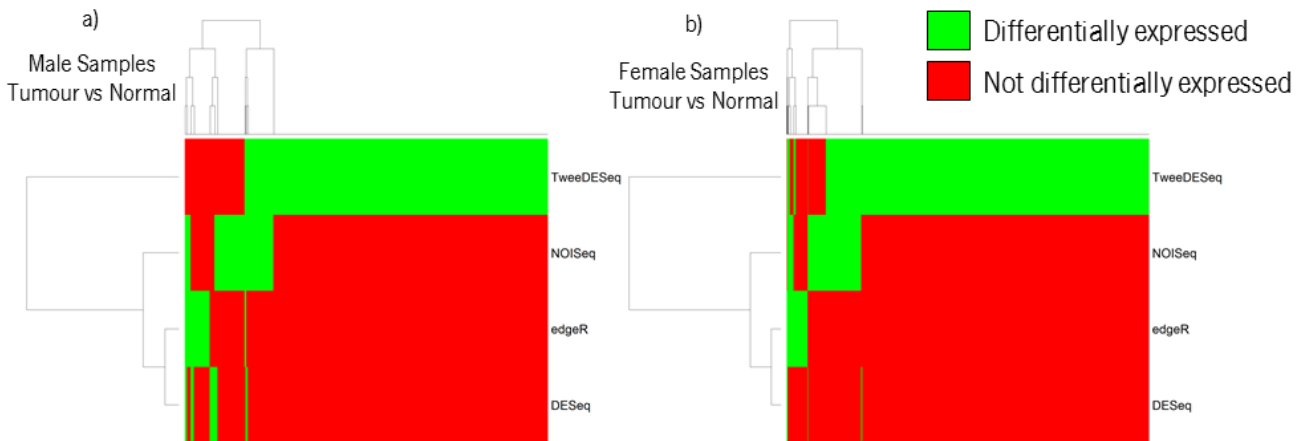


Figure 17 - Heat Maps with autosomal DEGS for each method in analysis 2

Each row contains the methods used in the analyses while columns represents all the genes found for each analysis.

a) Paired samples - Tumour vs Normal: Male - were found 152 genes in at least two methods for 1,362 genes; b) Paired samples - Tumour vs Normal: Female - were found 206 genes in at least two methods for 1,625 genes.

Next, we did the intersection of genes between analysis 1 and analysis 2 are shown in table 25.

Table 25 - Intersections between analysis 1 and analysis 2
Tumour - M vs F with Male/Female - Tumour vs Normal

Methods	<i>NOISeq</i> (n=4)	<i>tweeDEseq</i> (n=3)	<i>DESeq2</i> (n=45)	<i>edgeR</i> (n=401)
Genes in at least two methods (n=357)	0	0	2	11

Due to the large number of reported differentially expressed genes and the disagreement between methods, and to further understand the results in more detail, we chose to focus on the results provided by one of the methods. This allowed further analysing the relevance of this results. This next step included an enrichment analysis. The selected method was *edgeR* for the following reasons: i) this method was one of the quickest; ii) in general it was more sensitive, reporting more X and Y genes, which is a good indicator and obtained the best results in the intersection between studies; iii) Manual inspection of the expression values of DEGs showed a more clear difference in the fold-change of expression for genes reported by this method (see Figure 18).

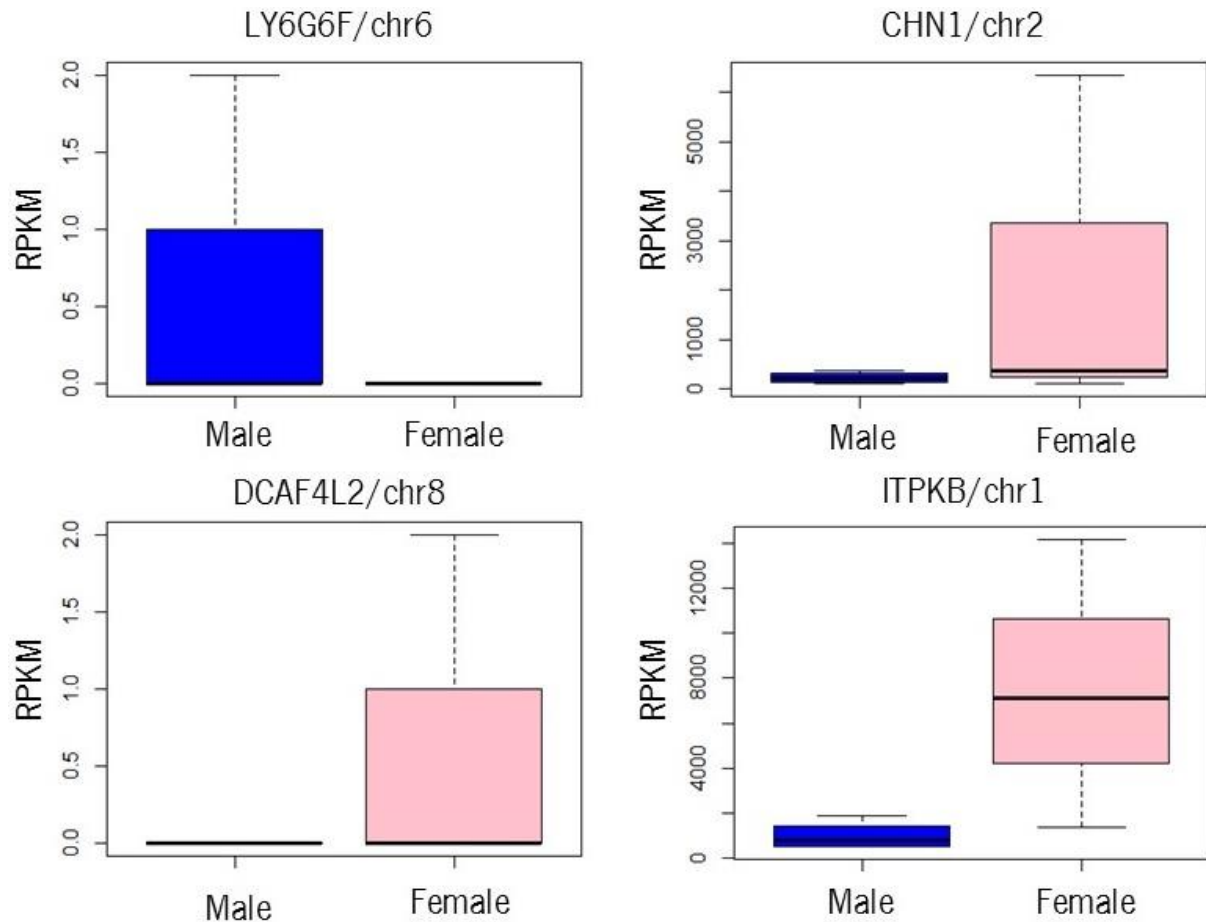


Figure 18 - Box plot with DEGs found in edgeR

These four DEGs found by edgeR and are examples of genes with differential expression between male and female and the other three methods could not catch them.

7.2.1 Functional enrichment analysis

In this section, we made the functional enrichment analysis of DEGs found in *edgeR*. We used the R package *clusterProfiler* to perform this analysis. Using the function *compareCluster*, we searched the functional enrichment between tumour and normal and between male and female. We divided each analysis into two groups (Tumour vs Normal and Male vs Female).

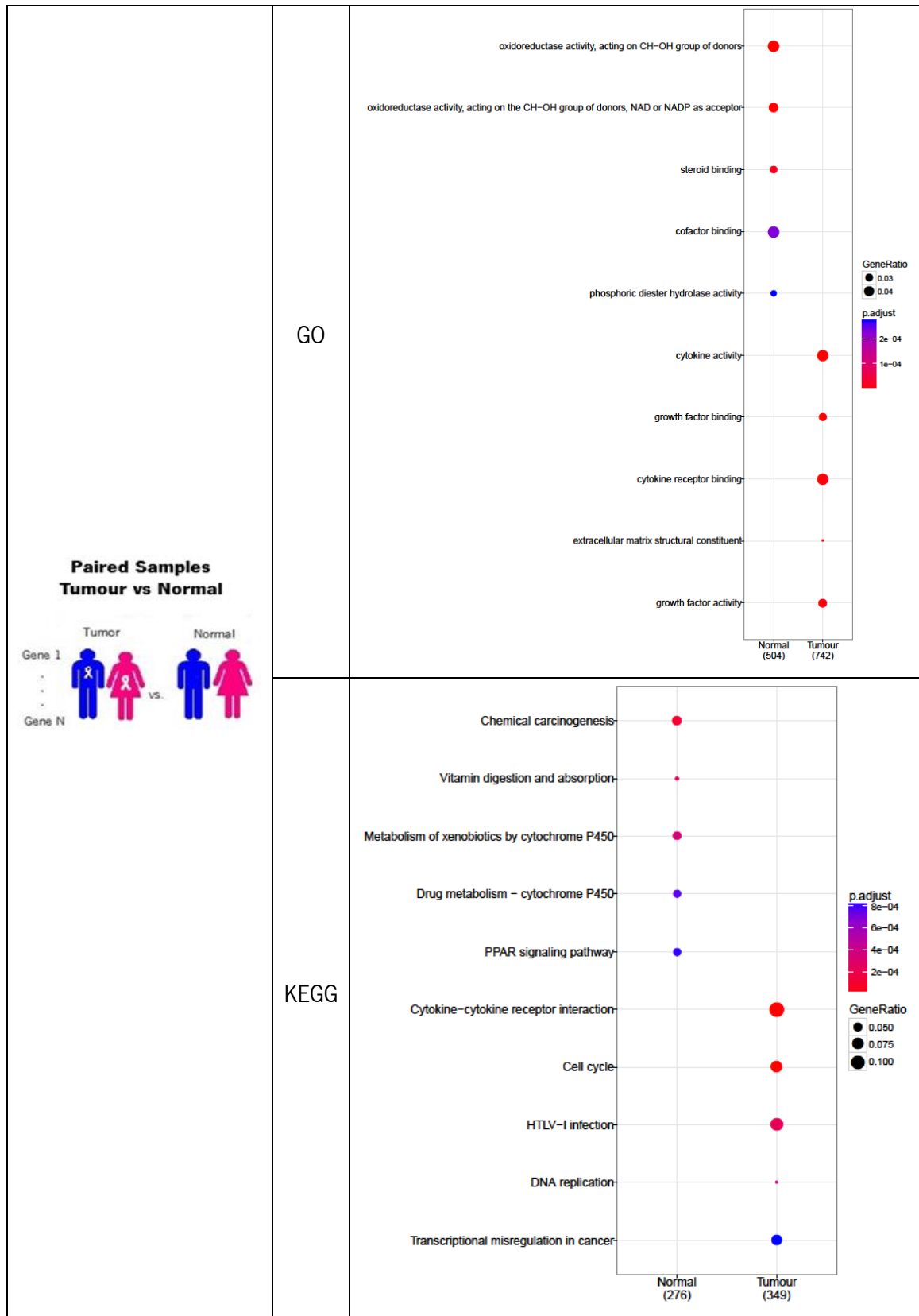


Figure 19 - GO and KEGG analysis for DEGs from tumour vs normal tissue samples

On a total for 536 genes upregulated in normal and 810 upregulated in tumour, 504 and 742 could be correctly annotated, respectively. Gene Ontology analysis revealed five enriched activities for genes upregulated in normal and five for tumour. For KEGG analysis, less genes were annotated for a total of 276 for normal and 349 for tumour.

Figure 19 shows the enrichment for ten molecular functions (five in normal and five in the tumour). We highlight the cytokine activity in tumour (n=34, p=2.7e-08) and oxidoreductase activity in normal (n=24, p=7.03e-13) as the more enriched. The cytokine-cytokine receptor interaction pathway was enriched in 35 genes (p=1.58e-04) and the most significant pathways enriched was chemical carcinogenesis in 18 genes (p=2.4e-04). Cytokines are a category of small proteins that are important in cell signalling. They are important in health and disease, specifically in host responses such as infection, immune responses and cancer. Cytokines are released by several cells in the body, commonly in response to an activating stimulus, and inducing responses through binding to specific receptors on the cell surface of target cells (Schreiber and Walter, 2011).

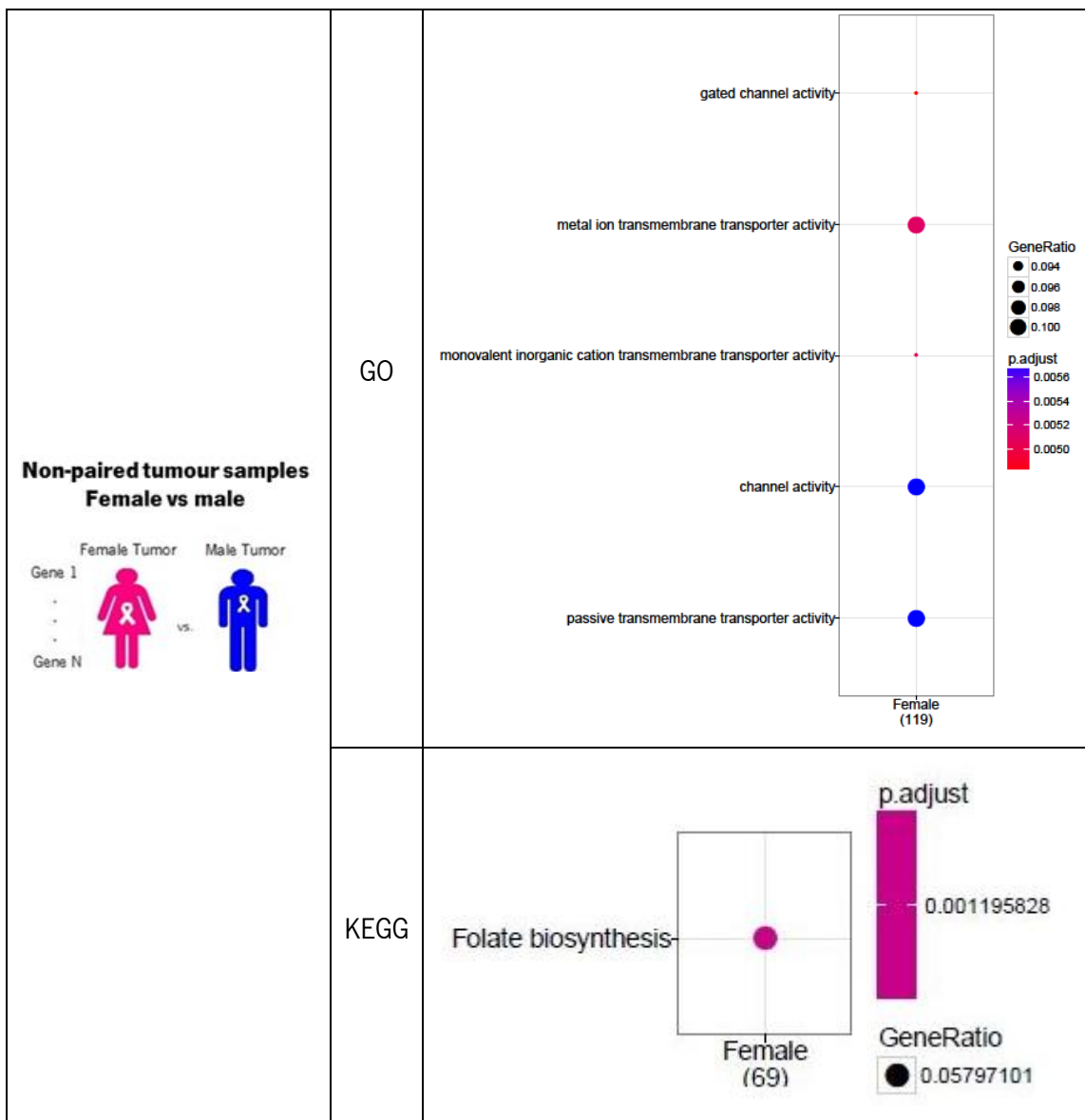


Figure 20 - GO and KEGG analysis for DEGs from male vs female – tumour tissue samples

For a total for 152 genes of female, 119 could be correctly annotated. Gene Ontology analysis revealed five enriched activities for female genes. For KEGG analysis, less genes were annotated for a total of 69 genes

The GO enrichment analysis revealed five molecular functions enriched in female genes (Figure 20). The more significant were metal ion transmembrane transporter activity (n=12 genes, p= 3.22e-05) and gated channel activity (n=11 genes, p=1.43e-05). Folate biosynthesis pathway (n=4 genes, p=0.001) was found enriched in female genes. This pathway (Figure 21) is involved in the metabolism of vitamins. Folate is one of the B vitamins and it is used as a supplement during pregnancy to prevent neural tube defects (NTDs). It can be also used to treat anaemia caused by folic acid deficiency. Folate is essential for the body to make DNA, RNA, and metabolise essential amino acids for cell division.

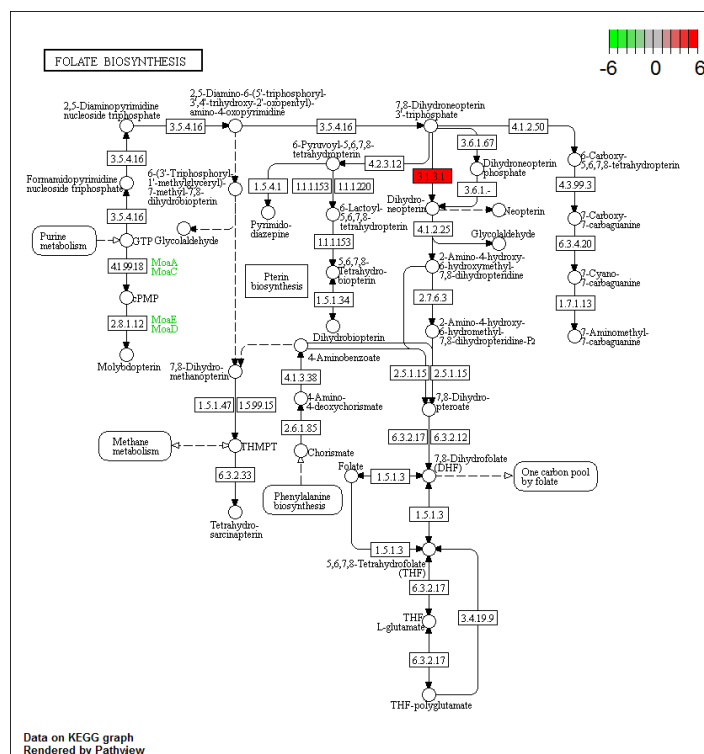


Figure 21 - Folate biosynthesis pathway

Scale values are represented by log₂-fold-change, positive for female (red) and negative in male (green). Four genes showed a positive fold change for alkaline phosphatase – ALPI, ALPL, ALPP and ALPLL2. These genes are involved in the metabolism of cofactors and vitamins.

The two analyses between male and female with paired samples (tumour and normal) were also tested, showing no enrichment in any of genes cluster. However, we were able to analyse the *groupGO* (Appendix I – Figures 25-30). Thus, we found that in the tumour, most genes have functions related to cellular compartments like membrane part (n=30) and plasma membrane (n=26). Twelve genes intervene in transferase activity and 25 in the regulation of metabolic processes. The normal tissue analyses shown function related with membrane and cell part (n=2), biological processes like single-organism cellular process (n=2) and intervention in protein binding (n=2).

Next, we made the enrichment analyses for analysis 2.

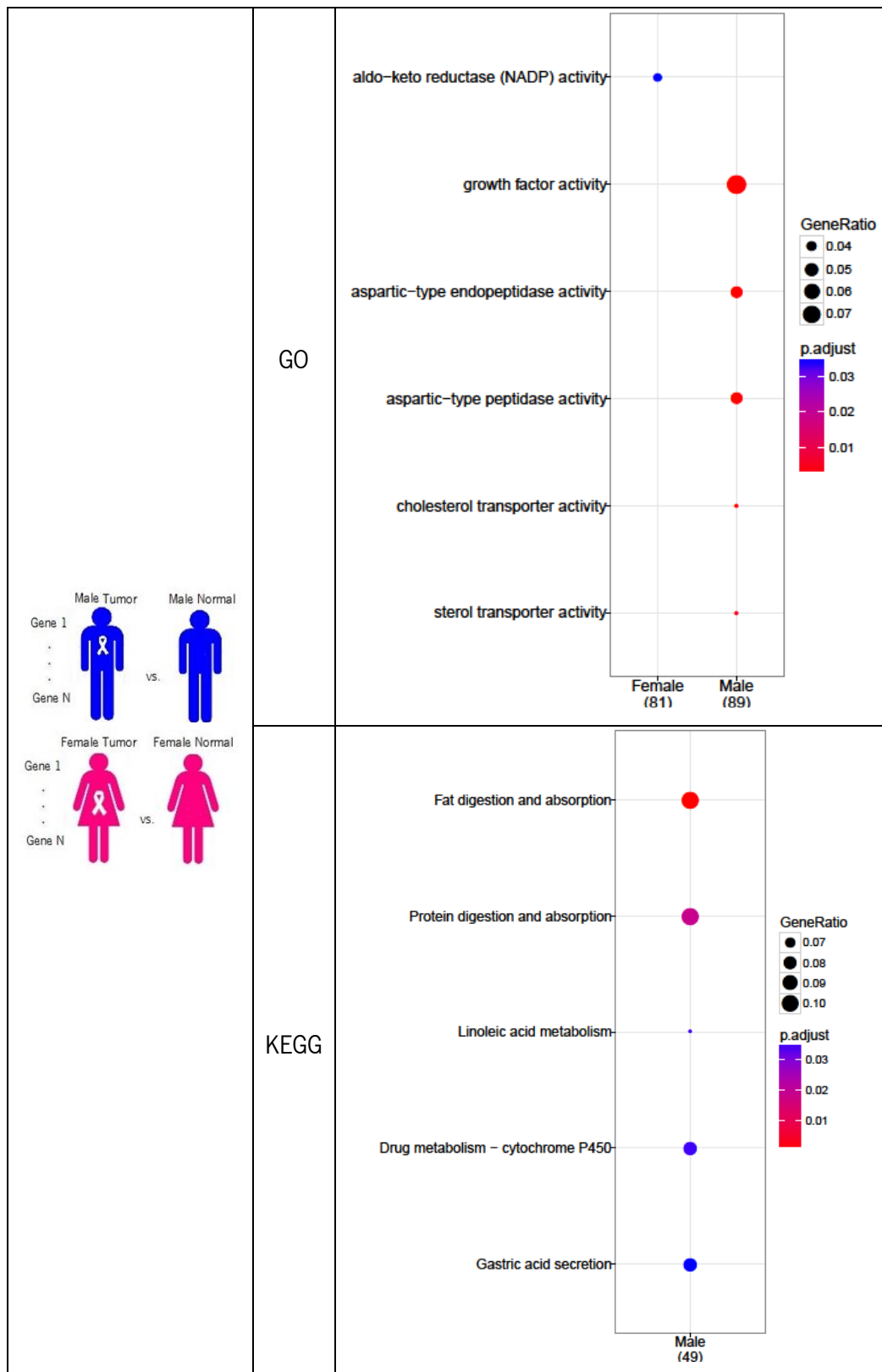


Figure 22 - GO and KEGG analysis for DEGs from tumour vs normal tissue samples by gender
 On a total for 93 genes of female and 99 of male, 81 and 89 could be correctly annotated, respectively. Gene Ontology analysis revealed one enriched activity for genes upregulated in female and five for male. For KEGG analysis, less genes were annotated for 49 for male and no enrichment found in female.

Gastric acid is a digestive fluid formed in the stomach. The acid plays a key role in the digestion of proteins by activating digestive enzymes and facilitates the digestion of protein and the absorption of iron, calcium, vitamin B12. Gastric acid, by lowering pH, kills ingested microorganisms and limits bacterial growth in the stomach and prevents intestinal infections such as *Clostridium difficile* (Schubert and Peura, 2008).

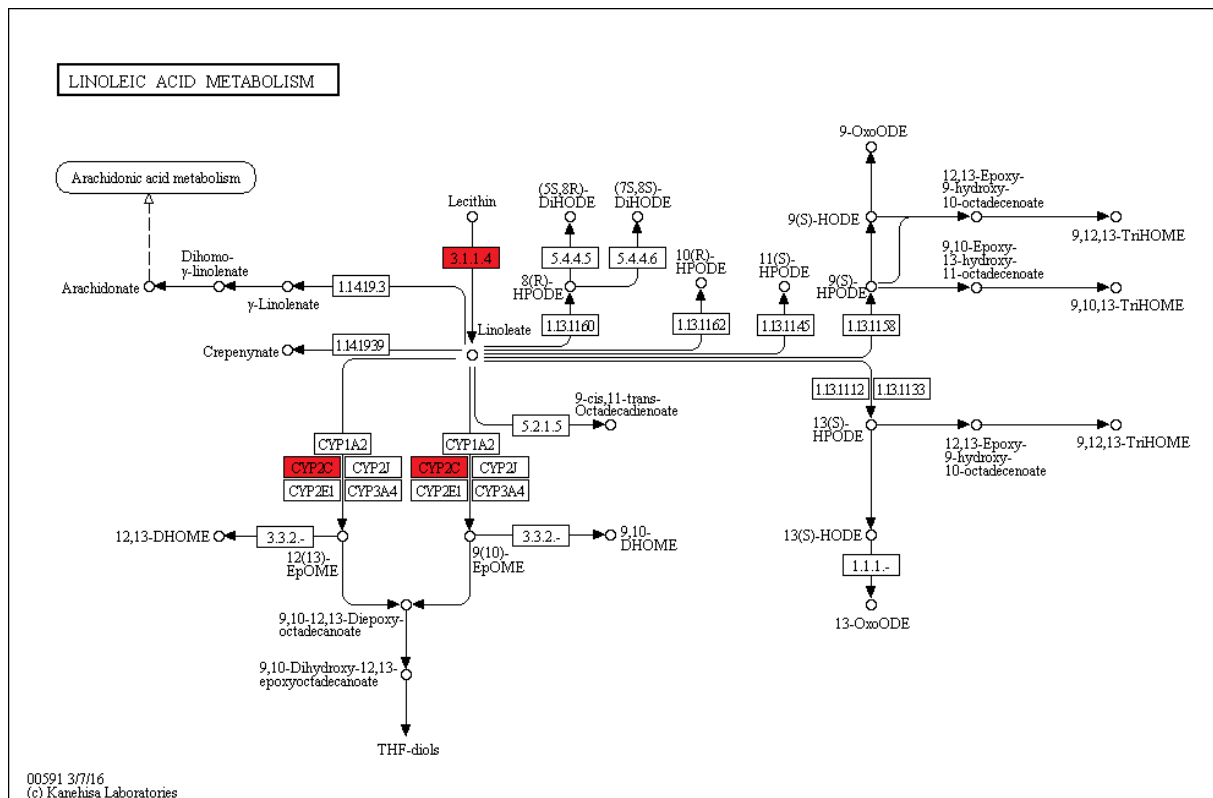


Figure 24 - Linoleic acid metabolism pathway
Three genes show enrichment in this pathway: *CYP2C19*, *CYP2C8* and *PLA2G3*.

Linoleic acid is a polyunsaturated omega-6 fatty acid. There is also proof of Conjugated linoleic acid (CLA) blocking the growth and spread of malignant tumours, primarily by influencing cell replication and mechanisms of carcinogenesis. On the other hand, it has been shown that CLA can induce insulin resistance and fatty liver (Field and Schley, 2004). Lee et al. (LEE et al., 2005) found that relatively small amounts of CLA inhibit tumour development. However, there is the need to perform studies to analyse the effects of CLA in clinical trials to humans.

8. DISCUSSION

In this dissertation, we followed two lines of analysis: i) use the TCGA cohort to recover and validate patterns from an internal study. For this first goal, we used genomic data (CNVs) and epigenomic data (DNA Methylation) from the TCGA gastric cancer cohort; ii) gender differential expression in gastric cancer. Here, we used transcriptomic data (mRNA expression) from the TCGA cohort.

With the genomic data in analysis i) our goal was to validate results found in GROUPSTUDY, which showed that when a particular gene was amplified there is co-amplification of regions on the same chromosome and other chromosomes. In human cancer, gene amplification offers a means of overexpression of oncogenes. These amplifications can work as a mechanism of resistance to therapies and are frequently related to poor prognosis. With the results obtained for this analysis, we found exactly the same patterns of co-amplification that in GROUPSTUDY. This provides a good evidence of validation of the initial results. It can be important to identify the genes or pathways that promote these amplifications since they can be targeted by a combination of therapy to avoid evolution of resistance to drugs planned to destroy the tumour.

Regarding epigenomic data, our main goal was to test data correlation between the TCGA and GROUPSTUDY datasets, since these two cohorts used different technologies to assess methylation status. We found a correlation of 0.6, which indicates a good equivalency between cohorts. We conclude that the data from the TCGA cohort is a good proxy for the methylation status in the GROUPSTUDY. Recall that in GROUSTUDY a set of genes were found to be hypermethylated in 80% of cohort samples.

As a subsequent task, we tried to validate these results with TCGA dataset. It is well known that DNA methylation can drive events in the pathogenesis of gastric cancer. Modifications in DNA methylation contribute to the molecular heterogeneity of gastric cancer. Detection of DNA methylation signatures can provide biomarkers for early detection, classification, assessment of the tumour prognosis, development of therapeutic strategies and patient follow-up. Validation of these results can be an important contribution to the development of novel biomarkers and to understand the biology of the cancer epigenome.

Our results showed that most genes (genes A, B, C, E, I, J, K, L, O and P) previously found as hypermethylated were hypermethylated in more than 80% of tumour samples. Results from adjacent normal samples should be interpreted with caution due to possible tumour contamination. Our results

suggest that could be a certain contamination level of normal samples in TCGA since comparison with independent normal gastric samples revealed different methylation patterns. Thus, from a total of 17 initial genes, we considered eight (genes A, B, C, E, I, J, L and O) as validated (44.6%). These genes can be fully tested as potential biomarkers of gastric cancer and may be important in the development of new therapies against this disease.

Our second major analysis consisted of gender differential expression analysis using a TCGA cohort, in order to better understand gender differential genomic patterns and their relevance to human cancer biology. We subdivided this study into two analyses.

We recall that the first analysis consists of four tests: i) tumour vs normal samples; ii) male non-paired tumour tissue samples vs female non-paired tumour tissue samples; iii) male paired tumour tissue samples vs female paired tumour tissue samples; and iv) male paired normal tissue samples vs female paired normal tissue samples. The second analysis consists of two tests: i) Male tumour samples vs Male normal samples, and ii) Female tumour samples vs Female normal samples). The second analysis was performed between tumour and normal for gender specific, we used each method individually (for each person) not taking into account the globality of samples.

To our knowledge, methods for differential expression based on RNA-Seq data cannot directly implement this feature. We decided for this approach since no better alternative was found, although we are aware that this analysis could be further explored with a statistical approach that implements paired testing, e.g. Wilcoxon paired rank test.

Focusing on the autosomal genes, for the first test of the first analysis, we found 1346 DEGs between tumour (810 upregulated in the tumour) and normal (536 upregulated in normal) tissue. *Cytokine activity* was found to be enriched, playing an important role in cell signalling, and being released by cells of the immune system, especially by monocytes and T lymphocytes. Functions such as *growth factor binding* and *growth factor activity* are recurrently found enriched in genes up-regulated in the tumour.

The analyses between male and female only with tumour samples revealed 401 DEGs specific for tumour tissue (268 upregulated in male and 133 upregulated in female). *Folate biosynthesis* pathway was found enriched (gene list). Folate is essential in the metabolism of nucleic acid precursors and several amino acids, which are required for cell division. Low levels of folate have been associated with specific cancers.

The analysis with paired samples in tumour and normal (between male and female) revealed 61 DEGs (24 in male and 37 in female) and 2 DEGs (2 in female), respectively. These two analyses show no enrichment for any gene set. For normal tissues (2 DEGs found), we expected no enrichment due to the low number of DEGs found, but in the tumour, with 61 DEGs we already expected some functional enrichment.

The second analyses between tumour and normal in gender specific showed 99 DEGs specific for male and 93 DEGs specific for female. *Growth factor activities* were found enriched in male. When secreted by tumour cells growth factors may play a major role in tumour cell progression stimulating cell division and cell migration. *Linoleic acid metabolism* pathways were also found to be enriched in male. There is literature evidence showing potential mechanisms in the effect of tumour metabolism and immune function at gender level (Field and Schley, 2004). Our results show that transcriptomic differences at gender level go beyond those expected to be found in sex chromosomes. We expect these results to shed light on the molecular differences and commonalities between genders and provide novel insights into the differential risk underlying this cancer.

As future work, we would like to pursue three goals:

i) Integrative analysis of multi-genomic data. By integrating multiple genomic data types, new fundamental information can arise to identify genomic alterations that characterise subtypes of biological and clinical significance. None of the data types can individually fully capture the complexity of the cancer genome or precisely identify the cancer-driving mechanism. However, together they can provide a new model for the discovery of new cancer subtypes and associated cancer genes (Shen et al., 2012);

ii) Extend the gender differential expression analysis to other cancer types, in particular, those with previously found with a significant difference in the incidence between genders. There are studies underway in the group for this purpose;

iii) Gene Co-Expression Networks. The correlation study of networks has increasingly risen in bioinformatics. Correlation networks enable network-based gene screening methods that can be used to find good candidate biomarkers or therapeutic targets (Langfelder and Horvath, 2008).

REFERENCES

- Akavia, U.D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H.C., Pochanard, P., Mozes, E., Garraway, L.A., and Pe'er, D. (2010). An Integrated Approach to Uncover Drivers of Cancer. *Cell* *143*, 1005–1017.
- Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol* *11*, R106.
- Anders, S., and Huber, W. (2012). Differential expression of RNA-Seq data at the gene level—the DESeq package. EMBL, Heidelberg, Ger.
- Ayala, B. (2013). TCGA Barcode. Retrieved from <https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode>
- Ayala, B. (2014). Understanding TCGA Biospecimen IDs. Retrieved from <https://wiki.nci.nih.gov/display/TCGA/Understanding+TCGA+Biospecimen+IDs>
- Bass, A.J., Thorsson, V., Shmulevich, I., Reynolds, S.M., Miller, M., Bernard, B., Hinoue, T., Laird, P.W., Curtis, C., Shen, H., et al. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* *513*, 202–209.
- Beaver, J.E., Tasan, M., Gibbons, F.D., Tian, W., Hughes, T.R., and Roth, F.P. (2010). FuncBase : a resource for quantitative gene function annotation. *Bioinformatics* *26*, 1806–1807.
- Boundless Microbiology (2016). Annotating Genomes.
- Bozeman, M.T. (2014). CNV Quality Assurance Tutorial. Golden Helix SNP Var. Suite User Guid. 2.
- Brambilla, E., Travis, W.D., Colby, T.V., Corrin, B., and Shimosato, Y. (2001). The new World Health Organization classification of lung tumours. *Eur. Respir. J.* *18*, 1059–1068.
- Broad Institute of MIT and Harvard (2016). Affymetrix SNP6 Copy Number Inference Pipeline.
- Chang, K., Creighton, C.J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., Ally, A., Balasundaram, M., Birol, I., Butterfield, Y.S.N., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* *45*, 1113–1120.
- Choi, Y.Y. (2015). Molecular Dimensions of Gastric Cancer - Translational and Clinical Perspectives. *J. Pathol. Transl. Med.* *50*, 1–9.
- Deraitus, M., and Freeman, K. (2001). Essentials of cell biology. In CHI '01 Extended Abstracts on Human Factors in Computing Systems - CHI '01, (New York, New York, USA: ACM Press), p. 475.
- Dorak, M.T., and Karpuzoglu, E. (2012). Gender Differences in Cancer Susceptibility: An Inadequately Addressed Issue. *Front. Genet.* *3*, 1–11.
- Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W.A., Hou, L., and Lin, S.M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* *11*, 587.
- Du, P., Huang, S., Kibbe, W. a, and Lin, S. (2014). Analyze Illumina Infinium methylation microarray data Major classes of Illumina methylation microarray data.
- Esnaola, M., Castelo, R., and González, J. (2013). tweedEseq: analysis of RNA-seq data using the Poisson-Tweedie family of distributions. *Dim* 1–12.
- Field, C.J., and Schley, P.D. (2004). Evidence for potential mechanisms for the effect of conjugated linoleic acid on tumor metabolism and immune function : lessons from n X 3 fatty acids 1 – 4. 1190–1198.
- Gene Ontology Consortium (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* *32*, 258D–261.
- Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge,

Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* *5*, R80.

Glickman, M.E., Rao, S.R., and Schultz, M.R. (2014). False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *J. Clin. Epidemiol.* *67*, 850–857.

GloboCan (2012). Stomach Cancer Estimated Incidence, Mortality and Prevalence Worldwide in 2012. Retrieved from http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx

Gresham, D., Dunham, M.J., and Botstein, D. (2008). Comparing whole genomes using DNA microarrays. *Nat. Rev. Genet.* *9*, 291–302.

Hale, R.L. (1981). Cluster analysis in school psychology: An example. *J. Sch. Psychol.* *19*, 51–56.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* *22*, 1760–1774.

Houle, J.L. (2008). A differential gene expression algorithm for comparative microarray analysis. Carleton University Ottawa, Canada.

Huang, W.-Y., Hsu, S.-D., Huang, H.-Y., Sun, Y.-M., Chou, C.-H., Weng, S.-L., and Huang, H.-D. (2015). MethHC: a database of DNA methylation and gene expression in human cancer. *Nucleic Acids Res.* *43*, D856–D861.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. (2002). The Ensembl genome database project. *Nucleic Acids Res.* *30*, 38–41.

Hudson (Chairperson), T.J., Anderson, W., Aretz, A., Barker, A.D., Bell, C., Bernabé, R.R., Bhan, M.K., Calvo, F., Eerola, I., Gerhard, D.S., et al. (2010). International network of cancer genome projects. *Nature* *464*, 993–998.

Jin, B., Li, Y., and Robertson, K.D. (2011). DNA Methylation: Superior or Subordinate in the Epigenetic Hierarchy? *Genes Cancer* *2*, 607–617.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* *38*, D355–D360.

Karolchik, D. (2003). The UCSC Genome Browser Database. *Nucleic Acids Res.* *31*, 51–54.

Klein, S.L., Hodgson, A., and Robinson, D.P. (2012). Mechanisms of sex disparities in influenza pathogenesis. *J. Leukoc. Biol.* *92*, 67–73.

Klinger, C. (2014). Data Levels. Retrieved from <https://wiki.nci.nih.gov/display/TCGA/Data+level>

Laddha, S. V, Ganesan, S., Chan, C.S., and White, E. (2014). Mutational Landscape of the Essential Autophagy Gene BECN1 in Human Cancers. *Mol. Cancer Res.* *12*, 485–490.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* *9*, 559.

LEE, K.W., LEE, H.J., CHO, H.Y., and KIM, Y.J. (2005). Role of the Conjugated Linoleic Acid in the Prevention of Cancer. *Crit. Rev. Food Sci. Nutr.* *45*, 135–144.

Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A., and Dewey, C.N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* *26*, 493–500.

Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D., and Darnell, J. (2000). *Molecular Cell Biology* (New York: W. H. Freeman).

Lokk, K., Modhukur, V., Rajashekar, B., Märten, K., Mägi, R., Kolde, R., Koltšina, M., Nilsson, T.K., Vilo, J., Salumets, A., et al. (2014). DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biol.* *15*, r54.

Ma, J., Malladi, S., and Beck, A.H. (2016). Systematic Analysis of Sex-Linked Molecular Alterations and Therapies in Cancer. *Sci. Rep.* *6*, 19119.

McLendon, R., Friedman, A., Bigner, D., Van Meir, E.G., Brat, D.J., M. Mastrogiannis, G., Olson, J.J., Mikkelsen, T., Lehman, N., Aldape, K., et al. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* *455*, 1061–1068.

National Institute of Health (2016a). Why are Tissue Samples Important to Cancer Genomics? Retrieved from <http://cancergenome.nih.gov/cancergenomics/tissuesamples>

National Institute of Health (2016b). Access Tiers. Retrieved from <https://tcga-data.nci.nih.gov/tcga/tcgaAccessTiers.jsp>

Norusis, M. (2009). Cluster analysis. In *SPSS 16.0 Statistical Procedures Companion*, (Prentice Hall), p. 648.

Pearce, M.S., and Parker, L. (2001). Childhood cancer registrations in the developing world: Still more boys than girls. *Int. J. Cancer* *91*, 402–406.

Pereira, B., and Rueda, O. (2015). Differential Expression Analysis using edgeR and DESeq2 edgeR Workflow. *6*, 1–7.

Pihl, T. (2013). RNASeq Version 2. Retrieved from <https://wiki.nci.nih.gov/display/TCGA/RNASeq+Version+2>

Pinheiro, I., Dejager, L., and Libert, C. (2011). X-chromosome-located microRNAs in immunity: Might they explain male/female differences? *BioEssays* *33*, 791–802.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* *35*, D61–D65.

Puig, P., and Valero, J. (2007). Characterization of count data distributions involving additivity and binomial subsampling. *Bernoulli* *13*, 544–555.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* *26*, 841–842.

Schreiber, G., and Walter, M.R. (2011). Cytokine receptor interactions as drug targets. *14*, 511–519.

Schrenzel, J., Kostic, T., Bodrossy, L., and Francois, P. (2009). Introduction to Microarray-Based Detection Methods. In *Detection of Highly Dangerous Pathogens*, (Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA), pp. 1–34.

Schubert, M.L., and Peura, D.A. (2008). Control of Gastric Acid Secretion in Health and Disease. *Gastroenterology* *134*, 1842–1860.

Shen, R., Mo, Q., Schultz, N., Seshan, V.E., Olshen, A.B., Huse, J., Ladanyi, M., and Sander, C. (2012). Integrative Subtype Discovery in Glioblastoma Using iCluster. *PLoS One* *7*, e35236.

Shmulevich, I. (2014). Large-scale molecular characterization and analysis of gastric cancer. *Chin. J. Cancer* *33*, 369–370.

Singh, M., and Salnikova, M. (2015). *Novel Approaches and Strategies for Biologics, Vaccines and Cancer Therapies* (Academic Press).

Soper, S.A., and Rasooly, A. (2016). Cancer: a global concern that demands new detection technologies. *Analyst* *141*, 367–370.

Stein, L. (2001). Genome annotation: from sequence to biology. *Nat. Rev. Gnetecis* *2*, 493–503.

Storey, J. (2002). A Direct Approach to False Discovery Rates on JSTOR. *Wiley Online Libr.* *64*, 479–498.

Tan, P., and Yeoh, K.-G. (2015). Genetics and Molecular Pathogenesis of Gastric Adenocarcinoma. *Gastroenterology* *149*, 1153–1162.e3.

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., and Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. *Genome Res.* *21*, 2213–2223.

Tarazona, S., Furi, P., Ferrer, A., and Conesa, A. (2014). NOISeq : Differential Expression in RNA-seq. *2*, 22.

- Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A. Di, Nueda, M.J., Ferrer, A., and Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res.* *43*, gkv711.
- Tipney, H., and Hunter, L. (2010). An introduction to effective use of enrichment analysis software. *Hum Genomics* *4*, 202–206.
- Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Współczesna Onkol.* *1A*, 68–77.
- Vandin, F., Upfal, E., and Raphael, B.J. (2012). Finding Driver Pathways in Cancer: Models and Algorithms. *Algorithms Mol. Biol.* *7*, 23.
- Wang, E. (2013). Understanding genomic alterations in cancer genomes using an integrative network approach. *Cancer Lett.* *340*, 261–269.
- Yim, O., and Ramdeen, K.T. (2015). Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data. *Quant. Methods Psychol.* *11*, 8–21.
- Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *Omi. A J. Integr. Biol.* *16*, 284–287.
- Zhang, W. (2014). TCGA divides gastric cancer into four molecular subtypes: implications for individualized therapeutics. *Chin. J. Cancer* *33*, 469–470.
- Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLoS One* *9*, e78644.
- Zhu, Y., and Ji, Y. (2014). TCGA-Assembler Quick Start Guide. 1–21.
- Zhu, Y., Qiu, P., and Ji, Y. (2014). TCGA-Assembler: Pipeline for TCGA Data Downloading, Assembling, and Processing. *Health.Bsd.Uchicago.Edu* 1–8.

APPENDIX I – *GROUPGO* RESULTS

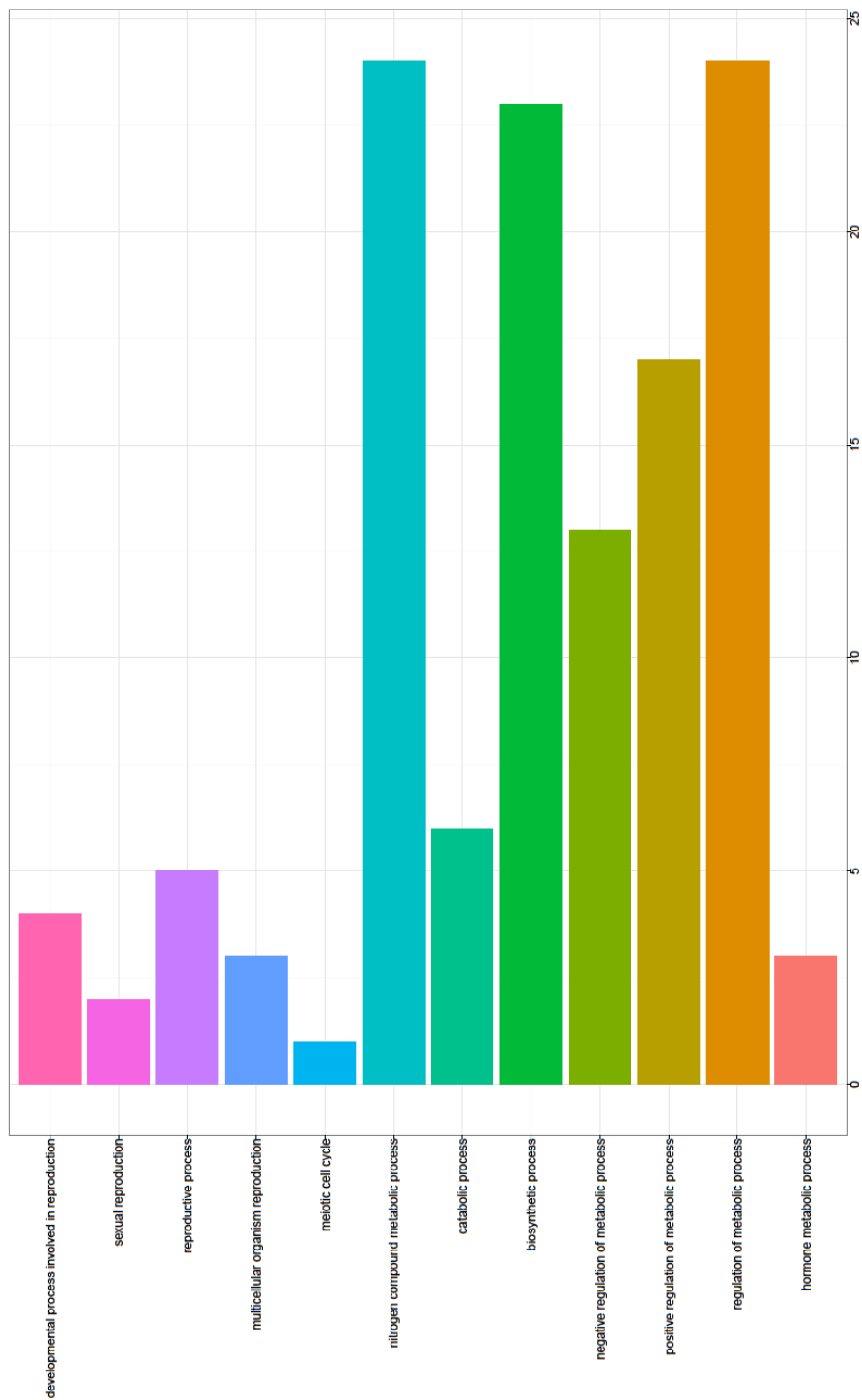


Figure 25 - GroupGO (Biological Processes) analysis for DEGs obtained by edgeR – Paired Tumour samples Male vs Female

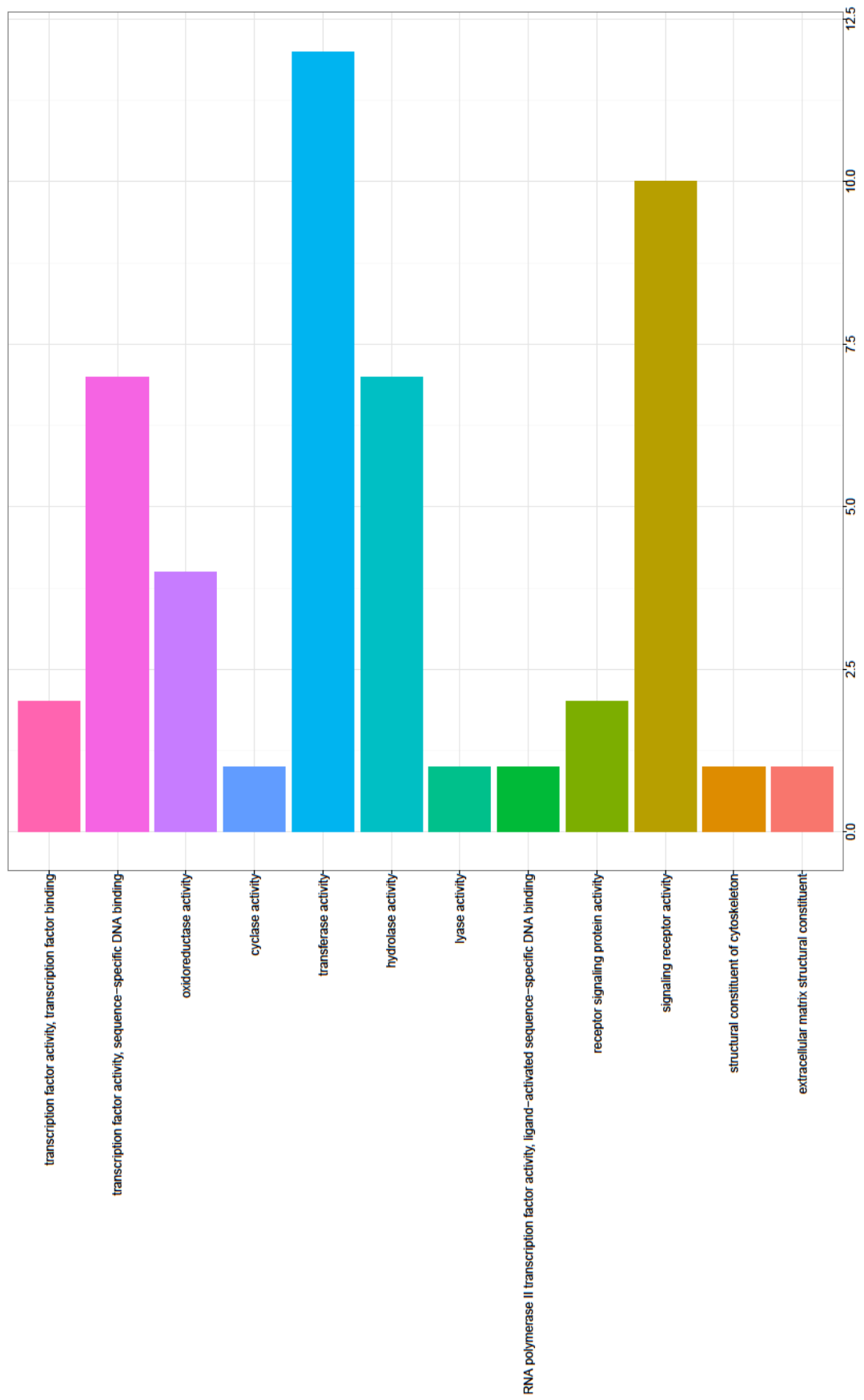


Figure 26 - GroupGO (Molecular Functions) analysis for DEGs obtained by edgeR – Paired Tumour samples Male vs Female

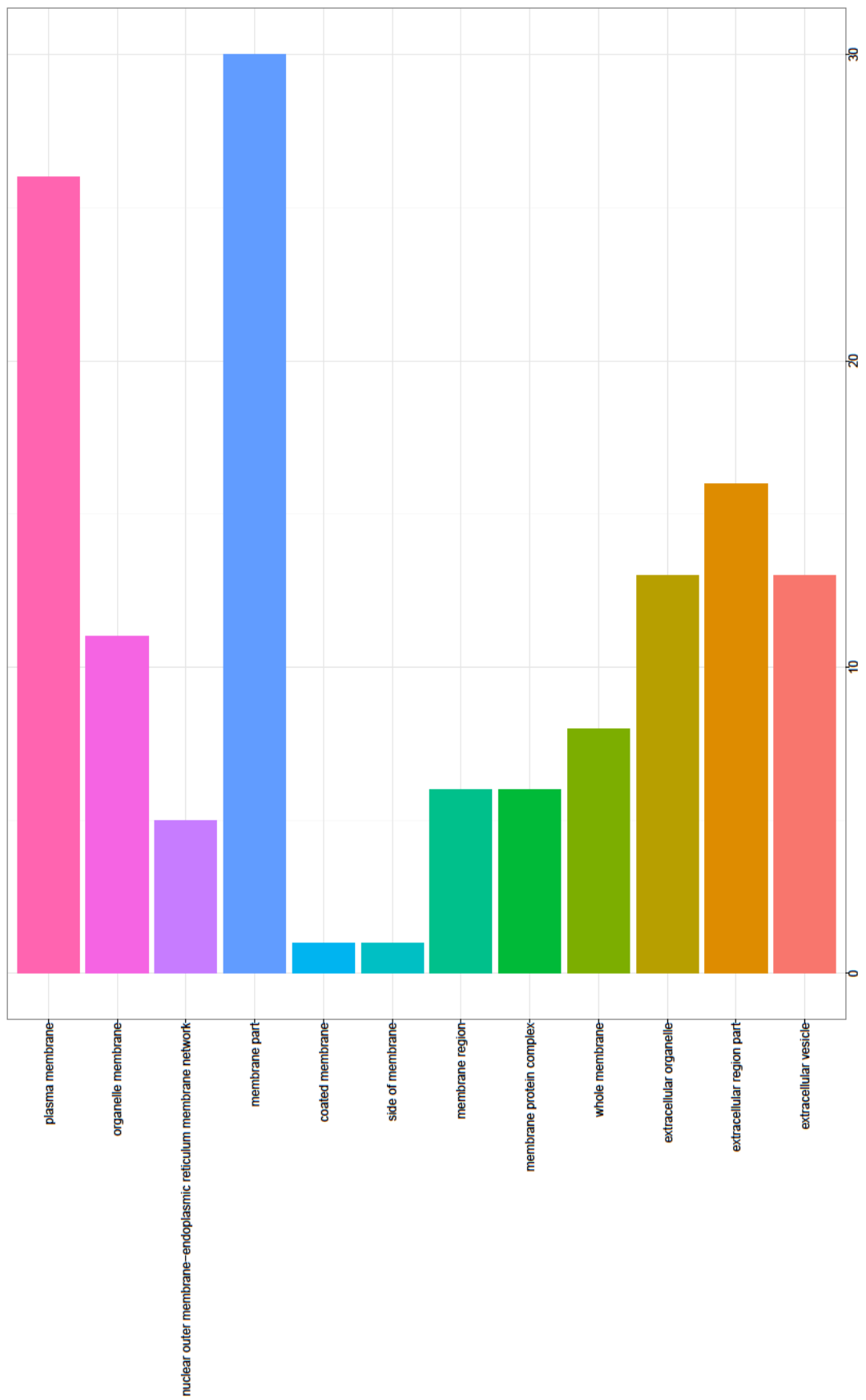


Figure 27 - GroupGO (Cellular Components) analysis for DEGs obtained by edgeR – Paired Tumour samples Male vs Female

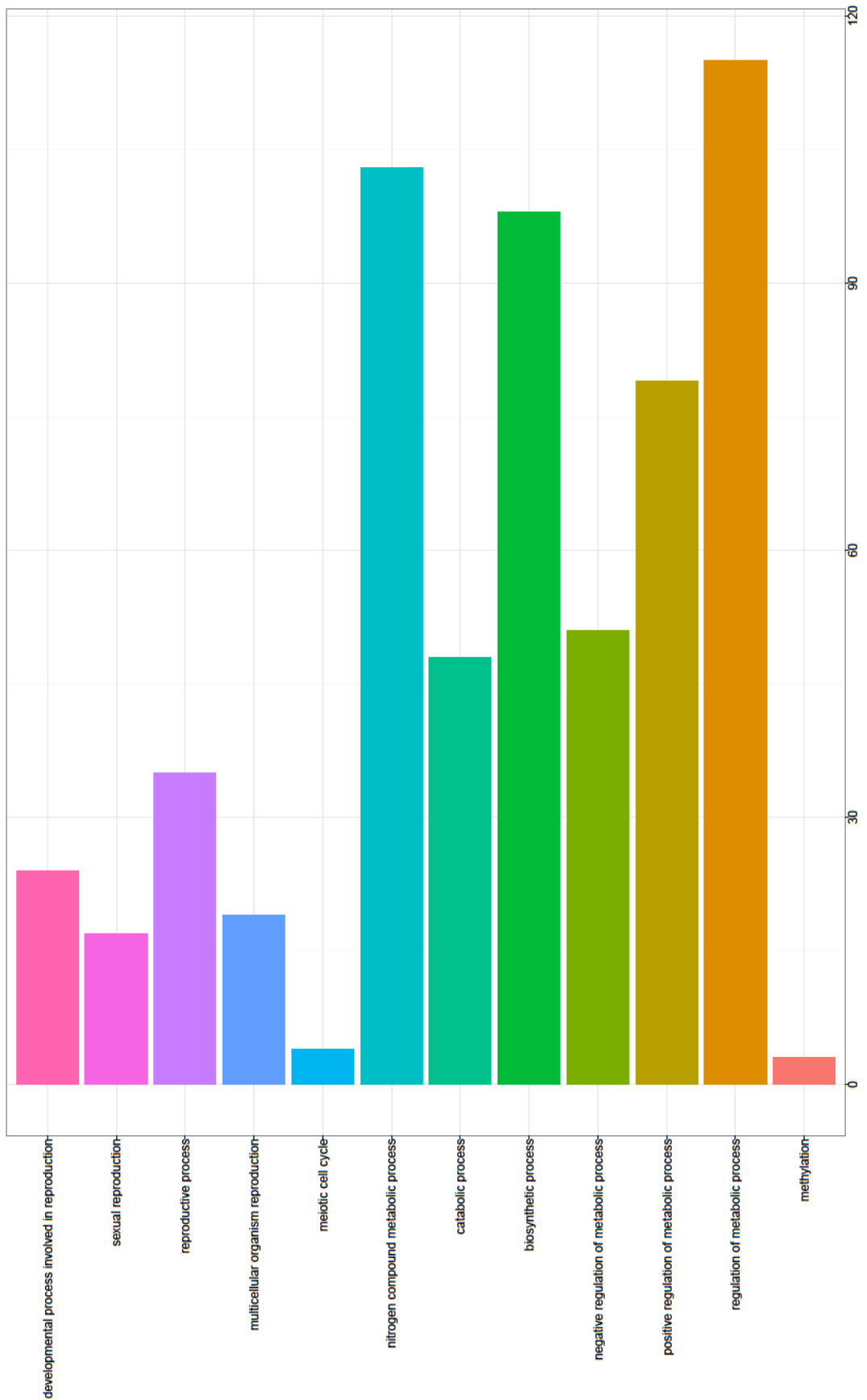


Figure 28 - GroupGO (Biological Processes) analysis for DEGs obtained by edgeR – Paired Normal samples Male vs Female

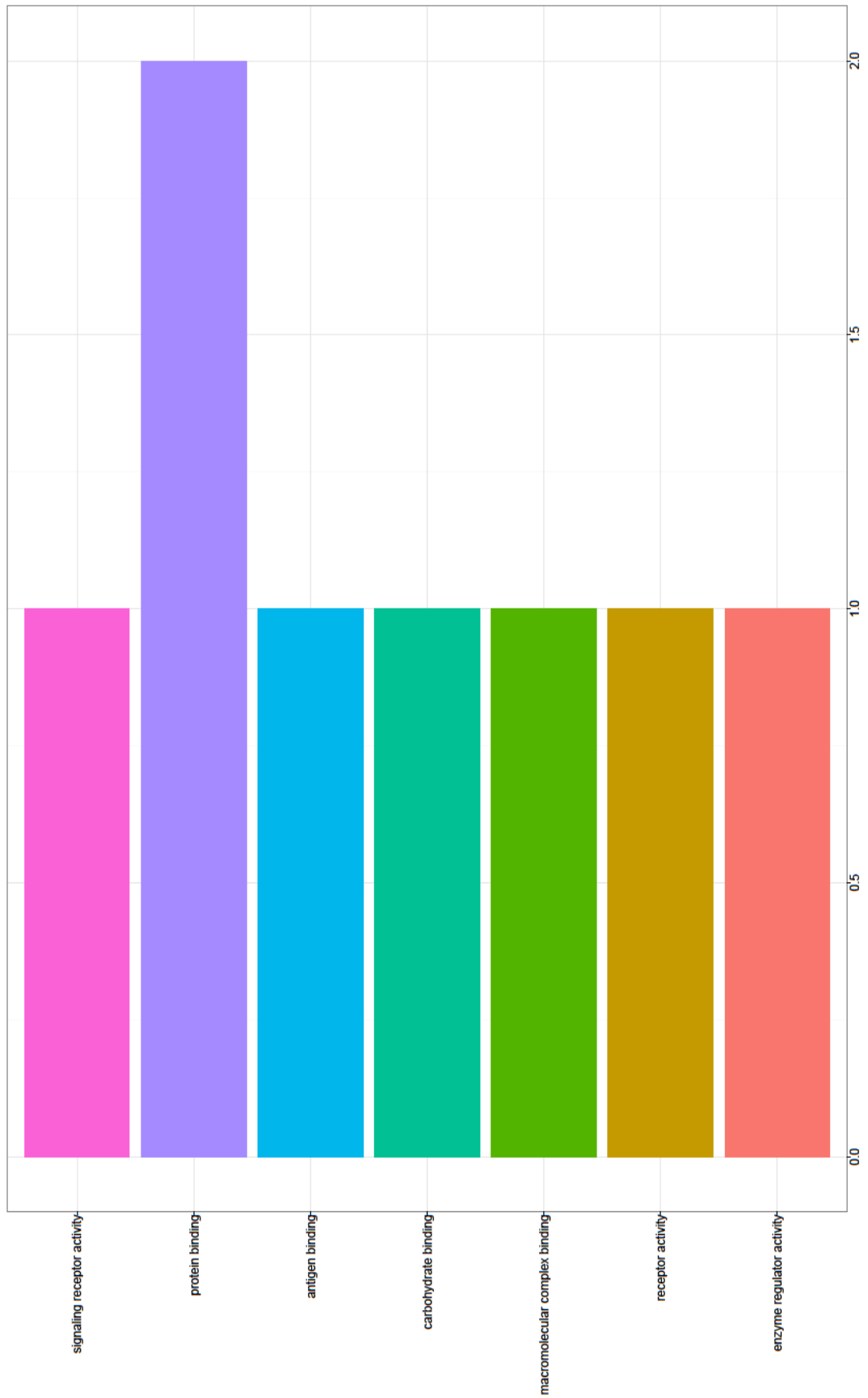


Figure 29 - GroupGO (Molecular Functions) analysis for DEGs obtained by edgeR – Paired Normal samples Male vs Female

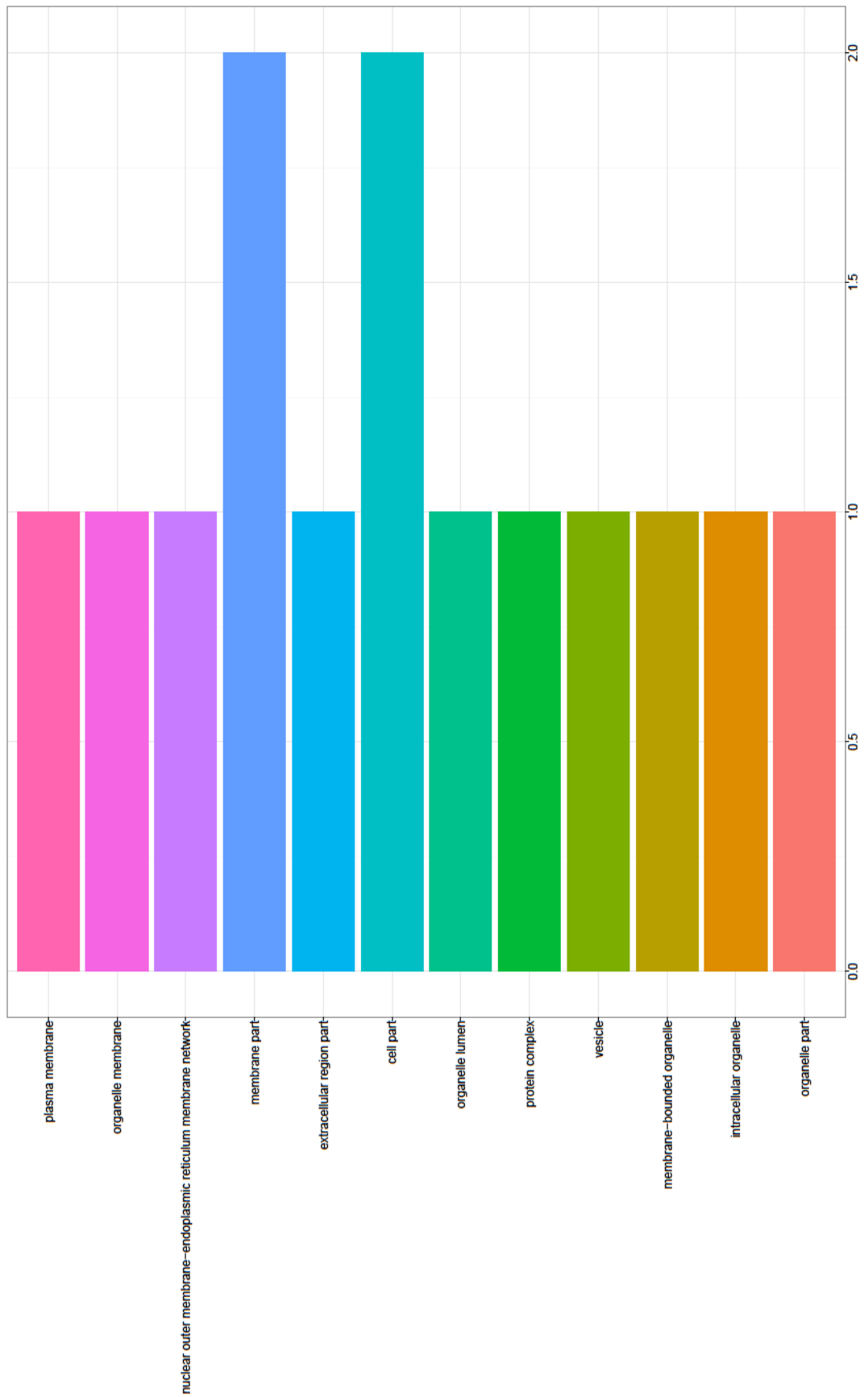


Figure 30 - GroupGO (Cellular Components) analysis for DEGs obtained by edgeR – Paired Normal samples Male vs Female