Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Medicine and Health Sciences
Department of Clinical and Molecular Medicine

Simon Geithus

Potential Biomarkers of Gastric Intestinal Metaplasia for Early Detection of Gastric Cancer: A bioinformatics study

Master's thesis in Master of Science in Molecular Medicine
Supervisor: Duan Chen
Co-supervisor: Chun-Mei Zhao
February 2022

NTNU
Kunnskap for en bedre verden

Simon Geithus

Potential Biomarkers of Gastric Intestinal Metaplasia for Early Detection of Gastric Cancer: A bioinformatics study

Master's thesis in Master of Science in Molecular Medicine
Supervisor: Duan Chen
Co-supervisor: Chun-Mei Zhao
February 2022

Norwegian University of Science and Technology
Faculty of Medicine and Health Sciences
Department of Clinical and Molecular Medicine

NTNU
Norwegian University of
Science and Technology

Acknowledgements

This thesis was actualized to complete my two years study in the master program Molecular Medicine at the Norwegian University of Science and Technology (NTNU). The thesis was performed in the Department of Clinical and Molecular Medicine in the research group Experimental Pharmacology and Surgery.

I extend my gratitude to my main supervisor Duan Chen for guiding me to the completion of this thesis with his patience, support, and feedback. You have been there when I have had any questions about the thesis and when I was uncertain on how to proceed. I will remember your guidance and carry it with me.

I extend my gratitude to my co-supervisor Chun-Mei Zhao for pushing me towards my goals. You have been there to maintain my ambitions and progress regarding the thesis.

I extend my gratitude to my fellow master student Mathilde Resell for keeping me in company and being a motivational factor when writing my thesis.

Finally, I extend my gratitude to my friends and family for always being there for me when I need them.

Abbreviations

AGA	American Gastroenterological Association
CAG	Chronic Atrophic Gastritis
CAR	Constitutive Androstane Receptor
DEGs	Differentially Expressed Genes
EGF	Epidermal Growth Factor
EIF2	Eukaryotic Initiation Factor 2
ER	Endoplasmic Reticulum
ESGE	European Society of Gastrointestinal Endoscopy
GC	Gastric Cancer
GHAI	Gastric Histological Activity Index
GIM	Gastric Intestinal Metaplasia
HER2	Human Epidermal growth factor Receptor 2
HID	High Iron Diamine
HNF	Hepatocyte Nuclear Factor
IMS	Severe Intestinal Metaplasia
IMW	Wild Intestinal Metaplasia
IPA	Ingenuity Pathway Analysis
M	gene stability value
MDS	MultiDimensional Scaling
NAG	Non-Atrophic Gastritis
NCGCs	Non-Cardia Gastric Carcinomas
OXPHOS	OXidative PHOSphorylation
PCA	Principal Component Analysis
PPIs	Proton Pump Inhibitors

PXR	Pregane X Receptor
RNA-seq	RNA-sequencing
ROS	Reactive Oxygen Species
scRNA-seq	single-cell RNA-sequencing
SCT	Single-Cell Transcriptome
SNPs	Single Nucleotice Polymorphisms
SPEM	Spasmolytic Polypeptide-Expressing Metaplasia/Pseudopyloric Metaplasia
SULT	SULfoTransferase
tSNE	t-distributed Stochastic Neighbor Embedding
UGT	UDP-GlycosylTransferase
UMAP	Uniform Manifold Approximation and Projection

Abstract

Gastric intestinal metaplasia (GIM) is an irreversible condition that carries an increased cancer risk. The development of GIM is generally unhindered due to a lack of clinical symptoms. Currently, there is no treatment for GIM, except for surveillance before GIM advances into dysplasia or cancer. The aim of this thesis was to discover biomarkers of GIM for early detection of gastric cancer by the means of bioinformatics.

The bioinformatic dataset used in this thesis included microarray data of 16 patients from whom surgical biopsies exhibited both GIM and gastric carcinoma in each patient (at St. Olavs Hospital) and single-cell RNA data of 13 patients from whom endoscopic biopsies showed wild superficial gastritis (3 patients), chronic atrophic gastritis (3 patients), GIM (6 patients), and early gastric carcinoma (1 patient)(at NCBI GEO under accession number GSE134520). The bioinformatic methods applied in this thesis included normalization with housekeeping genes, DESeq2, and Limma, data visualization with R, differential equations, and ingenuity pathway analysis.

The result showed that potential biomarkers of GIM encompassed signaling pathways (i.e., xenobiotic metabolism CAR signaling pathway, xenobiotic metabolism PXR signaling pathway, and sucrose degradation V), transcription factors (HNF1A, HNF4A, CDX2, PPARGGC1A, and GATA4) endogenous chemicals (elaidic acid and d-glucose), and the growth factor EGF.

In conclusion, the methodology applied in this thesis and the bioinformatic biomarkers of GIM found in this thesis may be useful for early detection of gastric cancer.

Graphical summary of four categories of biomarkers for gastric intestinal metaplasia for early detection of gastric cancer.

Samandrag

Gastrisk tarmmetaplasi er ein irreversibel tilstand som innehar ein auka kreftrisiko. Utviklinga av gastrisk tarmmetaplasi går generelt uhindra føre seg grunna mangelen på kliniske symptom. Der er foreløpig ingen behandling for gastrisk tarmmetaplasi, forutan overvaking før den utviklar seg til dysplasi eller kreft. Målet med denne avhandlinga var å finne biomarkørar for gastrisk tarmmetaplasi for tidleg oppdagelse av magekreft ved bruk av bioinformatiske metodar.

Det bioinformatiske datasettet som blei brukt inkluderte mikroarraydata frå 16 pasientar gjennomgått kirurgiske biopsiar med både gastrisk tarmmetaplasi og magekreft i kvar pasient (ved St. Olavs Hospital) og enkeltcelle RNA data frå 13 pasientar etter endoskopisk biopsiar med "wild" overfladisk gastritt (3 pasientar), kronisk atropisk gastritt (3 pasientar), gastrisk tarmmetaplasi (6 pasientar) og tidleg magekreft (1 pasient)(ved NCBI GEO under tilgangsnummer GSE134520). Dei bioinformatiske metodane brukt var normalisering med husholdningsgen, DESeq2 og Limma, data visualisering med R, differensial likning og ingenuity pathway analysis.

Resultatet viser at potensielle biomarkørar for gastrisk tarmmetaplasi inneber signaliseringsvegane: xenobiotisk metabolisme CAR signalveg, xenobiotisk metabolisme PXR signalveg og sukrosenedbrytning V; transkripsjonsfaktorane HNF1A, HNF4A, CDX2, PPARGGC1A og GATA4; endogene kjemikaliane elaidinsyre og d-glukose; og vekstfaktoren EGF.

Konklusivt viser avhandlinga at metoden og dei bioinformatiske markørane kan vere nyttige for tideleg funn av magekreft.

Contents

1 Introduction

Cancers originating in the stomach are the sixth most common type of cancer and the second leading cause of cancer-related deaths worldwide [1]. In Norway, gastric cancer (GC) is the seventeenth most common cancer and the twelfth leading cause of cancer-related deaths [2].

Apart from surgical removal of the tumor, there are no curative means of treating gastric cancer. Retrospective evidence indicates that a resection margin of 2–6 cm decreases the likelihood of remaining cancer cells [3].

1.1 Tumorigenesis of gastric cancer

Several genetic and environmental risk factors induce the development of GC. The development from normal epithelia to chronic gastritis is most commonly due to infection of H. pylori. Additional risk factors includes high salt consumption, smoking, gene polymorphisms, and immune responses. The risk factors increases the genetic instability and mutation rate, causing the condition to develop into GC (Figure 1.1) [4].

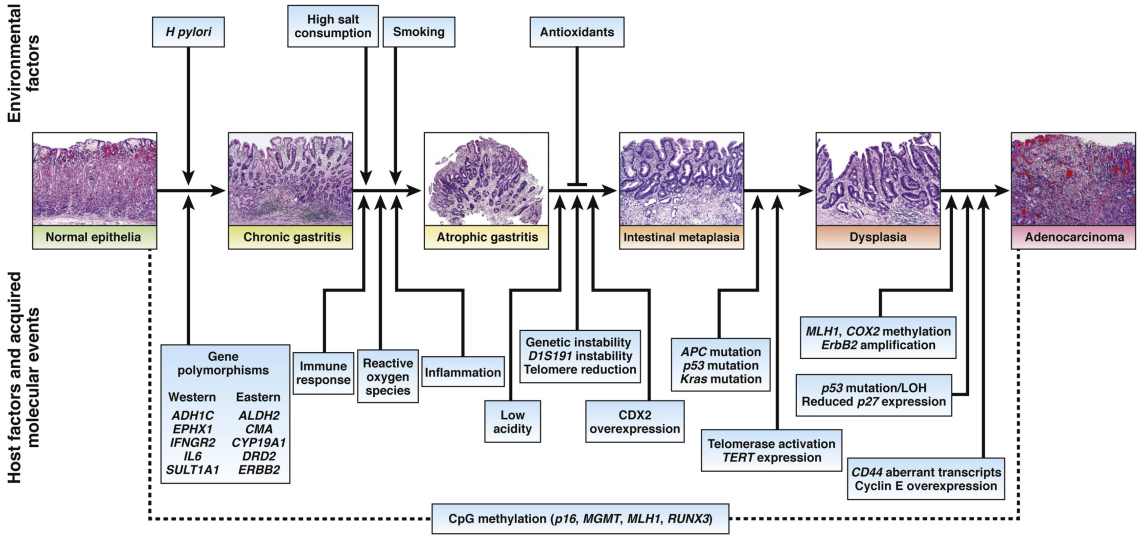


Figure 1.1: Tumorgenesis from normal epithelial tissue to adenocarcinoma of stomach. Note: environmental factors, host factors and acquired molecular events [4].

While H. pylori is the most investigated cause of GC [5, 6], estimated to contribute to 89.0 % of all non-cardia gastric carcinomas (NCGCs) [6], gastritis itself is likely the mechanism responsible for tumorigenesis [7]. Gastritis from hypergastrinemia from the long-term usage of proton pump inhibitors (PPIs), a commonly prescribed medication, is thus an additional risk factor that needs to be investigated [7, 8]. Since GC usually takes several decades to develop, it has been stipulated that the development of cancer can only be caused by stem cells and their niche [9]. Long-lived DCLK1$^+$ tuft cells [10], some self-replicating enterochroffamin-like cells, and probably other neuroendocrine cells can form cancer. Waldum et al. [7, 11] have suggested how enterochroffamin-like and stem cells can be responsible for diffuse-type and intestinal-type carcinoma respectively (Figure 1.2).

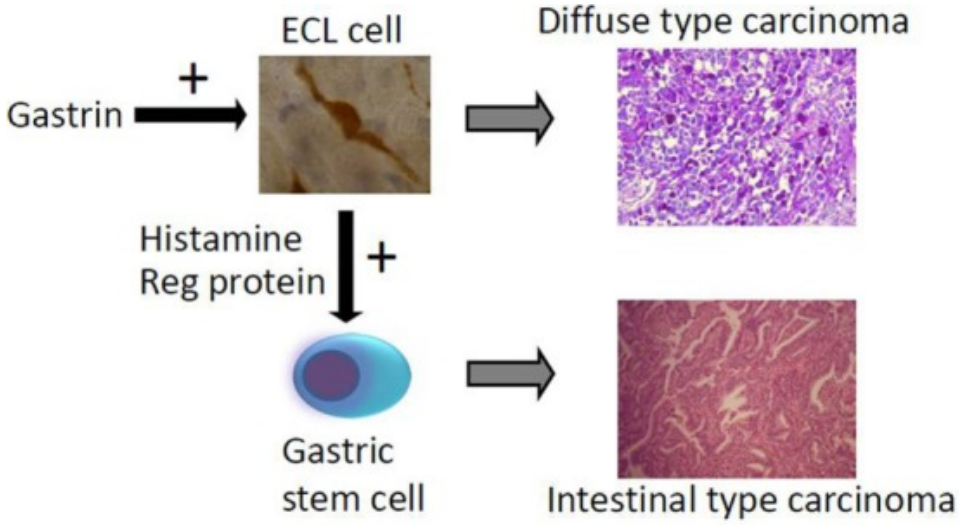


Figure 1.2: Cells potentially responsible for different types of GC through hypergastrinemia [7, 11].

Intestinal-type GC generally stems from long-term chronic inflammation. The diffuse-type GC seems to side-step the neoplastic progression, which the intestinal-type GC undergoes [12]. The diffuse-type GC consists of scattered solitary cells or small clusters. The appearance of diffuse-type GC is thus diffuse, small, and contains indistinct shapes. Any clusters diffuse-type GC creates appear more loosely attached. The intestinal-type GC consists of a more distinct and usually large glandular structure accompanied by

papillary fold formation and solid components [13].

The difference between these cancers indicates different epidemiology, as they do not transform into the other type [7]. About 15 % of carcinomas contain overlapping traits, making some cancers difficult to differentiate [13]. Note that numerous GC types have been discovered that do not fit the Lauren classification. Despite this, Lauren classification is still the most commonly used classification to date [14]. Trends in abnormal DNA aids in subdividing cancers, but no cancer biomarker can be used to categorize cancers without error [15]. Histological features in the tumor microenvironment are considered important for the prognosis and prediction of survival [14].

1.1.1 Metaplasia in general

Metaplasia is an act of transdifferentiation from one cell type to another [16]. Metaplasia in the stomach is classified as either gastric intestinal metaplasia (GIM), because its new cellular composition is composed of intestine-like gland structures, or pseudopyloric metaplasia/spasmolytic polypeptide-expressing metaplasia (SPEM), because of its expressed spasmolytic polypeptides. GIM can be divided into two subtypes, complete and incomplete [17]. The GIM subtypes are also known as small intestine metaplasia and colonic metaplasia, respectively. Incomplete metaplasia is the closest stage to dysplasia and GC [17, 18, 19].

While metaplasia in the stomach is undeniably associated with GC development, even after H. pylori eradication [20, 21, 22], there is no direct experimental evidence tying metaplasia to GC [9, 23]. SPEM is a fairly stable condition of metaplasia [23, 9] and are reversed in acute short-term injuries [23, 24]. SPEM appears closer to the stem cell zones in the stomach than GIM and is thought to be a fairly stable lesion due to TFF2, an anti-inflammatory and tumor suppressor gene. However, SPEM is potentially a precursor to GIM [9] as GIM often arises in pre-existing atrophic SPEM [24]. While GIM does not necessarily develop into GC [20, 21, 22], adjacent GIM to gastric dysplasia and GC does share a clonal origin [9]. Therefore, the presence of metaplasia can be a

useful biomarker for GC risk stratification [23], especially with compounding risk factors such as the severity and location of the metaplasia [20, 21, 22]. GIM has been shown to not be as reversible, even with the eradication of H. pylori. Thus, when the only curative treatment is surgical resection [25] it makes an interesting subject of investigation.

Currently, the American Gastroenterological Association (AGA) and the European Society of Gastrointestinal Endoscopy (ESGE) recommend no further action against GIM, except for the eradication of H. pylori. Routine surveillance (3-, 5-year interval) is only done in incomplete metaplasia or if there are any additional risk factors, like family history and increased GIM-affected area. There is still very little evidence to support the positive effects compared of longitudinal surveillance [26, 27].

1.1.2 Gastric Intestinal Metaplasia: molecular and cellular characteristics

Complete GIM shows a fairly stable expression, with paneth cells, well-defined goblet cells, crypt base columnar cells, enterocytes with brush borders, and villus organization mirroring the small intestine. Incomplete GIM, dubbed the colonic metaplasia, is more disorganized, harboring immature goblet cells, irregular mucin droplets, or glands with hybrid gastric/intestinal morphologies (Figure 1.3) [17, 24].

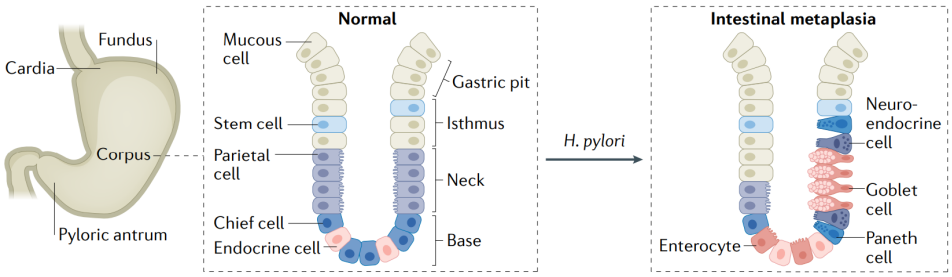


Figure 1.3: An example of the change in cellular composition from a normal tissue to intestinal metaplasia tissue due to the infection of H. pylori. Edited from Yeoh et al. [28].

General identifiers for GIM are the histological characteristics of intestine-like gland structures and intestine-specific cell lineages along with cell markers CDX1, CDX2,

MUC2, TFF3, and Villin. Sialic acid and/or sulfate residues can be detected by Alcian blue staining at pH 2.5. Specific markers for each condition are loss of gastric mucins MUC1, MUC5AC, and MUC6 for complete GIM and Brown staining by high iron diamine (HID) for incomplete GIM. HID detects sulfomucins containing sulfate residues but not sialic acid residues which are predominantly found in complete GIM [17].

1.2 Transcriptomics of gastric cancer and metaplasia

The transcriptomics encompasses the RNA sequences, structures, transcription, translation, and functions [29]. The field of transcriptomics allows an in-depth investigation on the upregulation and downregulation in tissue and cells for pathological identification and discovery of novel targets [30]. There are currently two leading techniques within transcriptomics, microarray, which can quantify a determined set of sequences; and RNA-sequencing (RNA-seq), which uses high-throughput sequencing to capture all the sequences [31]. Microarray was mainly used before 2015 and is usually replaced by RNA-seq in recent studies [32]. RNA-seq is a rapidly expanding field, with single-cell RNA-seq (scRNA-seq) becoming a common form of RNA-seq [33].

scRNA-seq can separate cell populations and highlight specific cell population biomarkers. A quantification through scRNA-seq found a significant increase in enterocytes, gland mucous cells (GMCs), and proliferative cells, and a decrease in enteroendocrine cells, when comparing GIM with non-atropic gastritis. Individual cells can be used for a biomarkers for GIM before any morphological distinguishable change has been made. The most notable example for biomarker for GIM is HES6 which is expressed specifically in early goblet cells [34]. However, not all GIM develop into GC [35], the previously mentioned genes CDX1, CDX2, and MUC2, which are general biomarkers for GIM [17], were differentially upregulated in GIM which didn't progress to GC in a study with 75 % negative H. pylori samples [35]. The genes KLF5, GATA4, and GATA6 are recurrently expressed ($\sim$ 30 %) where only one was expressed simultaneously. The common transcription regulator HNF4α downstream of all of these genes decreases cell proliferation

when inhibited [36]. Rare cancer variants can be discovered using scRNA-seq detecting cell subgroups. A chief cell subgroup expressing e.g. LIPF and PGC along with RNF43 with the Wnt/β-catenin signaling pathway are activated consistently with the fundic gland-type gastric adenocarcinoma [37].

1.3 Computational methods and challenges in transcriptomics

1.3.1 Background Correction

Background correction is used to remove noise, and hence strengthen or discover hidden expression differences between samples [38, 39]. However, a background correction is not without fault as it increases the rate of false positives [40, 41]. To avoid an increase in the false discovery rate some investigators refrain from using background correction, stating that the noise reduction is not sufficient reason to justify potential errors [42, 43].

1.3.2 Normalization

Normalization is the reduction of noise between samples and sample groups [44]. Currently, there is no standard for normalization within the scientific community in biology [45, 46, 47]. Reference genes used for normalization are constantly being scrutinized [47, 48, 49], and more sophisticated methods are continuously being developed and refined [50, 51, 52]. The constant development in the field led many investigators to use methods or genes that have been proven ineffective or sub-optimal compared to their newer counterparts. The downstream result of different normalization methods can diverge substantially [45, 46, 53]. However, no single method can be proven optimal and to be considered the gold standard today. Each normalization method has its strong points and drawbacks [53, 54, 55].

The normalization techniques can be separated into three main categories: 1) Samples normalized based on the entire sample, also known as global normalization [53]; 2) two

or more housekeeping genes being used as a basis of normalization [45]; and 3) a single housekeeping gene being used to normalize the samples [47]. Note that outputs within these main categories may differ greatly [45, 53].

Global normalization methods have become more prominent recently, with articles comparing the different methods [53, 54]. The more recent methods are shown to be more efficient and sophisticated than older methods. However, many older methods are still in use today [53]. Results from using the different versions of the same program diverge, displaying the sophistication and rapid developmental progress [54].

Multiple housekeeping gene normalization has been recommended compared to single housekeeping gene normalization. Several housekeeping genes should increase stability and reproducibility. However, an increase of housekeeping genes will not cover for badly selected genes [45, 56]. Multiple housekeeping gene normalization has been a benchmark when comparing the efficacy of global normalization methods [51, 53]. There are several ways of normalization using multiple housekeeping genes, from taking the average of the genes [57] to using Box-Cox [58].

A common method of normalization is single housekeeping gene normalization, where all results are based on the assumption of one constant gene. Unfortunately, the assumption that a housekeeping gene is constant and will not diverge based on biological differences is generally false and will give inaccurate numbers [59, 60]. Currently, there exists several traditional housekeeping genes used for normalization. Many of these genes are still in use today [45], but are disputed as non-ideal normalization factors when comparing them across tissues or tissue conditions [45, 56, 61, 62]. The general recommendation is to use multiple housekeeping genes to reduce the impact of a single gene [45, 56]. The verification of traditional genes and investigation after novel genes is still an ongoing research area [48, 63, 64]. Even though there has been no gold standard for housekeeping genes used as a reference across multiple tissues, it has been shown that some genes are more stable than others between certain tissue comparisons [65, 66].

Neither global normalization nor normalization based on reference genes should be ac-

cepted based on prior assumptions without validating the method for the specific sample [46, 67]. Housekeeping genes for normalization has been used when comparing GC and normal gastric tissue (Table 1.1) [64, 65, 66]. Several computational methods have been developed for this purpose. One of these programs is geNorm, which automatically calculates the gene stability measure (M) for all genes in a given set of samples [56].

Table 1.1: A list of stably expressed genes when comparing normal stomach tissue vs. GC tissue

Kwon et al.[64]		Rubie et al.[65]	Rho et al.[66]
CTBP1	OAZ1	AGPAT1	RPL29
CUL1	PAPOLA	B2M	RPL29-B2M
DIMT1L	SPG21	CAPN2	B2M
FBXW2	TRIM27	CYCC (CCNC)	B2M-GAPDH
GPBP1	UBQLN1	PMM1	
LUC7L2	ZNF207	SDHA	

1.3.3 Data visualization

A method for visual analysis is dimensional reduction. Dimensional reduction gives highly dimensional data a visual representation, while still preserving relevant information [68]. Dimensional reduction is commonly used by single-cell RNA sequencing studies [69, 70, 71, 72]. Whole tissue samples are investigated, [73], but no studies have been found performing whole tissue analysis with dimensional reduction in GIM or GC.

One of the initial methods available for dimensional reduction was principal component analysis (PCA) that compares the differences between objects [74, 75]. However, several methods have been developed since, such as Multidimensional scaling (MDS) that detects the similarity between objects [76]. The more recent powerful methods are t-distributed stochastic neighbor embedding (tSNE) and uniform manifold approximation and projection (UMAP) [77]. A common computational tool using tSNE and UMAP is Seurat in R. Seurat can be used for many pre-processing steps and visualization purposes. Seurat uses a weighted nearest neighbor procedure and focuses on preserving the spatial information [72]. UMAP is generally considered to be outperforming tSNE but

it is currently disputed [78], suggesting the need to test both.

A method for data visualization without dimensional reduction is heatmaps. A gene expression heatmap uses a selected set of genes and plots them toward the selected samples [79]. Clustering of the samples and genes are done with different correlation methods [80] and/or manually based on pre-evalueted factors [81].

1.3.4 Ingenuity Pathway Analysis (IPA)

IPA (Qiagen) is a web-based software application for the analysis of omics data. The database consists of a large repository of curated biological interactions [82]. The analysis uses a sophisticated prediction model for upstream and downstream biological interactions based on differentially expressed data. This prediction model gives specific pathways a p-value and a z-score and can suggest novel pathways based on correlated genes. The p-value is the statistical chance the pathway is significant and the z-score indicates how strongly the pathway is inhibited or activated [83]. RNA studies have previously been conducted on tissue samples with IPA investigating GC [84] and GIM [35]. A follow-up DNA study has investigated SNPs in GIM for potential targets [85].

2 Aims of thesis

2.1 Principal objective

Most current GC studies are focused on the identification and validation of cancer biomarkers from the tumor tissues that are usually not specific for the early stages but being detected in advanced stages of GC, and therefore cannot be used for early GC detection. The principal objective was to discover the early stage biomarkers for GC.

2.2 Secondary objectives

• To investigate the different methods of data preparation and their usefulness towards statistical differentiation and interpretation of the data.

• To find the potential biomarkers for GIM. Thus, two different methods of analysis was proposed:

o visualize the data with different computational techniques to see how well different developmental stages from GIM to GC can be differentiated;

o use Ingenuity Pathway Analysis (IPA) to data-mine, visualize and find potential biomarkers.

3 Method

3.1 Sampling of patients

The samples were surgically taken from 16 cancer patients during gastrectomy at St. Olav's hospital (REK 2012-1029). Each patient had 4 samples taken from them, except for patient no. 7 and 9. Patient no. 7 had 3 samples and patient no. 9 had 8 samples. The last 4 samples from patient no. 9 were taken at a later date and had an unknown origin as the patient originally underwent a total gastrectomy, therefore the last 4 samples were not analyzed further. The samples from patient no. 6 did not go through pathological evaluation and were not analysed further. The 4 sample locations were the curvatura minor, curvatura major, antrum, and cardia/corpus ventriculi. One or more of the sample sites in each patient contained cancer, the remaining contained control sites of normal tissue or GIM outside the cancers influence. The samples were evaluated twice through biopsy and pathology, and analyzed with Illumina microarray chips. The gene expression results were extracted as an excel document, while the control data results were extracted in in a txt format.

3.2 Sample data preparation

Data wrangling was performed using R version 4.1.1 [86] with the tidyverse package version 1.3.1 [87]. R code 3.1 shows how the gene expression data was imported and filtered by removing redundant samples using the sample_vector list containing patient samples that were not used for further analysis. The dataframe was then saved as a text file to be imported into lumi [88]. The control data was similarly imported from a text file to the dataframe df_control before it was filtered and saved as a text file.

R code 3.1: Data importing and preparation for lumi

```
df <- read_xlsx("path/data.xlsx", skip = 7)
df <- df %>% select(-contains(sample_vector))
write.table(df, path/data.txt, sep = "\t")
```

Lumi version 2.44.0 [88] was then used to filter the unexpressed genes from the data. Genes with a lower detection threshold than 1 % in 10 or less samples were removed.

R code 3.2: Gene filtering with lumi

```
lumiObject <- lumiR("path/data.txt",
                    sep = "\t",
                    lib.mapping = "lumiHumanIDMapping",
                    columnNameGrepPattern = list(
                    exprs="AVG_SIGNAL", detection="Detection"),
                    annotationColumn=c("TargetID", "ProbeID"))
presentCount <- detectionCall(lumiObject, Th = 0.01)
lumiMatrix <- lumiObject[presentCount >= 11,]
```

The lumi object's expression values were saved in a new dataframe and new gene annotations were added. The sample names were also renamed to remove redundant sample set numerals.

R code 3.3: Sample annotation preparation

```
df_lumi <- as.data.frame(exprs(lumiMatrix))
IDs <- as.data.frame(nuID2IlluminaID(lumiMatrix))
df_lumi <- cbind(IDs[c(2,5)], df_lumi)
df.colnames <- colnames(df_lumi)
df.colnames <- c(df.colnames[c(1,2)],
    paste0("S", str_sub(df.colnames[-c(1,2)], -5, -1), sep= ""))
colnames(df_lumi) <- df.colnames
```

The sample histology was imported as p_df and filtered to be used as a design matrix shown in R code 3.4. The samples in sample_vector were removed before the samples were filtered and categorized as GC, metaplasia or normal as p_df_c/m/n respectively. Samples were categorized as GC if they had a pathological evaluation as cancer or a biopsy classification of 1 or 2 out of 5. Samples were categorized as GIM if they did not classify as GC and they contained a scoring that indicated GIM. Samples were catego-

rized as normal if they were classified as neither GC or GIM. After the categorization a new column concerning the sample histology was made. The dataframes were then merged and arranged by SampleID.

R code 3.4: Patient histology data wrangling

```
p_df <- read_xslx("path/Patient_Histology.xlsx", skip = 1)
p_df <- p_df %>% filter(!Sample_ID %in%
    sample_vector)
p_df_c <- p_df %>% filter(Cancer == "yes" | Tumor_sample <=2)
p_df_m <- p_df %>% filter(Intestinal_metaplasia &
    (is.na(Cancer) & Tumor_sample > 2))
p_df_n <- p_df %>% filter(Intestinal_metaplasia == 0 &
    (Tumor_sample > 2 | is.na(Tumor_sample)) & is.na(Cancer))
p_df_c$Histology <- "C"
p_df_m$Histology <- "IM"
p_df_n$Histology <- "N"
p_df <- rbind(rbind(p_df_c, p_df_m), p_df_n)
p_df <- p_df %>% arrange(SampleID)
```

A single-cell transcriptome (SCT) atlas for premalignant lesions and early gastric cancer [34] were modified for data visualization shown in R code 3.5. All genes that had an adjusted p-value above 5 % and were present in more than one cell type were removed from the dataset. Genes without a match in the sample dataframe was cross-referenced with the UniProt knowledgebase [89]. The genes from the single cell transcriptome were then changed to coincide with the microarray data.

R code 3.5: Preparing the single cell transciptome

```
sca_df <- read_xlsx("path.xlsx", skip = 1)
sca_df <- sca_df %>% filter(p_val_adj < 0.05)
duplicate_counts <- data.frame(table(sca_df$"marker genes"))
duplicates <- duplicates[duplicates$Freq >1,][1]
sca_df <- sca_df %>% filter(
    !duplicates$"marker genes" %in% duplicates$Var1)
exc_genes <- setdiff(sca_df$"marker genes", df$TargetID)
sca_df$"marker genes"[which(sca_df$"marker genes" ==
    "Old_gene_name")] <- "New_gene_name"
```

The data from the SCT [34] was also used to confirm the findings from the tissue

data by using two different datasets. It was first uploaded into Seurat version 4.0.4 [72], and filtered identically to the original paper. I.e. cells expressing less than 400 genes, more than 7000 genes or cells containing more than 20% genes correlated with the mitochondria were filtered out. The data was extracted from Seurat and were kept as aggregated and non-aggregated dataframes. The aggregated data was kept for DESeq2 normalization for heatmap visualization and the non-aggregated data was kept for DESeq2 normalization and the differential expression equation.

R code 3.6: Preparing the single cell transcriptome data

```
SingleSeurat_df <- CreateSeuratObject(df, project = "filtering")
SingleSeurat_df[["percent.mt]] <-
        PercentageFeatureSet(SingleSeurat_df, pattern="^MT-")
Feat <- subset(SingleSeurat_df, subset=nFeature_RNA > 400 &
        nfeature_RNA < 7000 & percent.mt < 20)
sampleDF <- as.data.frame(t(FetchData(Feat, vars =
        temp@assays[["RNA"]]@counts@Dimnames[[1]])))
sampleDF_A <- rowMeans(sampleDF)
```

3.3 Normalization with housekeeping genes

The housekeeping genes from table 1.1 were extracted with the hk_gene_list and made into a new dataframe. The gene names were made unique for differentiation at a later stage.

R code 3.7: Creation of the housekeeping gene dataframe

```
hk_df <- df_lumi %>% filter(ILMN_Gene %in% hk_gene_list)
hk_df$ILMN_Gene <- make.unique(hk_df$ILMN_Gene)
```

The computational method geNorm, from ctrlGene version 1.0.1 [90], was used to check for gene stability. R code 3.8 shows the function applied to the hk_df from R code 3.7. All genes were inserted as rownames. The geNorm function ran using the transposed version of the matrix as the geNorm function uses the cols as its gene input and the rows as the sample input. geNorm was not performed on the whole dataset as this would

significantly increase the computational time.

R code 3.8: geNorm code for the most stably expressed genes

```
row.names(hk_df) <- hk_df[,1]
geNorm(t(hkg_df[-c(1,2)]), ctVal=FALSE)
```

A visual judgment was performed by creating graphs, shown in R code 3.9. An average for each gene row were calculated and divided upon. The new dataframe was plotted in ggplot and was visually judged for disparity. Different subsets of the housekeeping genes were created to confirm gene selection.

R code 3.9: Graph visualization of housekeeping genes stability

```
gene_avg_vector <- rowMeans(hk_df[-c(1,2)])
hk_df[-c(1,2)] <- hk_df[-c(1,2)] / gene_avg_vector
hk_df %>%
    pivot_longer(cols = 3:ncol(.),
        names_to = "samples", values_to = "values") %>%
    ggplot(data = . , aes(x = samples, y = values,
        color = ILMN_Gene, group = ILMN_Gene)) +
    theme_light() +
    theme(axis.text.x=element_text(angle=90)) +
    geom_line()
```

A new hk_df dataframe was made with a subset of the hk_gene_list called sub_hk_gene_list. sub_hk_gene_list consisted of the bottom 17 probes with a low geNorm value and an intensity beneath 1000. A mean of each sample from the subset in hk_df was calculated and divided upon each sample in the original dataframe. The normalized result was extracted in a dataframe called norm_df.

R code 3.10: Normalization of the dataset with housekeeping genes

```
hk_df <- filter(df_lumi$ILMN_Gene %in% sub_hk_gene_list)
sample_avg_vector <- colMeans(hk_df)
norm_df <- t(t(df[-1]) / sample_avg_vector)
```

3.4 DESeq2 and Limma normalization

The data was normalized with other methods for data visualization and differential expressions for later analysis. DESeq2 version 1.32.0 [91] was used for normalization through geometric mean, shown in R code 3.11. DESeq2 is a method made for RNA-seq and takes only count data. The microarray data underwent data wrangling for DESeq2 by multiplying and rounding it to make it indistinguishable from count data.

R code 3.11: Normalization through DESeq2

```
dds <- DESeqDataSetFromMatrix(countData = round(df_lumi*100),
        colData= p_df, design = ~Histology)
dds <- estimateSizeFactors(dds)
DESeq2_norm_df <- as.data.frame(counts(dds, normalized = T))
```

Limma version 3.48.3 [92] is a specialized program for microarray data, thus the lumi-filtered data was inputted directly. R code 3.12 shows how the data was exported from R and imported into the limma. normalizeBetweenArrays was used without background correction. neqc was used for Normexp background correction. The control data, for background correction, was prepared similarly to R code 3.3.

R code 3.12: Normalization through limma

```
write.table(df_lumi, "path/lumi.txt", sep = "\t",
            row.names=F, quote=F)
df_limma <- read.ilmn(files = "path/lumi.txt",
            ctrlfiles = "path/lumi_ctrl.txt",
            sep = "\t", annotation="ILMN_Gene",
            expr = "S")
limma_norm_df <- normalizeBetweenArrays(df_limma)
limma_norm_df <- neqc(df_limma)
```

3.5 Data visualization with R

Limma was used to dimentionally reduce the normalized data and create MDS plots. The MDS plot code is shown in R code 3.13.

R code 3.13: Creation of MDS plots

```
df_MDS <- plotMDS(df_norm, labels = notation)
    toplot <- data.frame(Dim1 = plot$x, Dim2 = plot$y,
                         histology = factor(p_df$Histology))
    plot <- ggplot(toplot, aes(Dim1, Dim2, colour = Histology))+
    scale_colour_manual(values =
        c("#FF6666", "#FF9933", "#3399FF")) +
    geom_point() + theme_minimal() +
    theme(legend.background=element_rect(
        fill="gray95", size=.3,linetype = "solid")) +
    ggtitle("Plottitle") +
    labs(x = "Leading logFC Dim1", y = "Leading logFC Dim2")
```

Seurat version 4.0.4 [72] was used to perform the dimensional reduction with tSNE and UMAP. Normalized data from the patients was inputted into Seurat, scaled, clustered, and plotted through Seurat functions: CreateSeuratObject, ScaleData, FindVariableFeatures, RunPCA, FindNeighbors, FindClusters, RunUMAP, and RunTSNE. The clustering methods performed were UMAP and tSNE. Specific parameters were used for plotting with tSNE and UMAP: 1) selection.method = vst, 2) nfeatures = 5000, 3) dims = 1:10 and 4) perplexity = 10. Additional plots were made with fewer genes, filtered with the SCT atlas dataframe sca_df, excluding potential noise from unrelated cells. The samples were annotated based on sample location and histology.

ComplexHeatmap version 2.8.0 [93] was used to create heatmaps. The normalized data was filtered based on the SCT from R code 3.5. A top annotation concerning the sample location and histology, and a row annotation separating the cell types were constructed. R code 3.14 shows how the annotation for the sample's histology were constructed.

R code 3.14: Heatmap annotation creation for histology

```
Histology = HeatmapAnnotation(Histology = p_df$Histology,
                                  col = list(Histology = c(
                              "Cancer" = "#FF0000",
                              "Intestinal Metaplasia" = "#FF9933",
                              "Normal" = "#0000FF")
                              ))
```

3.6 Differential equations

The differential equations were made with both DESeq2 and limma. The SCT was only tested in DESeq2. The gastric tissue samples were used in both DESeq2 and limma. R code 3.15 shows how the imported data from R code 3.11 were computed and readied in the dataframe res, before the final results were extracted to a separate excel document with writexl version 1.4.0 [94].

R code 3.15: DESeq2 differential equation

```
dds <- DESeq2(dds)
res <- results(dds, contrast = c("histology", "IM", "N"))
res <- res %>% arrange(padj)
res <- cbind(rownames(res), res)
colnames(res)[1] <- "Genes"
write_xlsx(res, "path/DESeq2_results.xlsx")
```

R code 3.16 shows how the design matrix to contrast the samples was created, and how limma inputs the results before they were extracted in a separate document with writexl.

R code 3.16: Limma differential equation

```
temp <- factor(all_patientNo6$Histology)
design <- model.matrix(~0+temp)
fit <- lmFit(df_norm, design)
fit <- ebayes(fit)
contrast <- makeContrasts(C-IM, N-C, N-IM, levels = design)
fit2 <- contrasts.fit(fit, contrasts)
fit2 <- eBayes(fit2)
res <- topTable(fit2, coef = 3, n = Inf)
res <- cbind(rownames(res), res)
colnames(res)[1] <- "genes"
write_xlsx(res, "path/limma_results.xlsx")
```

3.7 Ingenuity pathway analysis (IPA)

An individual identifier was given to each probe in the microarray data through Lumi in R [88, 95] using the lumiHumanIDMapping [96]. The R code 3.17 shows the conversion

of the xlsx file into an txt file. Lumi will only accept txt files. Afterwards lumi interprets and assigns an unique ILMN_ID to each unique Probe_ID. The new ILMN_ID was then bound to the original dataframe.

R code 3.17: Applying ILMN_ID to df through lumi

```
temp_df <- read_xlsx("path.xlsx", skip = 7)
write.table(temp_df, "path.txt")
lumiObject <- lumiR("path.txt", sep = " ",
    lib.mapping = "lumiHumanIDMapping)
lumiObject <- nuID2IlluminaID(lumiObject)
df <- cbind(lumiObject[5], df[-1])
```

A core analysis ran on each dataset with the criteria of adjusted P value < 0.05 and an absolute log2FoldChange of 1. Each dataset was investigated for their individual signaling pathways. A comparison analysis was made between the tissue and the SCT datasets for common signaling pathways and an upstream analysis.

4 Results

4.1 Normalization using housekeeping genes

Gastric housekeeping genes have several probes detecting identical genes with varied detected intensity (Table 4.1). The gene stability value (M) from geNorm showed increased stability in genes with lower expression, while highly expressed genes were more unstable (Figure 4.1). The genes with the lowest M were FBXW2-LUC7L2 with an M of 0,086. RPL29, PMM1, GAPDH, and B2M showed an M ranging from 0,36 to 0,27. There were 18 probes over an M of 0,2 and 18 below with CCNC.2 (ILMN_1676423) and PAPOLA being 0,192 and 0,185 respectively. OAZ1 was the only high intensity gene with a low M of 0,16. Among the 17 remaining probes below 0,2 M did CCNC.2 and PAPOLA show the greatest instability in the graphical plots before and after normalization, respectively (Figure 4.2 and 4.3). For all housekeeping genes in a graphical visualization, both with and without RPL29 that had the highest instability in geNorm, see appendix A.

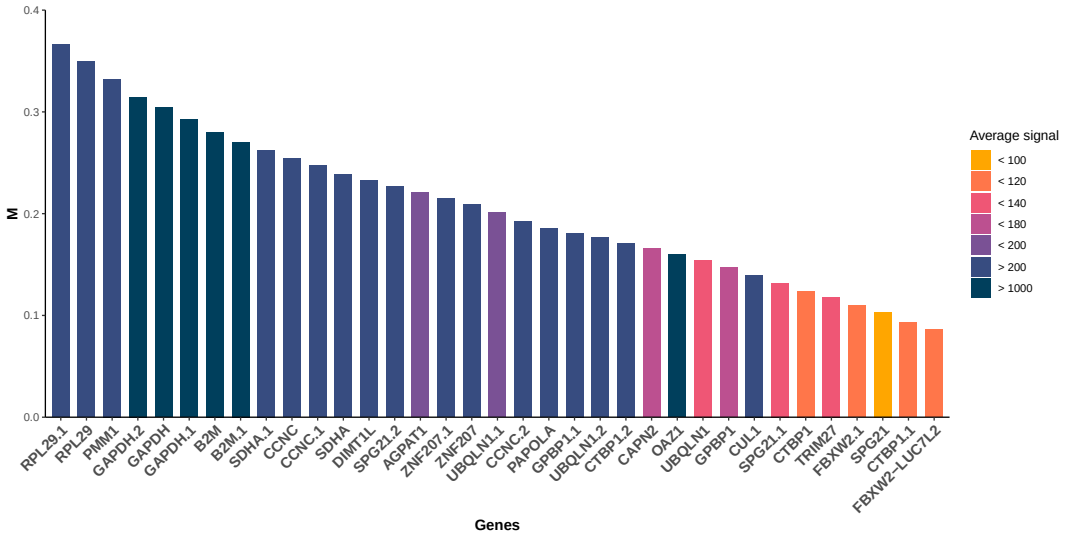


Figure 4.1: The M of the housekeeping genes when compared with each other according to geNorm (ctrlGene_1.0.1) before normalization.

Table 4.1: Corresponding gene and illumina_ID to their average measured intensity. Gene ID is modified with a number from additional gene probes investigated

Gene	Illumina_ID	Intensity	Gene	Illumina_ID	Intensity
AGPAT1	ILMN_1679520	180,1	GPBP1.1	ILMN_1711792	391,0
B2M	ILMN_2148459	7494,8	LUC7L2	ILMN_1747099	114,3
B2M.1	ILMN_1725427	9268,6	OAZ1	ILMN_1773080	6112,2
CAPN2	ILMN_1716057	178,1	PAPOLA	ILMN_1798354	632,4
CCNC	ILMN_1798705	249,0	PMM1	ILMN_1780236	401,2
CCNC.1	ILMN_2409395	318,2	RPL29	ILMN_1737517	245,6
CCNC.2	ILMN_1676423	265,0	RPL29.1	ILMN_1771051	252,4
CTBP1	ILMN_2278235	110,5	SDHA	ILMN_1744210	355,4
CTBP1.1	ILMN_2379734	105,6	SDHA.1	ILMN_2051232	592,8
CTBP1.2	ILMN_1719158	430,0	SPG21	ILMN_1653876	97,4
CUL1	ILMN_1749629	260,0	SPG21.1	ILMN_1733560	125,1
DIMT1L	ILMN_1803312	422,0	SPG21.2	ILMN_1657423	209,5
FBXW2	ILMN_1775753	121,8	TRIM27	ILMN_1655482	123,4
FBXW2.1	ILMN_3251567	109,1	UBQLN1	ILMN_1798380	134,5
GAPDH	ILMN_1802252	1053,8	UBQLN1.1	ILMN_2351611	193,7
GAPDH.1	ILMN_2038778	2629,0	UBQLN1.2	ILMN_1688622	588,3
GAPDH.2	ILMN_1343295	1493,6	ZNF207	ILMN_1670895	475,4
GPBP1	ILMN_2233493	164,3	ZNF207.1	ILMN_1778177	592,3

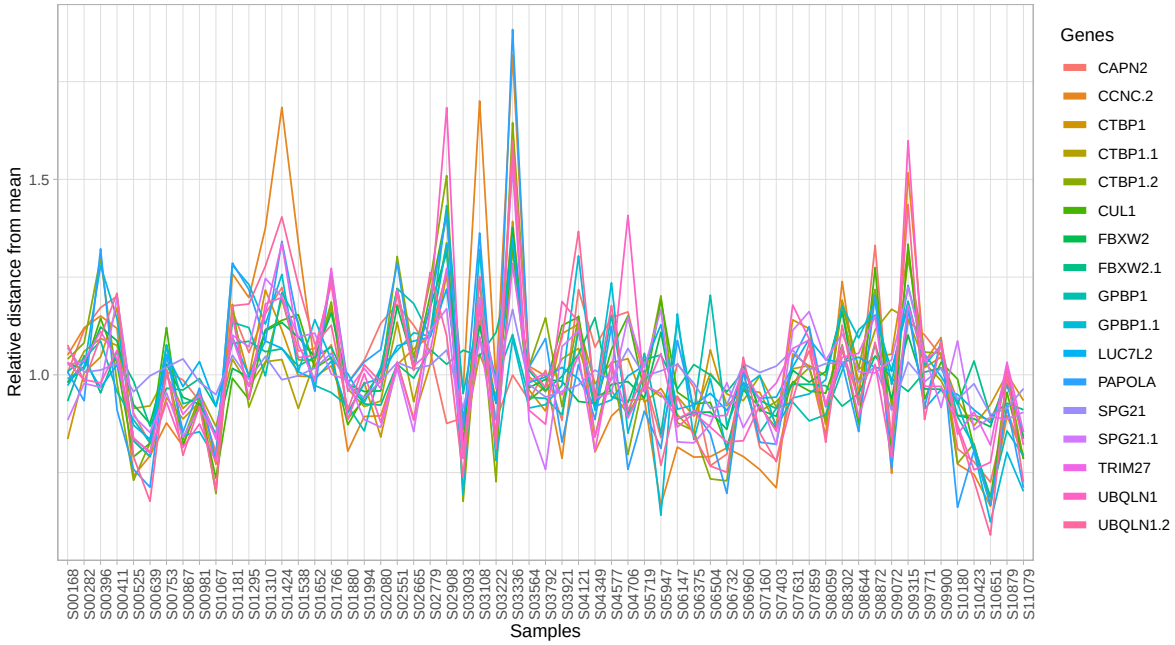


Figure 4.2: Relative variance of housekeeping genes between samples. Containing geNorm results with less than 0.2 M, excluding OAZ1.

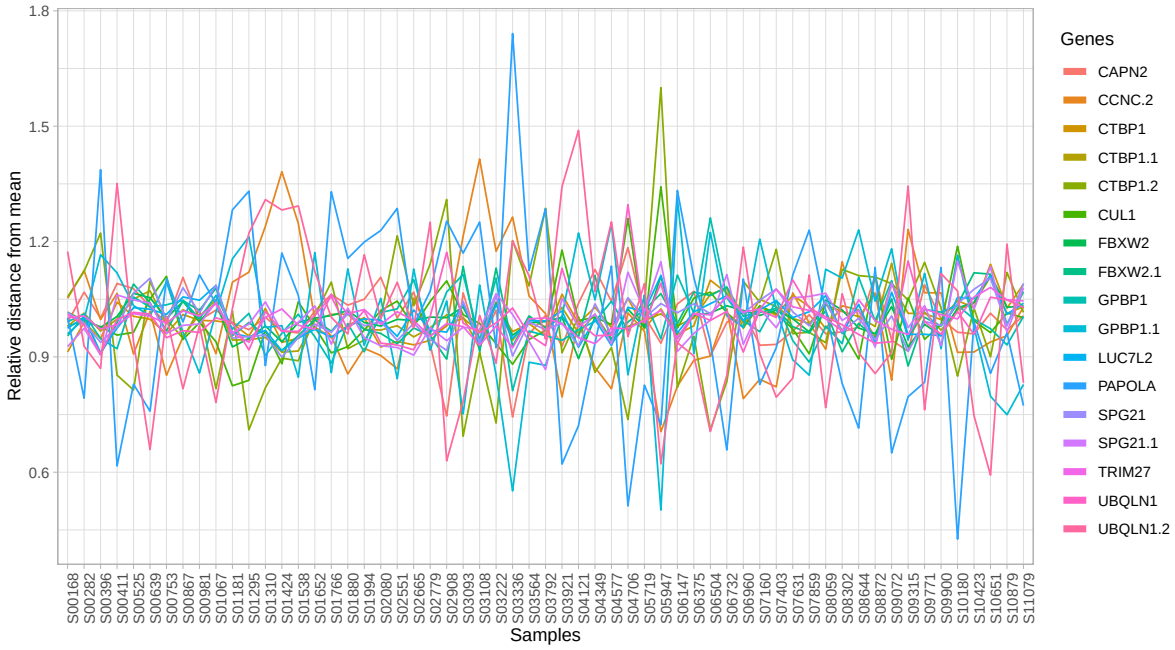


Figure 4.3: Relative variance of housekeeping genes after housekeeping gene normalization between samples. Contaning geNorm results less than 0.2 M, exculding OAZ1.

The SCT went through global normalization with DESeq2, specialized for single cells. The differential expression between the metaplastic and non-atropic gastitis revealed that ACTB, B2M, GAPDH, OAZ1, RPL29, and SDHA were differentially expressed with a log2 fold change of 1 or above. The genes AGPAT1, CAPN2, CCNC, CTBP1, PAPOLA, PMM1, UBQLN1, and ZNF27 were significantly differentially expressed with a log2 fold change between 0,164-0,729. The genes CUL1, FBXW2, GPBP1, SPG21, TRIM27, and HPRT1 were not significantly differentially expressed. Expressions for DIMT1L, LUC72L, and S18 rRNA were not found.

4.2 Datasample analysis with R

4.2.1 SCT filtering

The SCT had 1299 genes after excluding genes present in more than one cell cluster. Comparing the tissue data to the SCT left 1625 genes which were reduced to 1158 after aggregating duplicate genes. The unique genes per cell type ranged from 3 genes for

neck-like cells to 179 for endothelial cells (Table 4.2).

Table 4.2: Number of identifiers corresponding to the SCT.

Cell type	Unique single cell transcriptome genes [34]	IDs after lumi filtering	Genes after removal of duplicates
B cell	69	81	62
Cancer cell	29	38	27
Chief cell	13	20	12
Endothelial cell	197	254	179
Enterocyte	138	180	128
Enteroendocrine	72	75	60
Fibroblast	112	132	100
Antral basal gland mucous cell	12	10	9
Goblet cell	14	14	13
Macrophage	138	178	129
Mast cell	69	78	55
Metaplastic stem-like cell	33	49	31
Neck-like cell	3	5	3
Proliferative cells	50	73	47
Pit Mucous cell	107	146	100
Smooth muscle cell	121	142	102
T cell	122	150	101
Total	1299	1625	1158

4.2.2 Dimensional reduction

Data visualization of MDS plots created through Limma, which underwent log2 transformation (Figure 4.4 B-D, F) displayed a uniform better spread than their non-log2 transformed counterpart (Figure 4.4 A, E). The differences between the log2 transformed graphs are minor (Figure 4.4 B-D, F). While the log2 transformed MDS plots exhibit some differentiation, they were unable to properly differentiate between cancer, gastric intestinal metaplasia and normal samples.

tSNE and UMAP plots, using background correction and normalization from the limma package, showed better separation than the MDS plots when using all genes. Some separation was detected but no clear clustering of based on histology was observed (Figure 4.5, A). The enterocyte genes gave two distinct clusters, however, both clusters contained all histologies (Figure 4.5, B). The goblet cell genes showed similar clustering as the enterocyte genes but to a lesser extent (Figure 4.5, C). The chief and cancer cell

genes showed similar clustering potential as using all genes, with chief cells showing a little bit better separation (Figure 4.5, A, D-E). All subfigures showed some clustering of antrum, cardia and corpus ventriculi samples (Figure 4.5). For different data preparation methods, see appendix B-F.

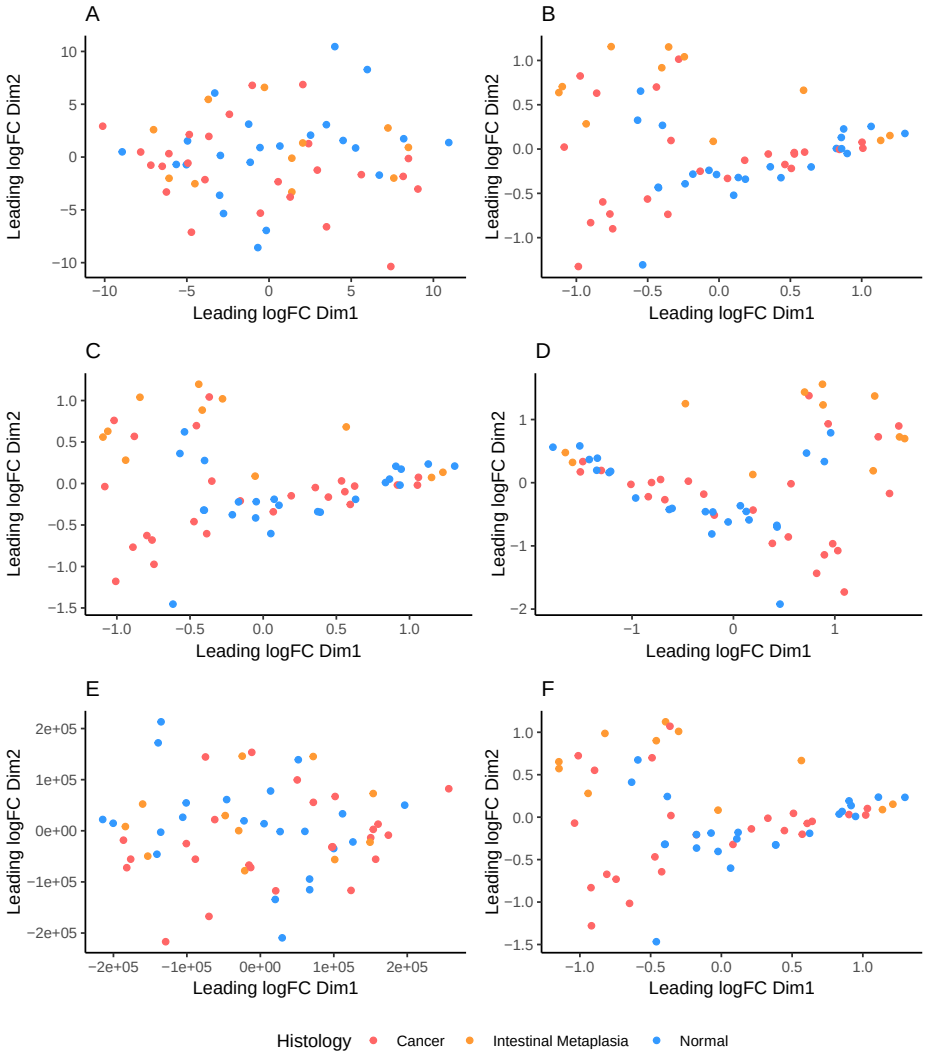


Figure 4.4: MDS plot of the patient samples. A) The data was normalized based on housekeeping genes. B) The data was normalized based on housekeeping genes and log2 transformed. C) The data was quantile normalized and log2 transformed through limma. D) The data was normexp background corrected, quantile normalized and log2 transformed through limma. E) The data was normalized with geometric mean through DESeq2. F) The data was normalized with geometric mean through DESeq2 and log2 transformed.

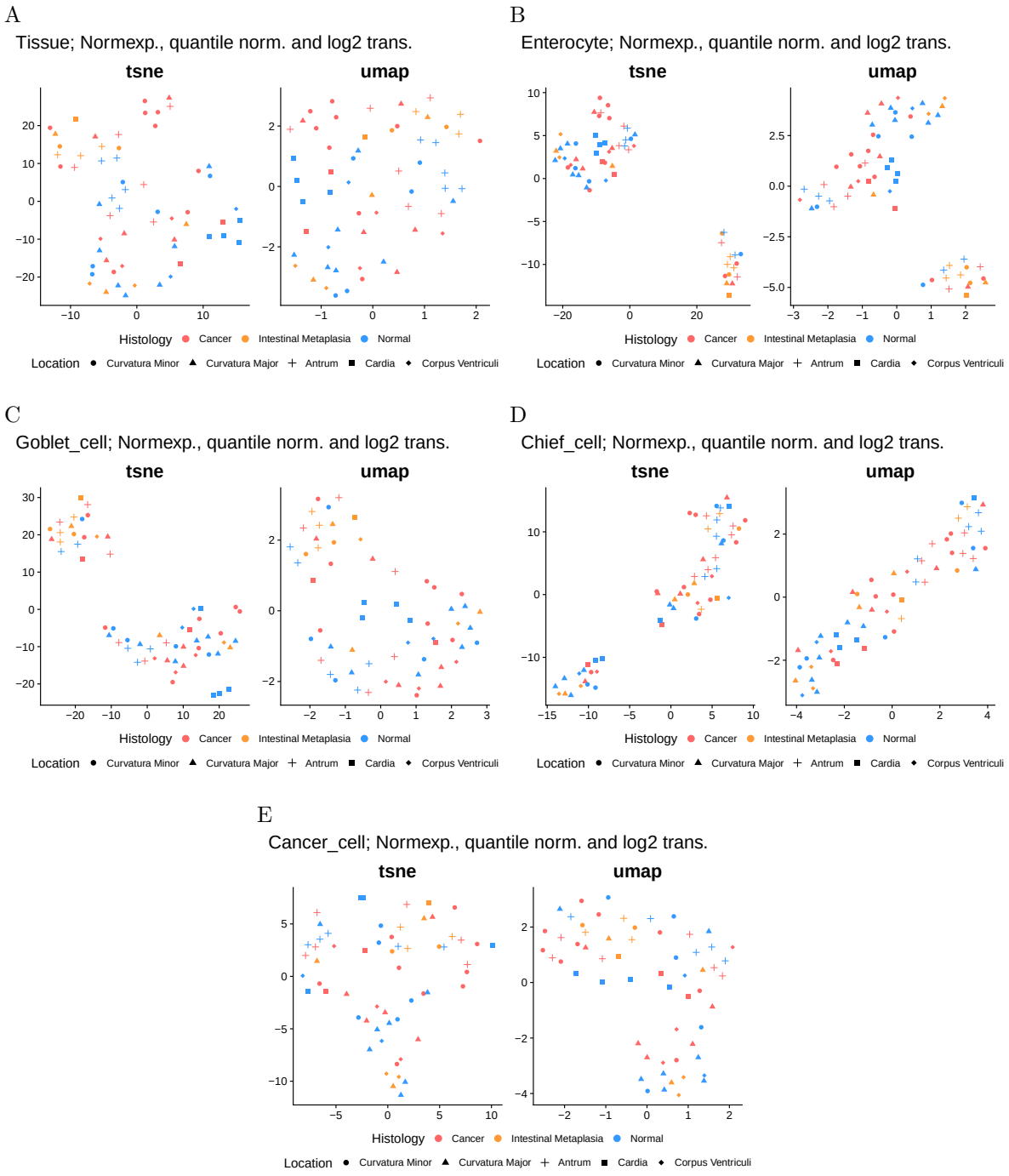


Figure 4.5: Dimensional reduction with tsne and umap using Seurat. Genes corresponding to each cell type were extrapolated from the SCT from table 4.2.

4.2.3 Cellular heatmap

The heatmaps from the tissue sample data processed through limma were manually clustered by rows using cell types and separated by columns using either histology (Figure 4.6) or patient number (Figure 4.7). The remaining clustering was done computationally. The sample location and GHAI score annotation were not clustered. For sample classification, see appendix G.

All histological states contain samples whose genetic expression differ more within their states than outside . The normal tissue was clustered in the middle of the two pathological states, indicating that the cancer and metaplastic tissue was more similar to the normal tissue than each other. The normal samples were separated into three groups. Group no. 1 had an upregulated expression of pit mucous and proliferative cell related genes. Group no. 2 had some genes related to enterocytes and all chief cells upregulated. Group no. 3 had no single distinct group of upregulated genes. The GIM samples were separated into two distinct groups and an outlier more similar to the normal broad normal sample category. Group no. 1 had almost all enterocyte and goblet cell related genes upregulated. Group no. 2 had a similar expression to the normal sample's group no. 2 where some enterocyte and all chief cell related genes were upregulated. The GC samples were a lot more varied, with certain samples showing more similar expression to the GIM samples. Cancer-related genes visibly upregulated in 8 samples coincided with 7 out of 12 pathologically classified cancers and 7 of 22 biopsy-classified cancers, see appendix G. Cancer-related genes were also upregulated in 3 normal samples and 2 metaplasia samples. Several other samples of all histological classifications had an upregulation of cancer-related genes, but this upregulation was confined to a subset of the cancer-related and was slightly visually distinct. GC samples with upregulated cancer-related genes showed a general upregulation in neck-like cells, proliferative cells, and macrophages. Non-cancer samples did not follow this general upregulation, except for proliferative cells. Some GC samples showed similar expression to group no. 2 in the normal and GIM samples with chief cell related genes being upregulated. All histological classifications showed some tendency toward clustering based on sample site with

more pathogenic tissue showing less clustering. The GHAI score showed some tendency toward clustering. The sample sites with the most clustering were the antrum followed by curvitura major and cardia (Figure 4.6).

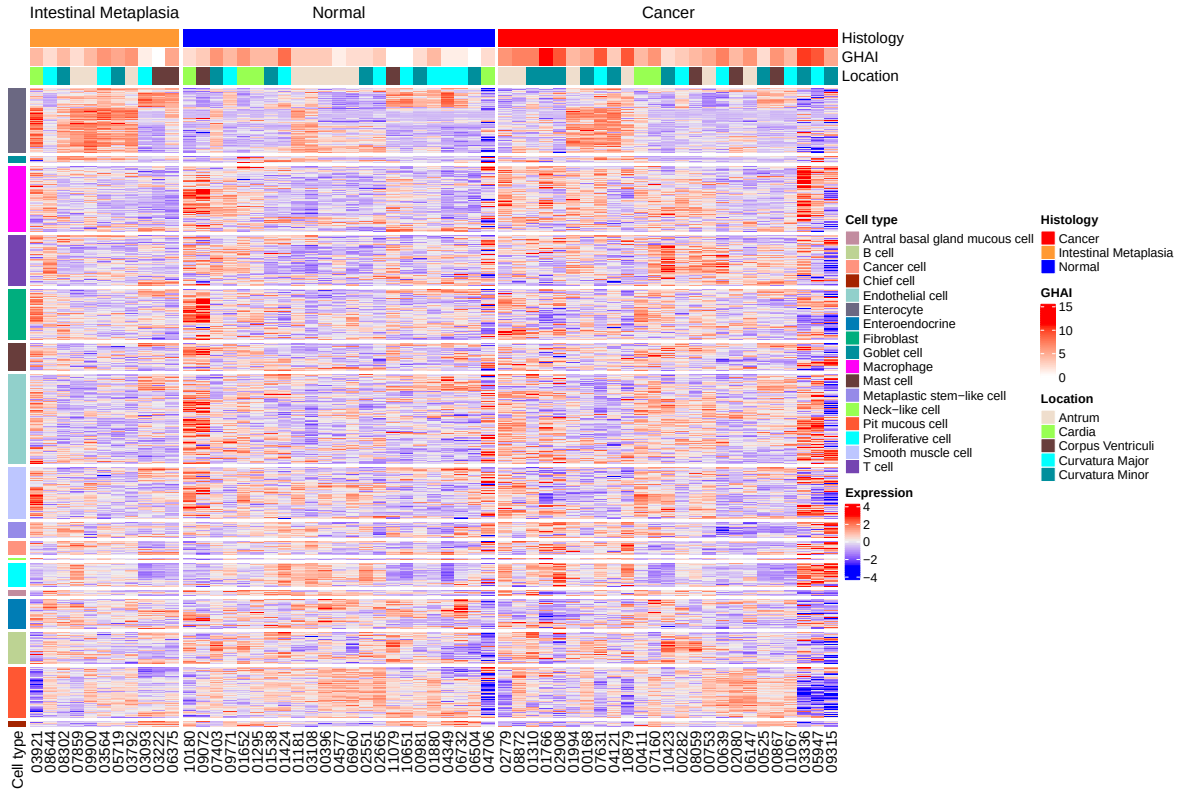


Figure 4.6: Heatmap of the tissue data using the SCT to separate the row by cell types. The top annotations GHAI, location, and histology were made using appendix G.

Patient-based clustering revealed many similarities within each patient (Figure 4.7). Intestinal metaplasia is present within 6 patients with intestinal metaplasia clustered together as 2 (patient no. 8 and 11) or 4 (patient no. 9, 13, 14, and 15). Patient no. 2, which had all 4 samples classified as cancer through biopsy and one through pathology, showed no increased expression in cancer-related genes compared to the normal samples. For heatmaps normalized with housekeeping genes, see appendix H.

The aggregated data from the SCT shows a separation of all pathological states without the use of manual clustering. The wild intestinal metaplasia (IMW) showed an increased expression in several cell types, but it differed between the two samples. The severe

intestinal metaplasia (IMS) showed a clear increase in enterocyte and goblet cell related genes. The enterocyte and goblet cells were not among the increased cell-related genes in IMW. Chronic atrophic gastritis (CAG) separated IMW and IMS, showing a more similar expression to IMW and NAG. The wild superficial gastritis (NAG) samples used as controls showed an increase in either pit mucous cell or enteroendocrine cell related genes (Figure 4.8).

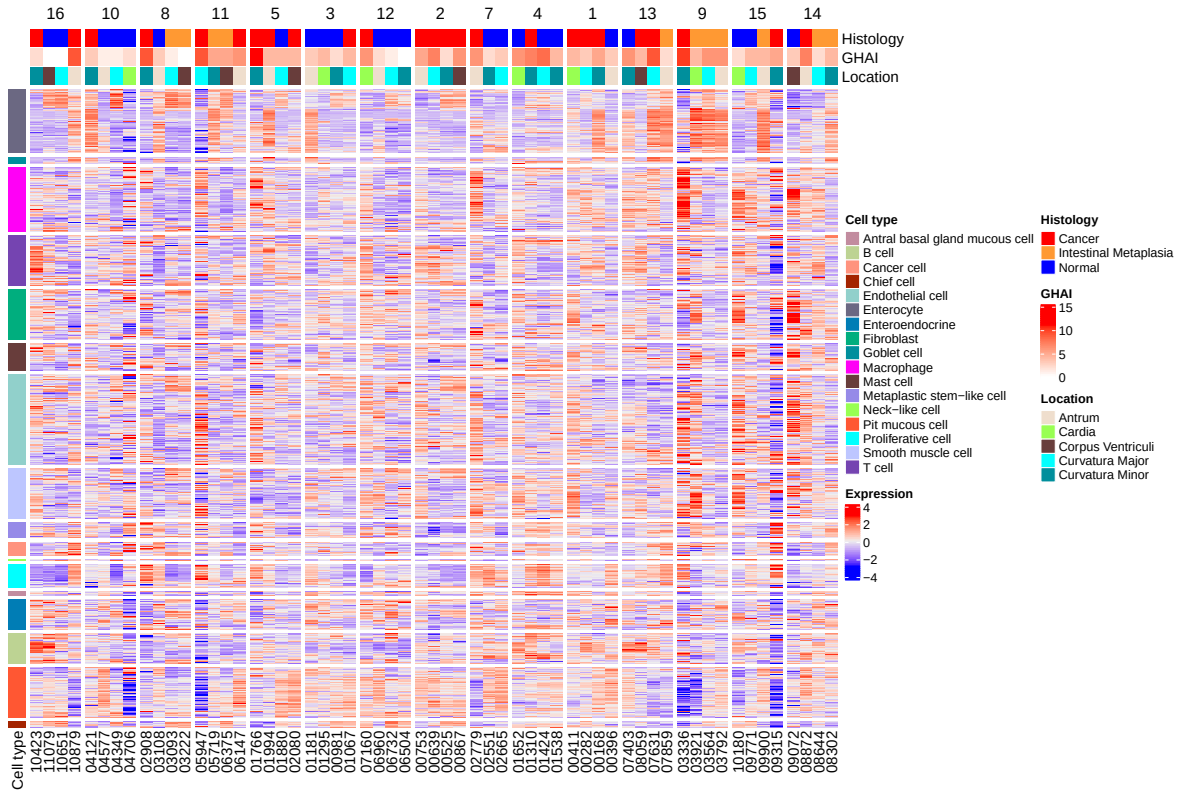


Figure 4.7: Heatmap of the tissue data using the SCT to separate the row by cell types. The top annotations GHAI, location, histology, and patient number were made using appendix G.

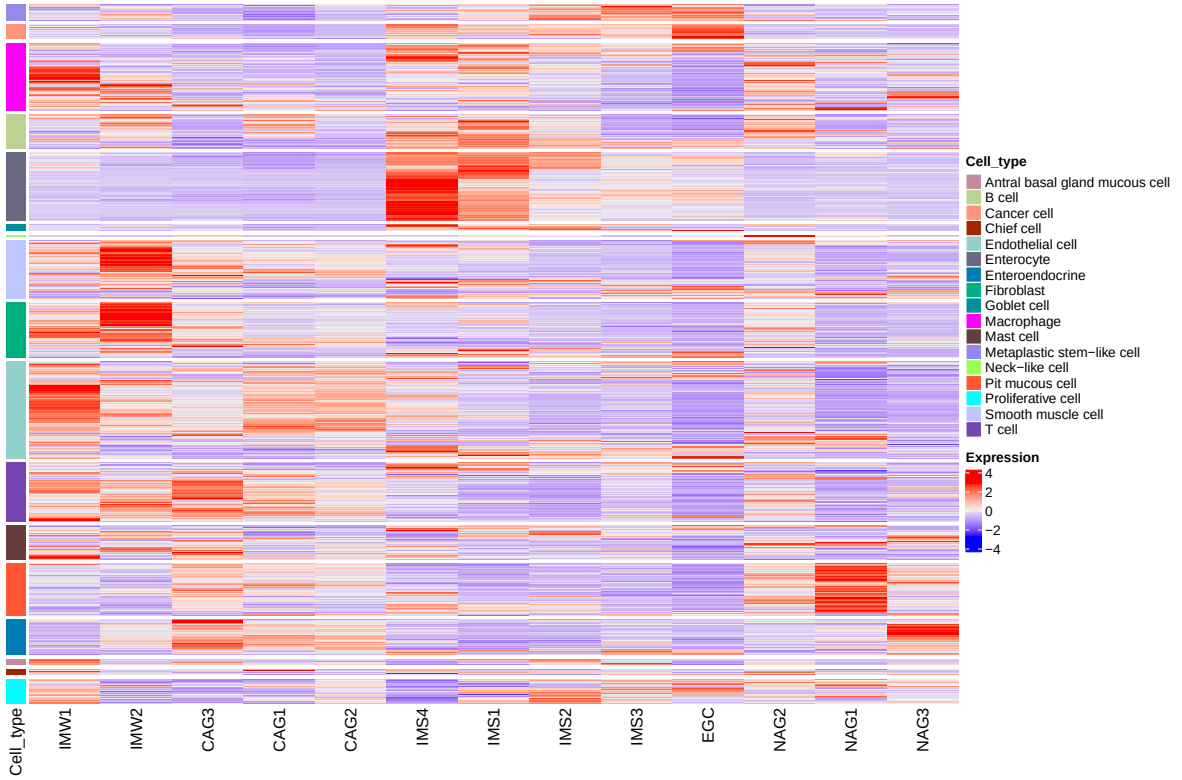


Figure 4.8: Heatmap of the SCT RNA-seq data from [34], aggregated and plotted for visual purposes.

4.3 Ingenuity pathway analysis

4.3.1 Significant Pathways

The tissue data comparing GIM to normal tissue samples was interpreted and made into differential expressions through 5 different methods. The significant value for the signaling pathways was a p-value below 0.05 and an absolute z-score above 2. DESeq2 normalized data calculated for differential expressed genes through limma resulted in 5 significant canonical pathways. Housekeeping gene normalized data calculated for differential expressed genes through limma gave 9 significant canonical pathways. Limma normalization without background correction and differential expression calculation resulted in 10 significant canonical pathways akin to DESeq2 normalization and differential expression calculation. The dataset LimmaFC was background corrected, normalized, and calculated through Limma with 19 significant canonical pathways. The three

common significantly activated and expressed signaling pathways were the Xenobiotix metabolism PXR signaling pathway, Xenobiotix metabolism CAR signaling pathway and the LXR/RXR activation. 4 out of 5 methods had serotonin degradation, superpathway of melatonin degradation and xenobiotic metabolism general signaling pathway significantly expressed and activated (Figure 4.9).

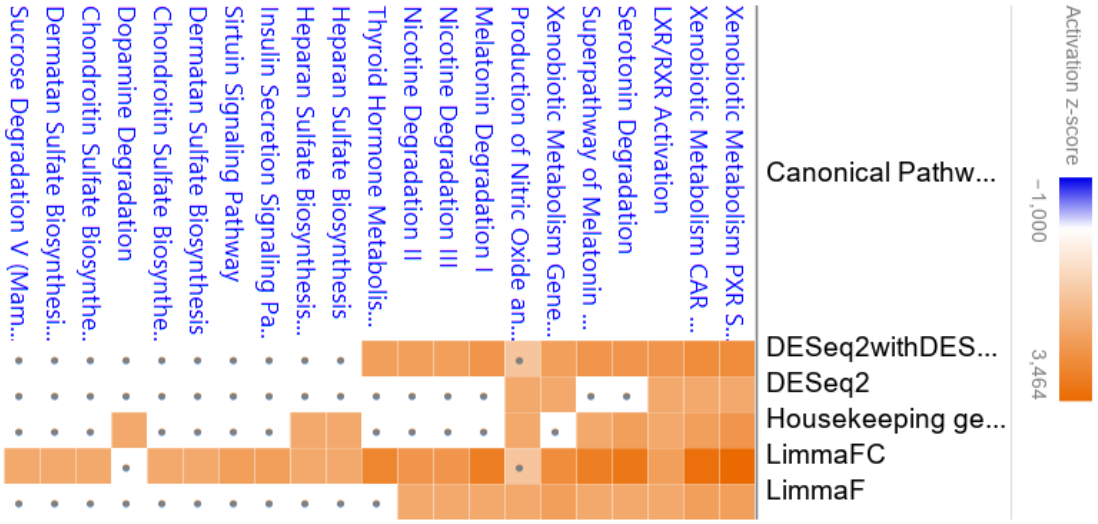


Figure 4.9: Top pathways for all normalization methods with DEGs calculation of metaplasia vs normal tissue. Orange and blue means activation or inhibition respectively. A dot indicates a Z-score below 2. Pathways with a Z-score below 2 were removed.

The top pathways from the tissue data (LimmaFC) were serotonin degradation, thyroid hormone Metabolism II, melatonin degradation, superpathway of melatonin degradation, and FXR/RXR activation, sorted after p-value (Table 4.3). Serotonin degradation consist of 5 enzymes responsible for breaking down serotonin. The tissue dataset upregulated three enzymes and estimated an activation of the pathway (Figure 4.10). Thyroid hormone metabolism II consisted of 3 enzymes where 2 of them are upregulated, predicting an upregulation. Both of the upregulated enzymes were identical to serotonin degradation (Figure 4.11). Melatonin degradation consisted of 3 enzymes where all of them contained upregulated genes. Two of the enzymes overlapped with serotonin degradation (Figure 4.12). Superpathway of melatonin degradation added two additional pathways of melatonin degradation but they were not activated (Figure 4.13). FXR/RXR activation consisted of separate cascades that jointly predicted an increased

cholesterol activity, the total prediction were neither up- nor downregulated (Figure 4.14).

The top pathways of the SCT dataset were oxidative phosphorylation, mitochondrial dysfunction, sirtuin signaling pathway, EIF2 signaling, and glucocorticoid receptor signaling sorted after p-value (Table 4.4). Oxidative phosphorylation (OXPHOS) consisted of 5 complexes and were predicted upregulated with all complexes containing upregulated genes (Figure 4.15). Mitochondrial dysfunction encompasses OXPHOS but was not predicted up- or downregulated (Figure 4.16). Sirtuin signaling pathway was predicted downregulated, but the SIRT genes were not predicted (Figure 4.17). EIF2 signaling was predicted upregulated leading to effects such as increased endoplasmic reticulum (ER) stress, amino-acid transport/biosynthesis, and vascularization (Figure 4.18). Glucocorticoid receptor signaling consisted of separate cascades that jointly predicted a decrease in an inflammatory response, the total prediction were neither an up- nor downregulation (Figure 4.19).

Table 4.3: Top 5 canonical pathways from the LimmaFC tissue data sorted by P-value. An absolute Z-score of two indicates significant significant activation or inhibition.

Top Pathways	P-value	Z-score	Overlapp
Serotonin Degradation	5,76E-12	3,162	10/57 (17,5%)
Thyroid Hormone Metabolism II	5,03E-11	2,828	8/33 (24,2%)
Melatonin degradation	1,1E-10	3,000	9/54 (16,7%)
Superpathway of Melatonin Degradation	2,56E-10	3,000	9/59 (15,3%)
FXR/RXR Activation	1,07E-9	NaN	11/125 (8,8%)

Table 4.4: Top 5 canonical pathways from the SCT sorted by P-value. An absolute Z-score of two indicates significant significant activation or inhibition.

Top Pathways	P-value	Z-score	Overlapp
Oxidative Phosphorylation	1,52E-84	8,485	72/111 (64,9%)
Mitochondrial Dysfunction	2,35E-77	NaN	81/165 (49,1%)
Sirtuin Signaling pathway	7,92E-44	-2,949	69/292 (23,6%)
EIF2 Signaling	2,47E-39	5,014	58/224 (25,9%)
Glucocorticoid Receptor Signaling	1,85E-26	NaN	73/540 (13,5%)

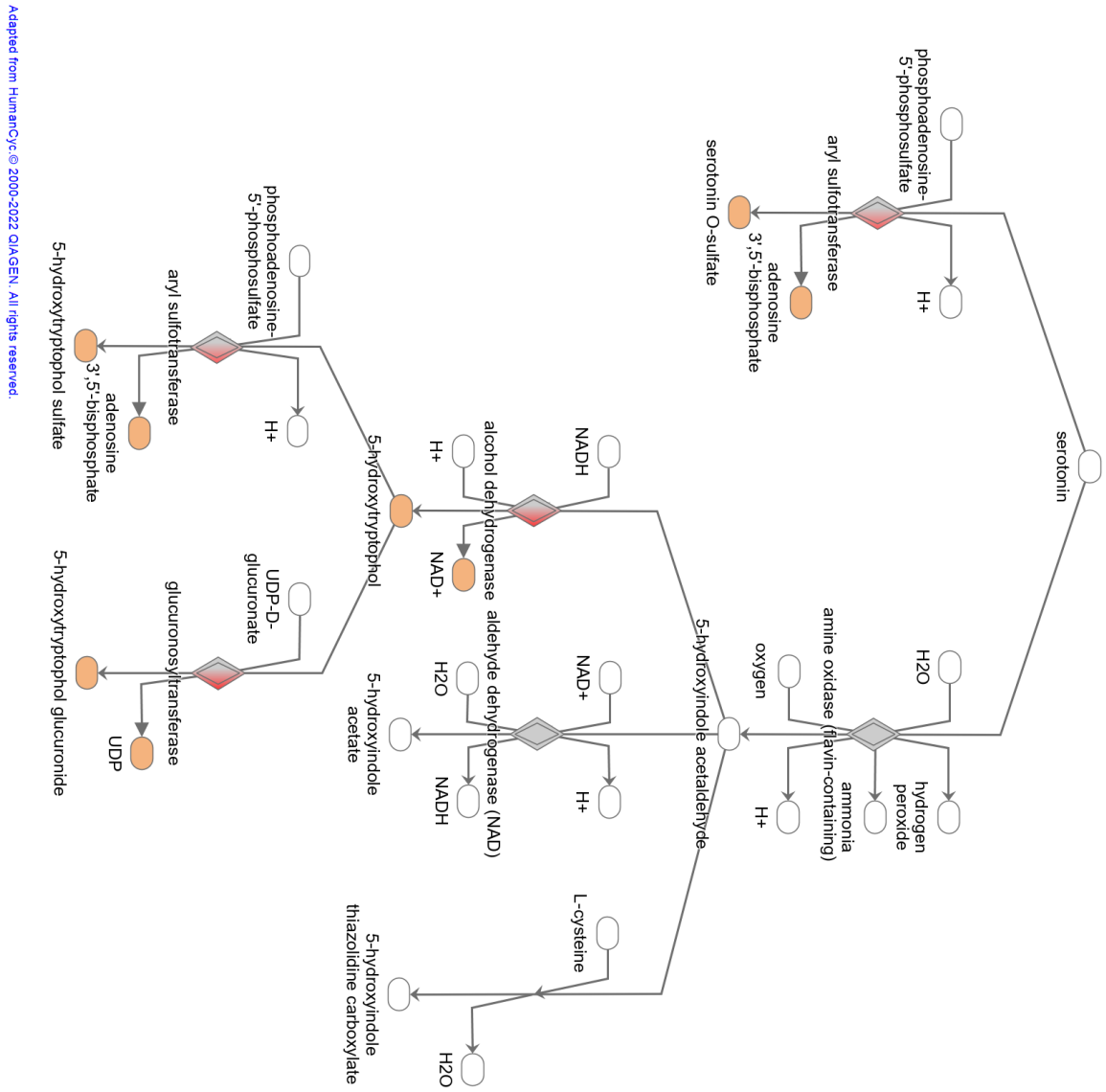


Figure 4.10: Serotonin degradation pathway overlayed with the LimmaFC gene expression with the Molecule Activity Predictor (MAP) activated. Red molecules are known as upregulated genes from the dataset. Orange components are predicted upregulated components based on the known expression. Gray components are present, but insignificantly differentially expressed, and excludes a prediction. White components not present and exclude a prediction. Mixed colors indicate a complex with components containing differing information.

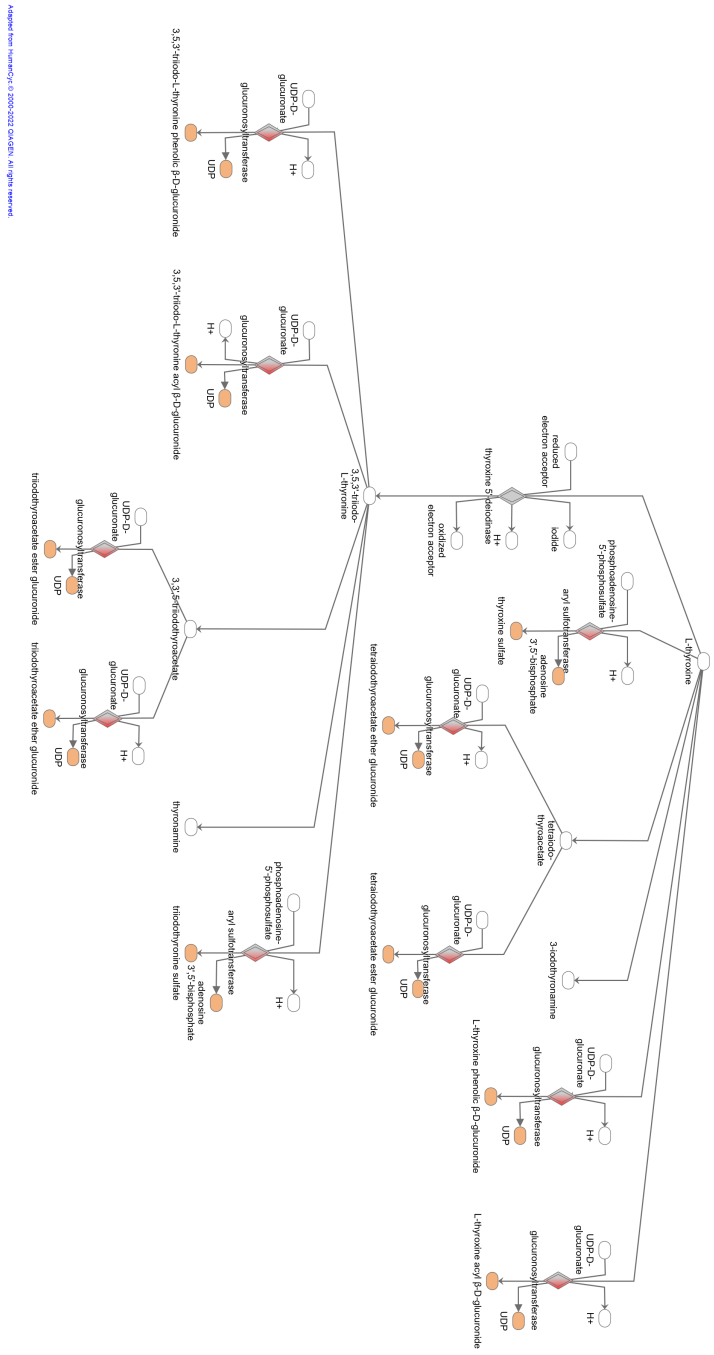


Figure 4.11: Thyroid hormone metabolism II overlayed with the LimmaFC gene expression with the Molecule Activity Predictor (MAP) activated. Red molecules are known as upregulated genes from the dataset. Orange components are predicted upregulated components based on the known expression. Gray components are present, but insignificantly differentially expressed, and excludes a prediction. White components not present and exclude a prediction. Mixed colors indicate a complex with components containing differing information.

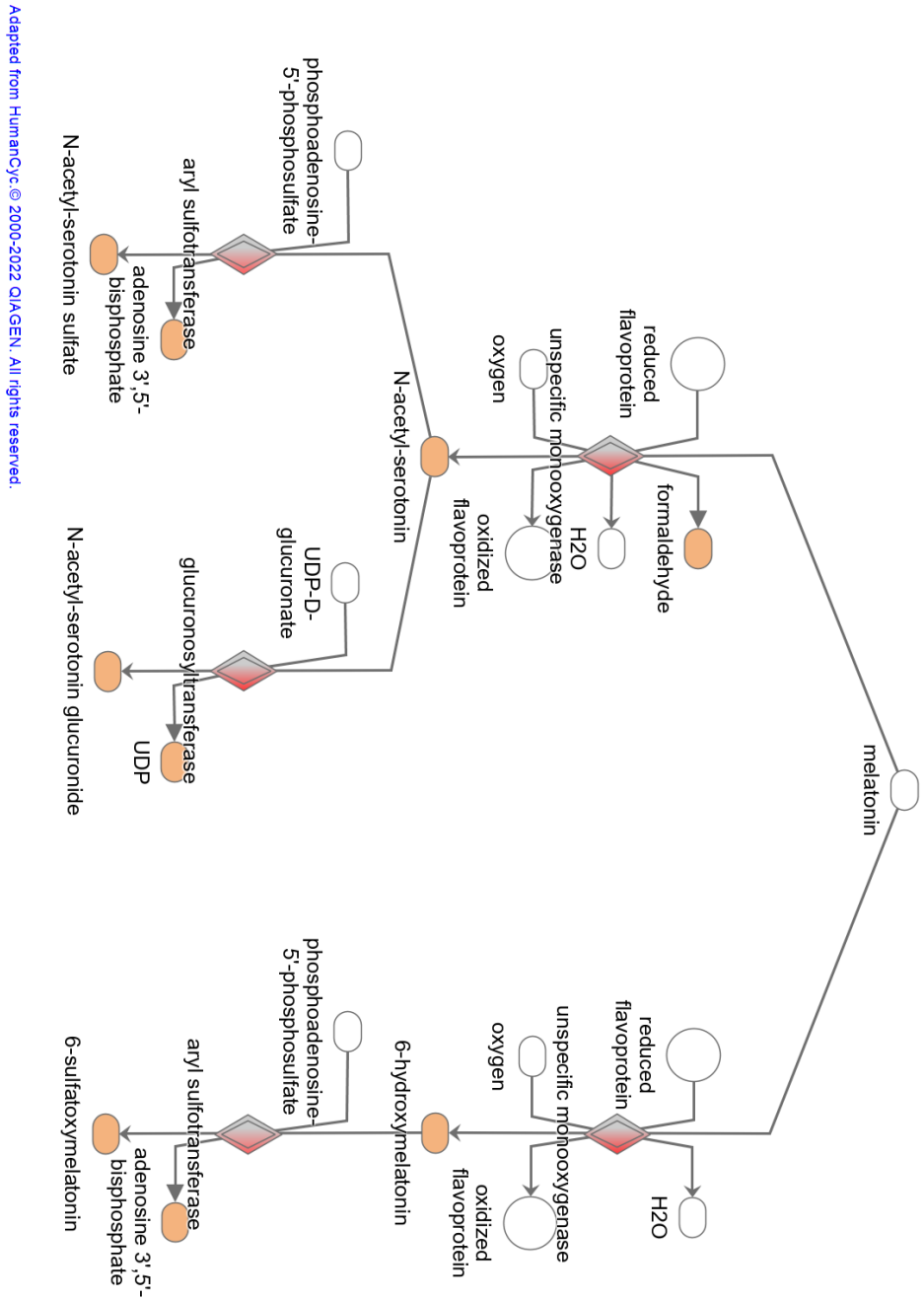


Figure 4.12: Melatonin degradation overlayed with the LimmaFC gene expression with the Molecule Activity Predictor (MAP) activated. Red molecules are known as upregulated genes from the dataset. Orange components are predicted upregulated components based on the known expression. Gray components are present, but insignificantly differentially expressed, and excludes a prediction. White components not present and exclude a prediction. Mixed colors indicate a complex with components containing differing information.

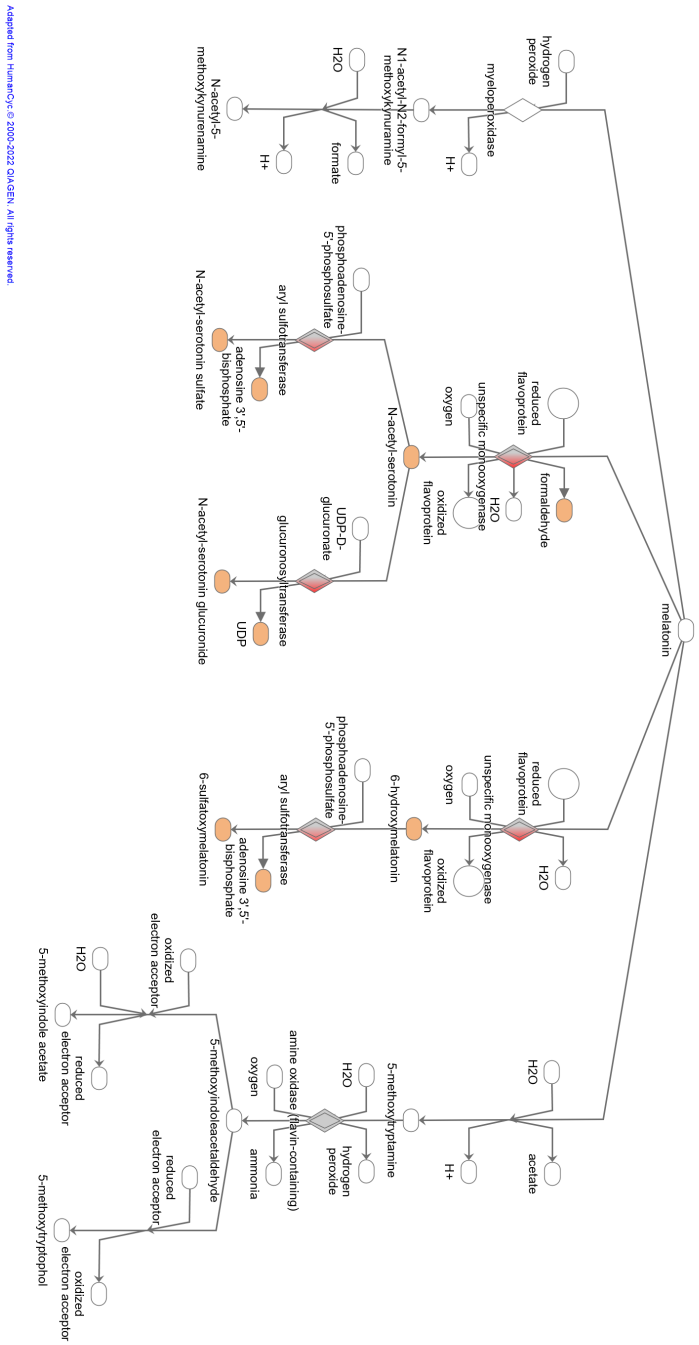


Figure 4.13: Superpathway of melatonin degradation overlayed with the LimmaFC gene expression with the Molecule Activity Predictor (MAP) activated. Red molecules are known as upregulated genes from the dataset. Orange components are predicted upregulated components based on the known expression. Gray components are present, but insignificantly differentially expressed, and excludes a prediction. White components not present and exclude a prediction. Mixed colors indicate a complex with components containing differing information.

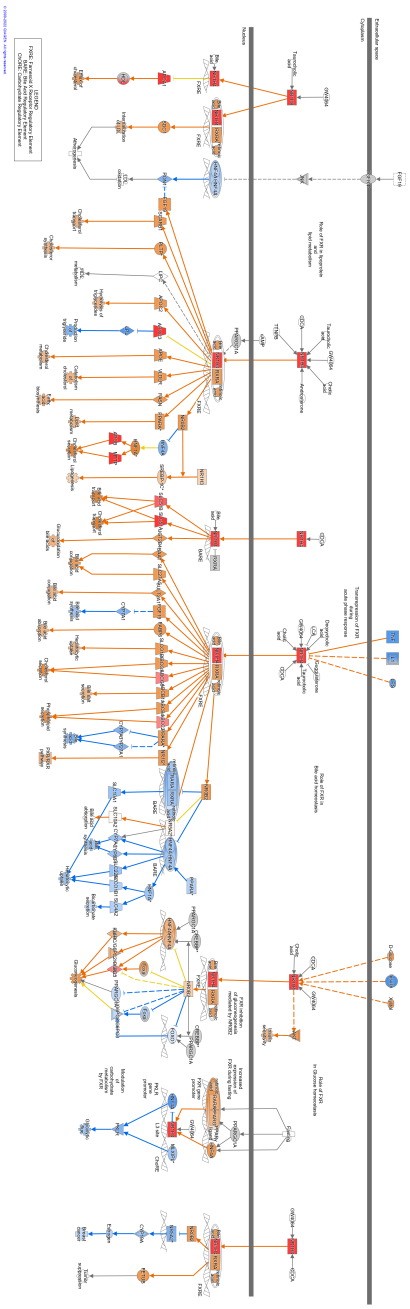


Figure 4.14: FXR/RXR activation overlayed with the LimmaFC gene expression with the Molecule Activity Predictor (MAP) activated. Red molecules are known as upregulated genes from the dataset. Orange components are predicted upregulated components based on the known expression. Blue components are predicted inhibited based on the known expression. Gray components are present, but insignificantly differentially expressed, and excludes a prediction. White components not present and exclude a prediction. Mixed colors indicate a complex with components containing differing information.

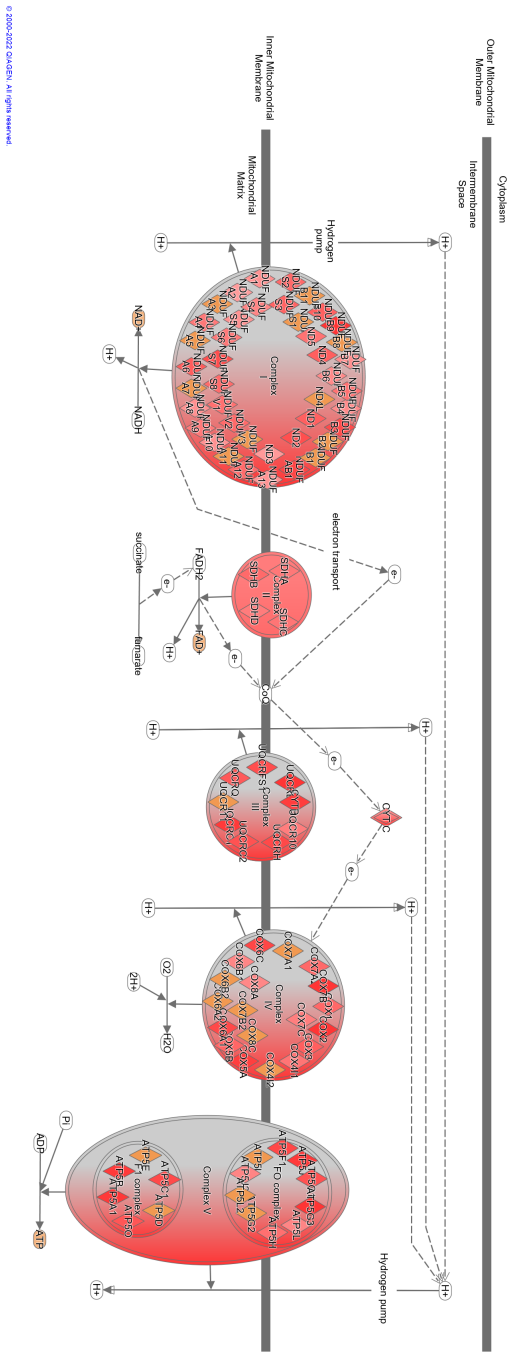


Figure 4.15: Oxidative phosphorylation pathway overlayed with the SCT gene expression with the Molecule Activity Predictor (MAP) activated. Red molecules are known as upregulated genes from the dataset. Orange components are predicted upregulated components based on the known expression. Gray components are present, but insignificantly differentially expressed, and excludes a prediction. White components not present and exclude a prediction.

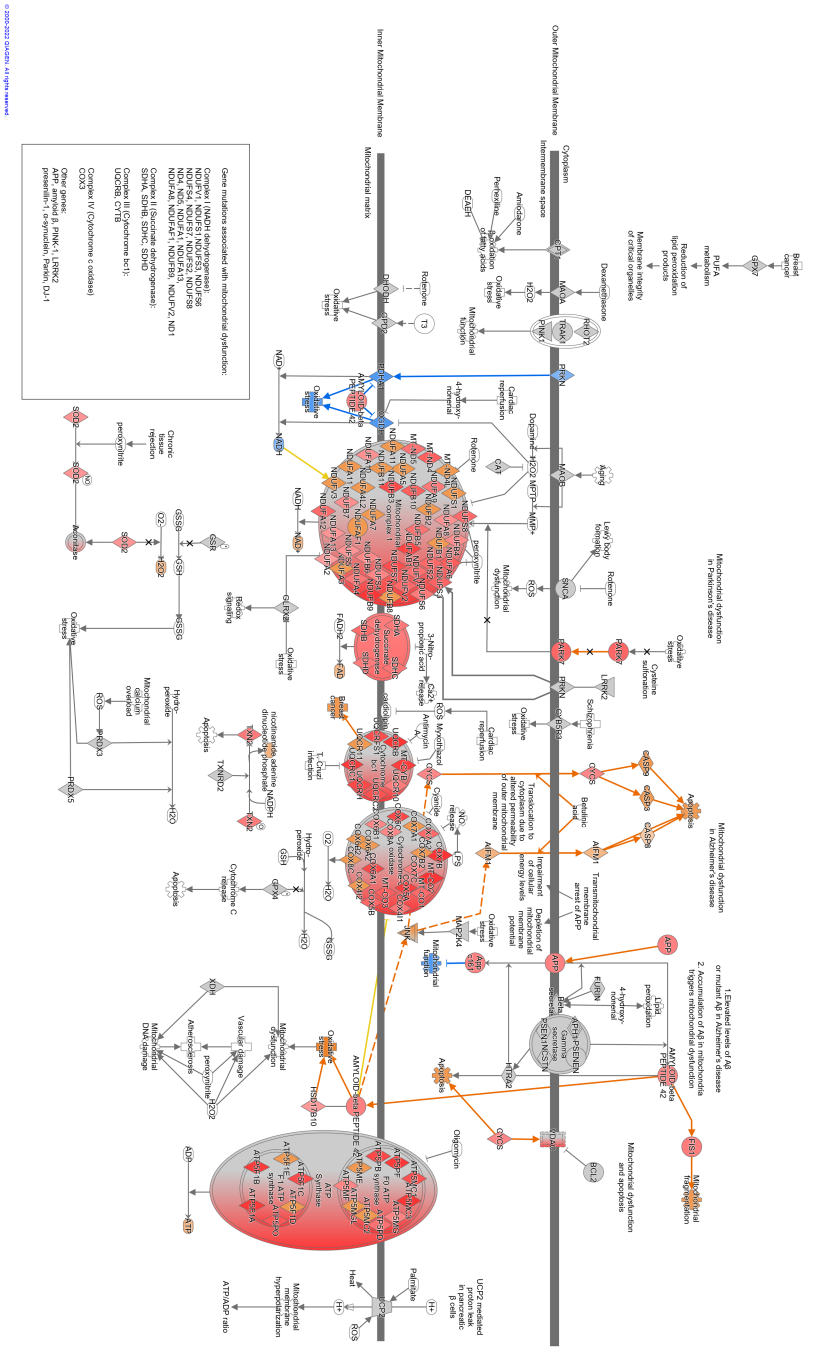


Figure 4.16: Mitochondrial dysfunction overlayed with the SCT gene expression with the Molecule Activity Predictor (MAP) activated. Red molecules are known as upregulated genes from the dataset. Orange components are predicted upregulated components based on the known expression. Blue components are predicted inhibited based on the known expression. Gray components are present, but insignificantly differentially expressed, and excludes a prediction. White components not present and exclude a prediction.

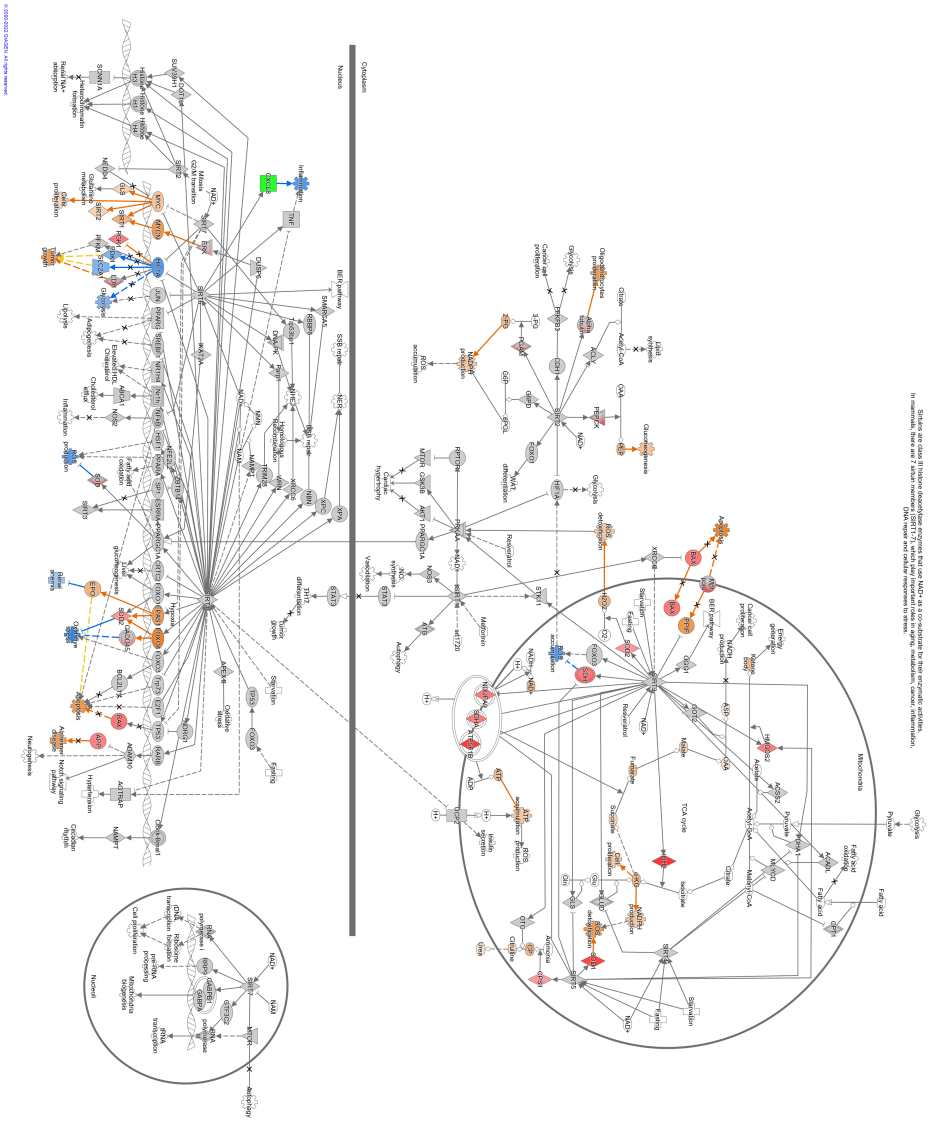


Figure 4.17: Sirtuin signaling pathway overlayed with the SCT gene expression with the Molecule Activity Predictor (MAP) activated. Red molecules are known as up-regulated genes from the dataset. Orange components are predicted upregulated components based on the known expression. Green molecules are known as down-regulated genes from the dataset. Blue components are predicted inhibited based on the known expression. Gray components are present, but insignificantly differentially expressed, and excludes a prediction. White components not present and exclude a prediction.

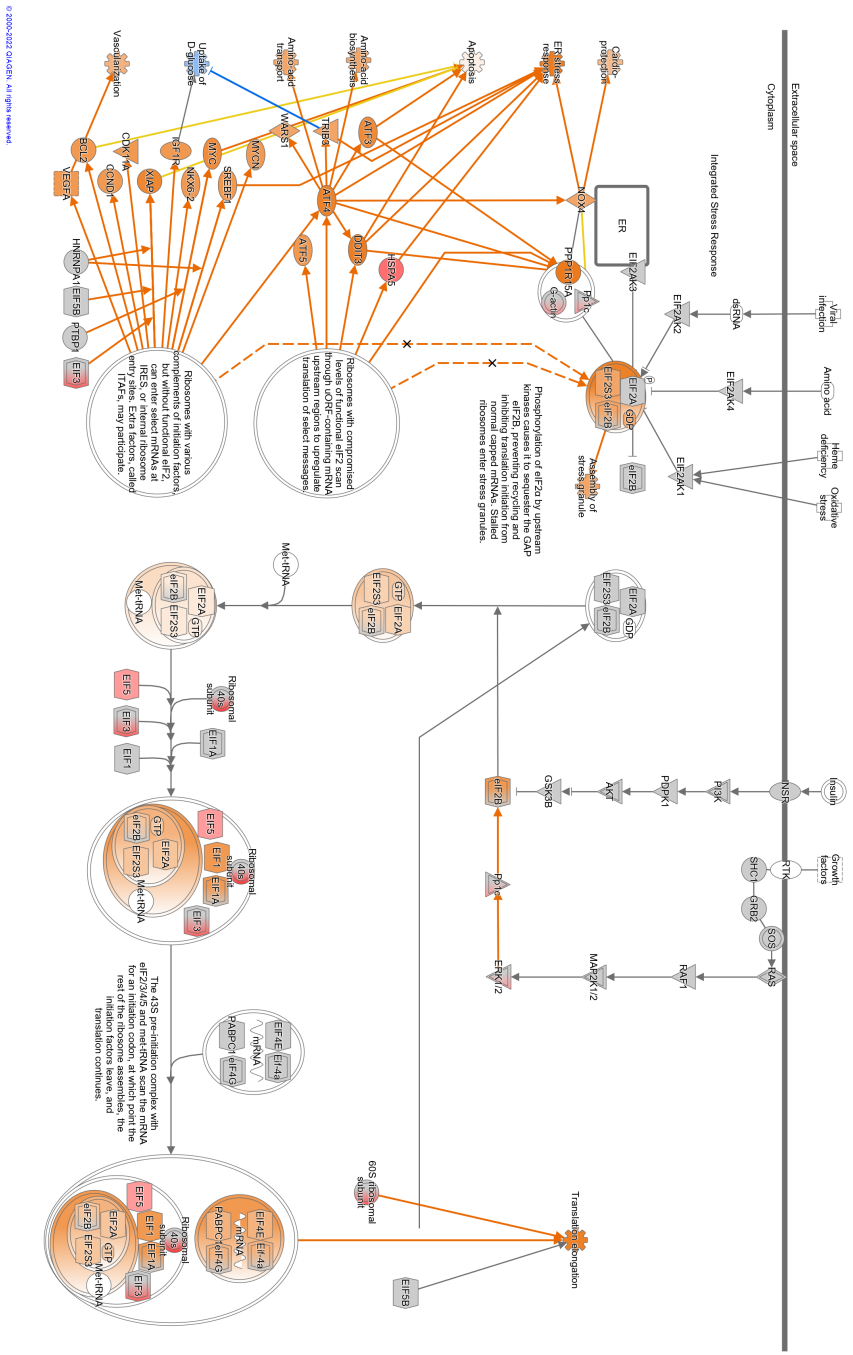


Figure 4.18: EIF2 signaling overlayed with the SCT gene expression with the Molecule Activity Predictor (MAP) activated. Red molecules are known as upregulated genes from the dataset. Orange components are predicted upregulated components based on the known expression. Blue components are predicted inhibited based on the known expression. Gray components are present, but insignificantly differentially expressed, and excludes a prediction. White components not present and exclude a prediction.

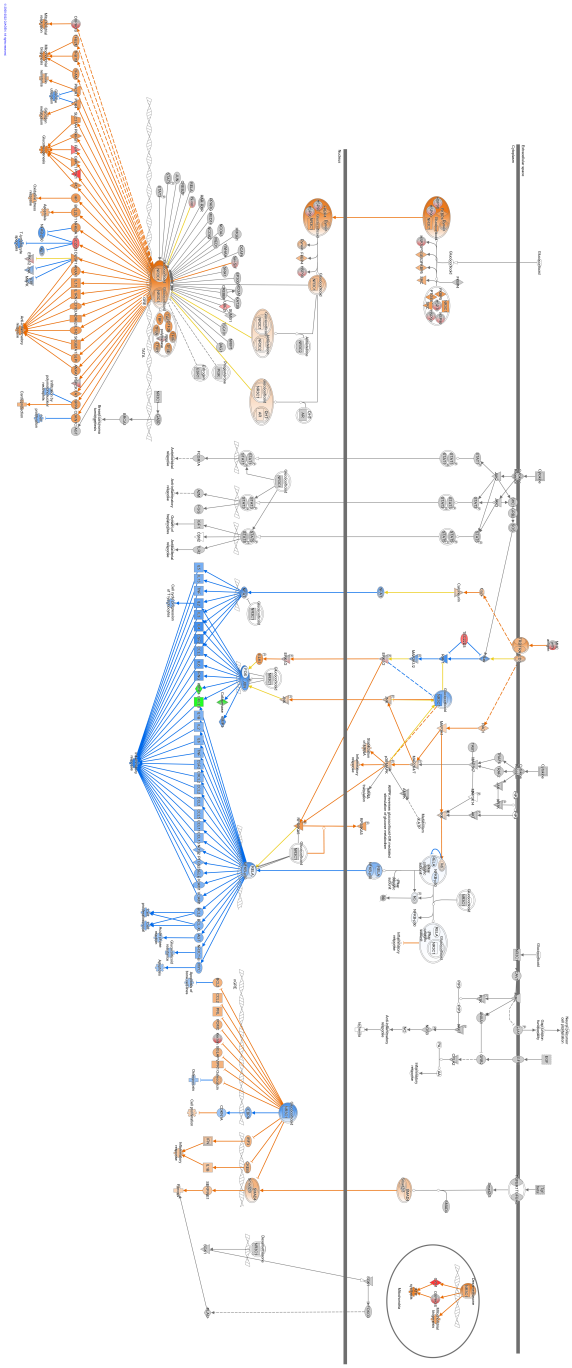


Figure 4.19: Glucocorticoid receptor signaling overlayed with the SCT gene expression with the Molecule Activity Predictor (MAP) activated. Red molecules are known as upregulated genes from the dataset. Orange components are predicted upregulated components based on the known expression. Green molecules are known as down-regulated genes from the dataset. Blue components are predicted inhibited based on the known expression. Gray components are present, but insignificantly differentially expressed, and excludes a prediction. White components not present and exclude a prediction.

The two datasets shared 4 pathways of significance in terms of p-value and z-score. The xenobiotic metabolism CAR signaling pathway was activated in both datasets. The xenobiotic metabolism PXR signaling pathway fell short in terms of z-score in the SCT data, but was included due to its proximity. Sucrose degradation V was activated in both datasets, but only in the LimmaFC dataset for the tissue data that was used for the IPA prediction. The sirtuin signaling pathway showed contradictory results from the two datasets. The tissue data and the SCT had 5 and 69 matching molecules respectively. The xenobiotic metabolism CAR signaling pathway, xenobiotic metabolism PXR signaling pathway, sucrose degradation V, and sirtuin signaling pathway were ranked 7, 6, 8, and 50, respectively of 52 pathways in the tissue data and 45, 48, 20, and 3, respectively of 98 pathways in the SCT (Table 4.5). The pathways were only filtered and sorted based on p-values.

Table 4.5: Common canonical pathways between the tissue data and the SCT. An absolute Z-score of two indicates significant activation or inhibition.

Top Pathways	Z-score (Tissue)	Z-score (SCT)	Predicted activation
Xenobiotic Metabolism CAR Signaling Pathway	3,317	2,496	Activated
Xenobiotic Metabolism PXR Signaling Pathway	3,464	1,941	(Mixed)
Sucrose Degradation V (Mammalian)	2	2,236	Activated
Sirtuin signaling pathway	2,236	-2,949	Contradictory

The xenobiotic metabolism CAR and PXR signaling pathways have an identical outcome but differ in the cascade leading up to the biological functions metabolism of xenobiotics, drug transport and inflammatory response. The cascades ended in cell survival after metabolism of xenobiotics and drug transport. The cell survival was unpredicted in all datasets, but IPA included no associative relationship between cell survival and the upstream cascade (Figure 4.20-4.23). The xenobiotic metabolism CAR signaling pathway encompassed 172 genes and contained 11 and 13 DEGs for the tissue data and SCT data respectively. The xenobiotic metabolism PXR signaling pathway encompassed 176 genes and contained 12 and 13 DEGs for the tissue data and SCT data respectively. SULT1A1 is the only overlapping gene between the two datasets (Table 4.6).

The sucrose V (Mammalian) had upregulated sucrose alpha-glucosidase, ketohexoki-

nase and trikinase and had upregulated components in fructose-biophosphatase aldolase (Figure 4.24).

Table 4.6: Differentially expressed molecules in the xenobiotic metabolism CAR signaling pathway and the xenobiotic metabolism PXR signaling pathway using the LimmaFC data and the SCT data.

CAR (LimmaFC)	PXR (LimmaFC)	CAR (SCT)	PXR (SCT)
ABCC2	ABCC2	ALDH2	ALDH2
CHST5	CHST5	ALDH3A1	ALDH3A1
CYP3A4	CYP3A4	GSTA1	CES2
GSTA2	GSTA2	GSTK1	GSTA1
	PPP1R14D	GSTO1	GSTK1
SULT1A1	SULT1A1	HSP90B1	GSTO1
SULT1A2	SULT1A2	MAP3K	HSP90B1
SULT1A3/SULT1A4	SULT1A3/SULT1A4	MGST2	MGST2
UGT1A1	UGT1A1	MGST3	MGST3
UGT2B7	UGT2B7	PP2R1A	NCRO1
UGT2B11	UGT2B11	RACK1	PP1CA
UGT2B17	UGT2B17	SCAND1	SCAND1
		SULT1A1	SULT1A1

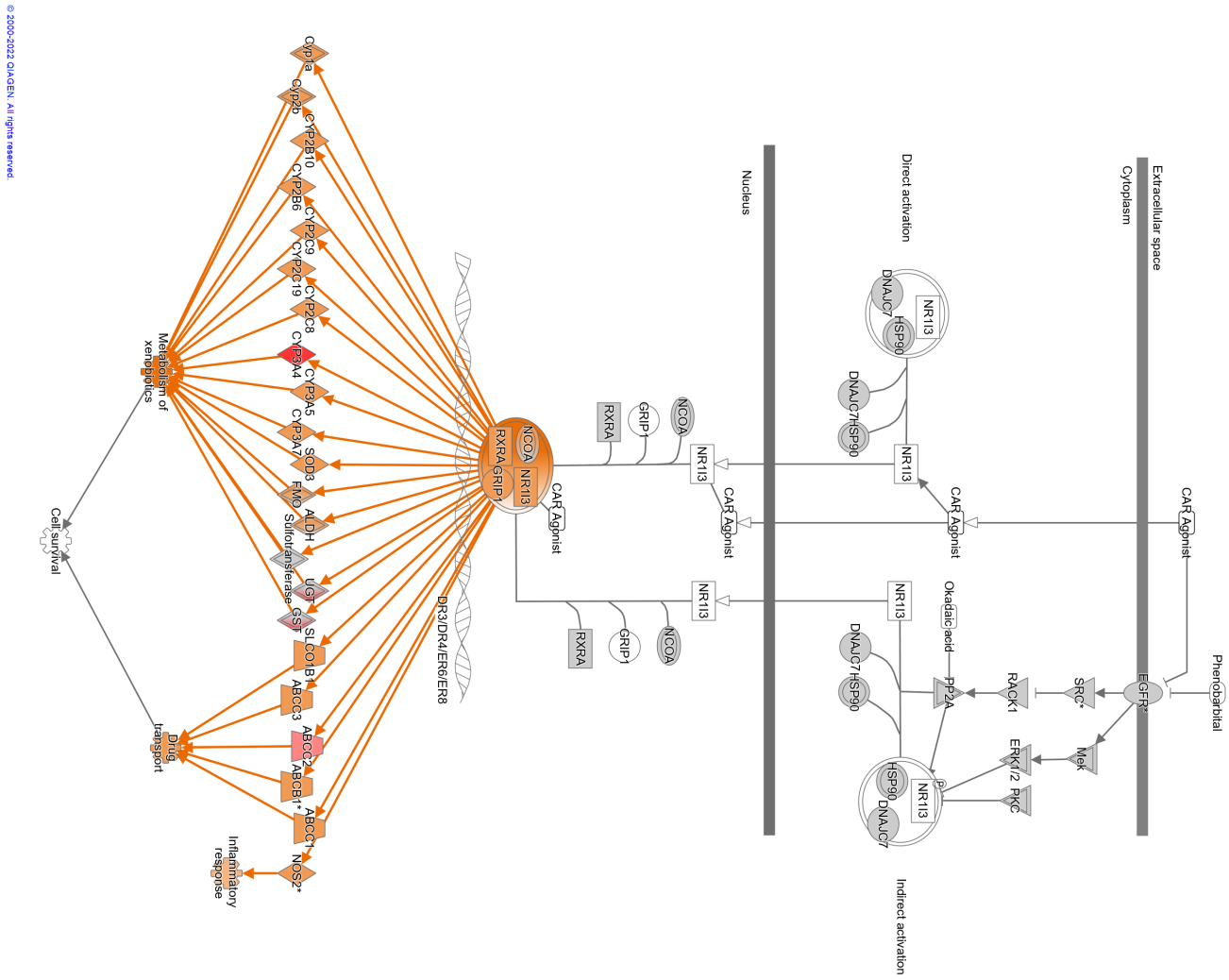


Figure 4.20: Xenobiotic metabolism CAR signaling pathway overlayed with the LimmaFC gene expression with the Molecule Activity Predictor (MAP) activated. Red molecules are known as upregulated genes from the dataset. Orange components are predicted upregulated based on the known expression. Gray components are present, but insignificantly differentially expressed, and excludes a prediction. White components not present and exclude a prediction. Mixed colors indicate a complex with components containing differing information. The CAR antagonist is white and can't be activated or inhibited during prediction.

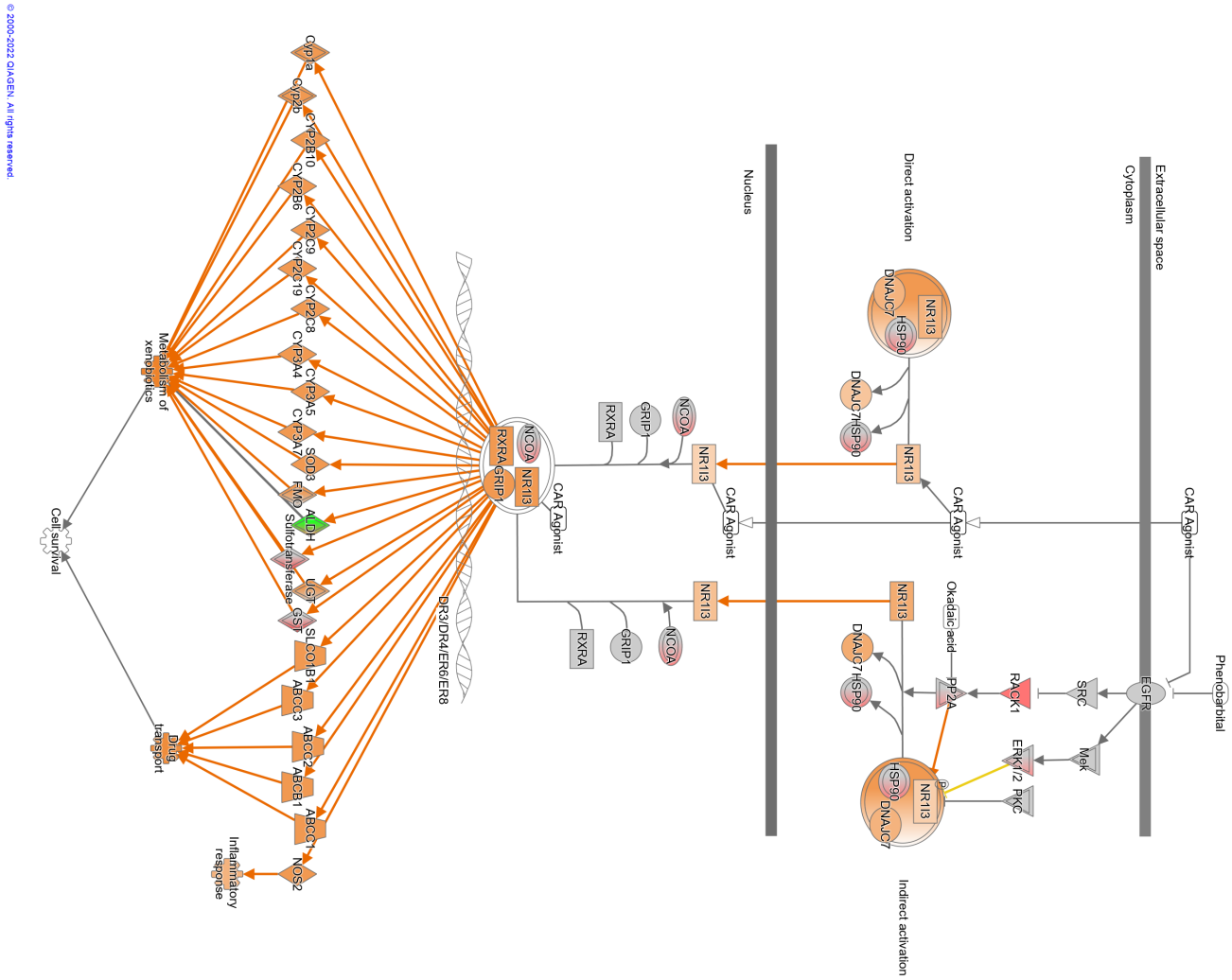


Figure 4.21: Xenobiotic metabolism CAR signaling pathway overlayed with the SCT gene expression with the Molecule Activity Predictor (MAP) activated. Red molecules are known as upregulated genes from the dataset. Orange components are predicted upregulated based on the known expression. Green molecules are known as down-regulated genes from the dataset. Blue components are predicted inhibited based on the known expression. Gray components are present, but insignificantly differentially expressed, and excludes a prediction. White components not present and exclude a prediction. Mixed colors indicate a complex with components containing differing information. The CAR antagonis is white and can't be activated or inhibited during prediction.

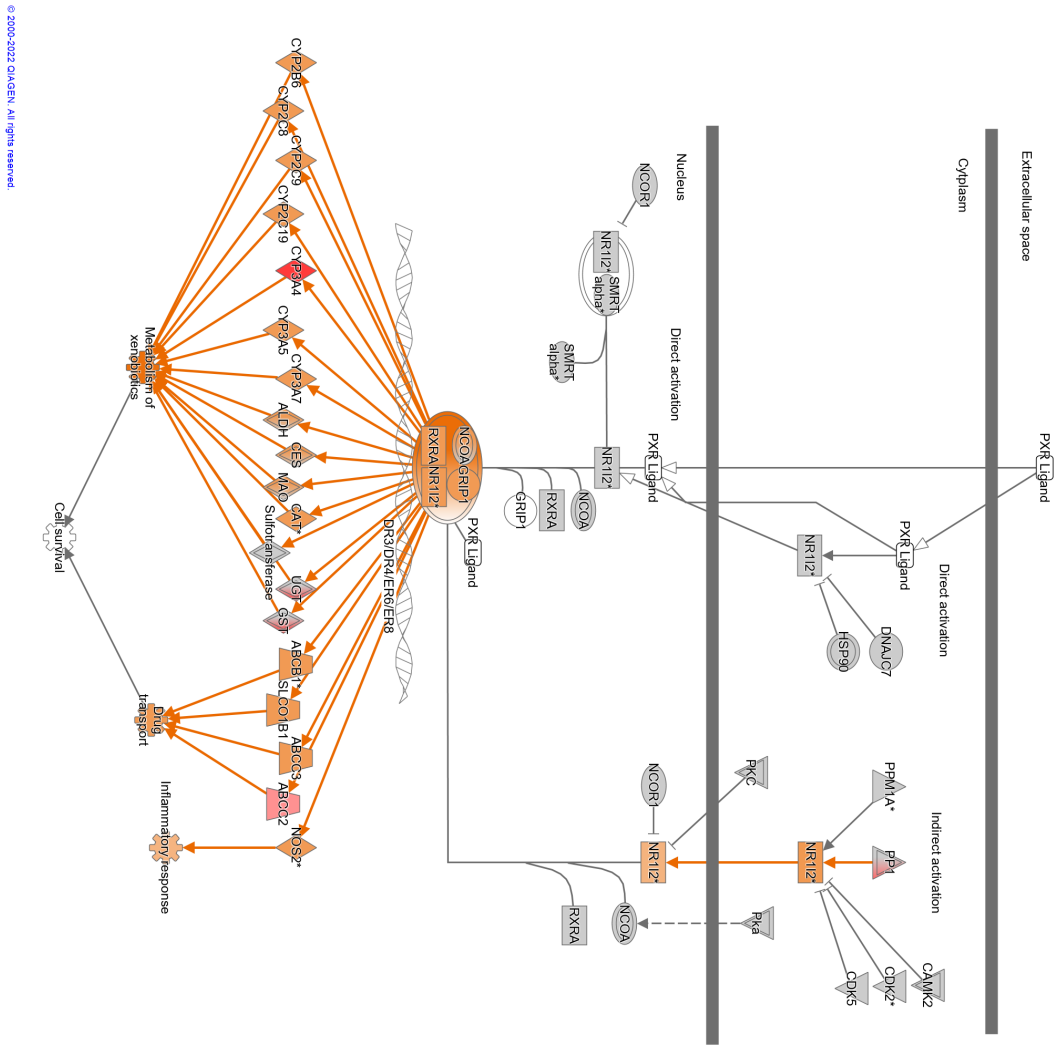


Figure 4.22: Xenobiotic metabolism PXR signaling pathway overlayed with the LimmaFC gene expression with the Molecule Activity Predictor (MAP) activated. Red molecules are known as upregulated genes from the dataset. Orange components are predicted upregulated based on the known expression. Gray components are present, but insignificantly differentially expressed, and excludes a prediction. White components not present and exclude a prediction. Mixed colors indicate a complex with components containing differing information. The PXR ligand is white and can't be activated or inhibited during prediction.

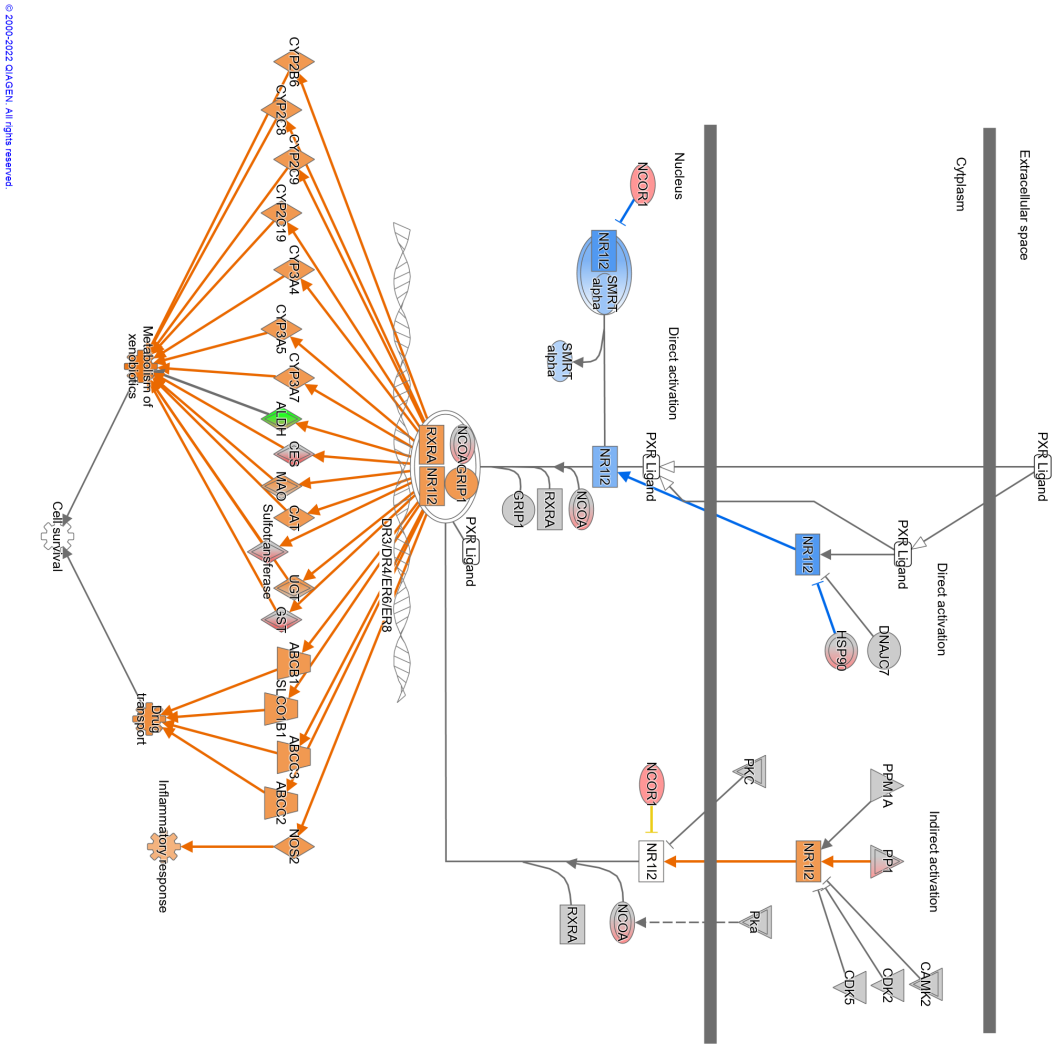


Figure 4.23: Xenobiotic metabolism PXR signaling pathway overlayed with the SCT gene expression with the Molecule Activity Predictor (MAP) activated. Red molecules are known as upregulated genes from the dataset. Orange components are predicted upregulated based on the known expression. Green molecules are known as down-regulated genes from the dataset. Blue components are predicted inhibited based on the known expression. Gray components are present, but insignificantly differentially expressed, and excludes a prediction. White components not present and exclude a prediction. Mixed colors indicate a complex with components containing differing information. The PXR ligand is white and can't be activated or inhibited during prediction.

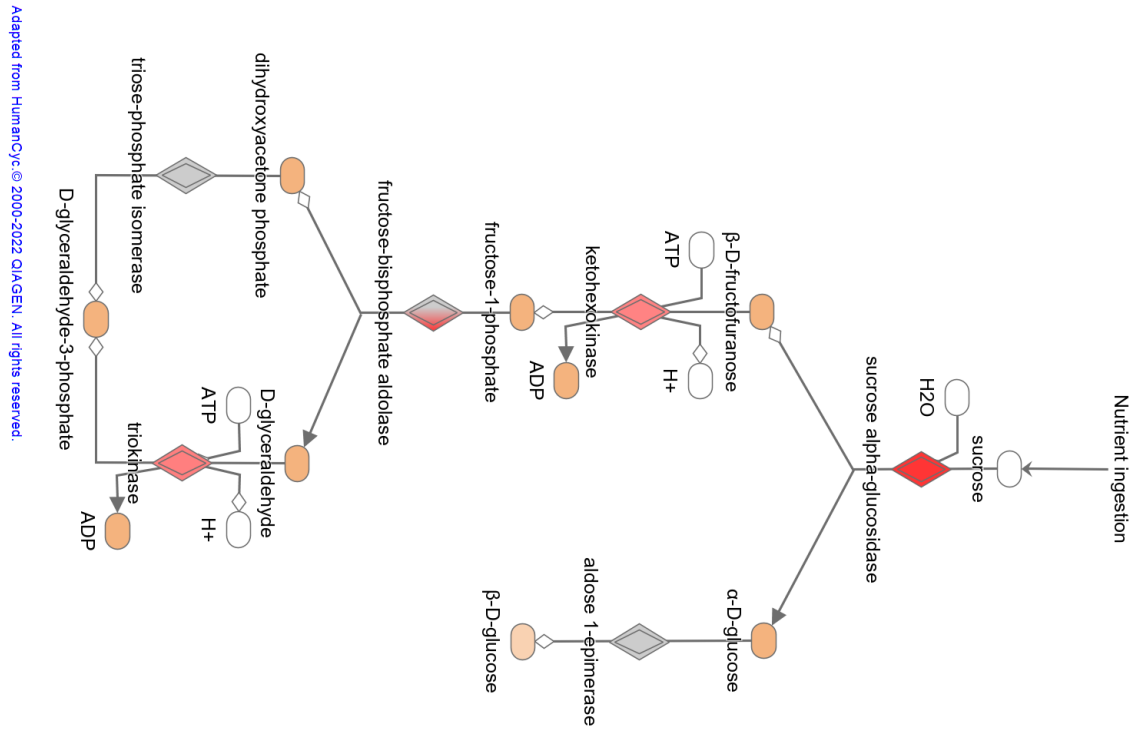


Figure 4.24: Sucrose V (mammalian) pathway overlayed with the LimmaFC gene expression with the Molecule Activity Predictor (MAP) activated. Red molecules are known as upregulated genes from the dataset. Orange components are predicted upregulated components based on the known expression. Gray components are present, but insignificantly differentially expressed, and excludes a prediction. White components not present and exclude a prediction. Mixed colors indicate a complex with components containing differing information.

4.3.2 Upstream analysis

When comparing the tissue and SCT datasets through IPA, and their common upstream regulators, HNF1A were the most prominent. HNF1A, elaidic acid, HNF4A, and CDX2 were indicated as activated in both datasets with Z-scores above 2. HNF4α dimer, EGF, and D-glucose were only activated in a single datasets. PPARGC1A, GATA4, and L-triiodothyronine had a positive, but statistically insignificant z-score. Most of the upstream regulators were transcription regulators, but IPA also highlighted elaidic acid, L-triiodothyronine, and D-glucose as significant endogenous chemicals (Table 4.7).

Table 4.7: Top 10 common upstream regulators between tissue sample and the SCT. Exogenous chemical compounds removed from list.

Top Upstream Regulator	P-value (Tissue)	P-value (SCT)	Predicted activation	Molecule type
HNF1A	5,34E-15	1,07E-4	Activated	Transcription regulator
elaidic acid	4,40E-4	7,19E-4	Activated	chemical - endogeneous mammalian
HNF4A	1,81E-5	6,14E-15	Activated	Transcription regulator
CDX2	3,72E-8	3,08E-2	Activated	Transcription regulator
HNF4α dimer	2,14E-6	8,04E-5	Mixed Activated	complex
PPARGC1A	9,14E-3	3,00E-3	(Insignificant)	Transcription regulator
GATA4	2,48E-3	2,03E-2	(Insignificant)	Transcription regulator
EGF	8,35E-3	1,27E-7	Mixed Activated	Growth factor
L-triiodothyronine	1,79E-4	4,86E-2	(Insignificant)	chemical - endogeneous mammalian
D-glucose	5,47E-3	3,71E-12	Mixed Activated	chemical - endogeneous mammalian

5 Discussion

5.1 Normalization methods and outcomes

This thesis presents several different normalization methods. The different methods were used to investigate differences between potential results.

5.1.1 Housekeeping genes

The housekeeping genes were not uniformly expressed in each sample. Thus, using only one housekeeping gene would have a large effect on the normalization outcome. The usage of several housekeeping genes in normalization is needed to reduce biological noise [45]. The genes with a lower expression were the most stable in geNorm and the visual analysis. It is unknown how much the background noise sway the results with lowly expressed genes. OAZ1 was removed to negate its impact on the normalization process, because of its high expression. The common reference genes ACTB, HPRT1, and 18S rRNA were disregarded in the analysis because of their suboptimal performance as single reference genes [66]. However, they can be beneficial when using multiple housekeeping genes and should not necessarily be disregarded in future studies.

The differentially expressed genes from the SCT dataset using DESeq2 global normalization mostly coincided with the geNorm results for the tissue dataset, except for OAZ1. The difference between stably expressed genes in the dataset can be based on biological discrepancies, indicating the need to double-check individual datasets when normalizing with housekeeping genes. Note that the SCT contained only antrum samples and the differential expression analysis used GIM vs non-atrophic gastritis (NAG), while the tissue dataset was composed of more biological diversity in its samples regarding sample location and pathology.

The most unstable genes in the samples were RPL29, PMM1, GAPDH, and B2M that

are, except for PMM1, widely known housekeeping genes [66, 97]. RPL29 and B2M gave a varied expression throughout all the samples, while GAPDH spiked at some cancer samples, indicating an increase in GAPDH in GC. Findings confirms that GAPDH are dysregulated in cancer [61, 98] and that GAPDH is not a suitable normalization factor for GC [64]. GAPDH is more recently being considered as a potential biomarker for GC [99].

5.1.2 Comparison of outcomes using different normalization methods

The differences in housekeeping gene normalization and between the different global normalization methods can be seen when comparing the significant pathways with DEGs in GIM vs normal tissue samples. The removal of background noise in Limma (LimmaFC) gave the most significant pathways, but is more prone to false positives than uncorrected data [100]. DESeq2 normalized data differed from limma normalized data (LimmaF) when both went through the limma differential expression equation, but were quite similar when DESeq2 went through its pipeline which is originally meant for single-cell data. The cause of this difference is unknown and could be an interesting subject to investigate in the future. The housekeeping gene normalization had fewer pathways compared to the limma pipeline, but the pathways it had coincided with the other methods, confirming its efficacy.

While background correction increases the rate of false positives [100], it can be useful for novel findings in data mining. Using multiple datasets will reduce the effect of false positives, as identical false positives are less likely to happen in both datasets simultaneously. Two datasets were therefore chosen to proceed with the tissue dataset being background corrected.

5.2 Visualization through dimensional reduction and heatmap

5.2.1 Dimensional reduction of tissue

All dimensional reduction methods showed similar results. No clear clusters based on diagnosis were made with either MDS, tSNE, or UMAP plots using the whole gene set. The tSNE and UMAP plots displayed better separation than the MDS plots. Some subclusters with an identical diagnosis could be extrapolated but the borders were ambiguous. These subclusters were predominantly clustered by sample location. The antrum and cardia demonstrated, while imperfect, more desirable clustering than the normal, GIM, and GC tissue separation. The sample location clustering is likely caused by the different gastric cell compositions [101, 102, 103]. A random sample without annotation would range from impossible to easy to place, making the usefulness inconsistent. However, the noteworthy effect the sample location has on dimensional reduction indicates a need for consideration for future research. MDS plotting was not considered for gene subgroup visualization, but should not be neglected for its potential.

5.2.2 Dimensional reduction of cell-related genes

Cell-related genes, made with tSNE and UMAP, separated the samples more than using all genes, except for cancer-related genes. However, the samples separated only into two distinct clusters. An ideal separation would illustrate three or more clusters separating normal tissue, GIM, and GC, which would indicate separate conditions. Additional clusters could include different stages of metaplasia or different types of GC, i.e. intestinal-type and diffiuse-type GC. Both clusters contained every classification, making them impossible to separate.

Neither tSNE nor UMAP was deemed superior to the other, but tSNE separated the clusters to a larger extent. Both the goblet and chief cell related genes were unable to separate properly using UMAP clustering. The enterocyte related genes were separated

in both methods, but tSNE clustered the samples tighter together, making the separation within the cluster harder. However, if the samples separated into distinct clusters based on histology a tight cluster would not pose an issue. The chief cell related genes did not separate the antrum samples, regardless of histology. This is likely because chief cells are not typically present in the antrum [103]. While some variation can be seen with the other normalization methods, the difference was negligible. However, the limma background corrected and normalized data was deemed to have the best visually distinct clusters by a small margin across most figures.

5.2.3 Heatmap

The heatmap coincided with the findings in the dimensional reduction. All three histological states had samples that could be placed into subgroups, but the subgroups within each histological state were more related to subgroups from other histological states than each other. The sample location displayed some clustering within the different histological states, but the GIM and GC samples did not substantially cluster.

The GIM samples were the most consistent and contained similar expressions in 7 of 11 samples. Note that GIM was the smallest sample pool. Within these 7 samples were goblet and enterocyte related genes consistently upregulated, congruent with the SCT report [34]. However, no downregulation was extrapolated from the enteroendocrine cells.

The pathologically classified cancer samples overlapped with cancer-related genes in most of its samples but not all. The biopsy samples classified as cancer had an overlap but contained several samples without a clear upregulation. The upregulation of cancer-related genes was, however, not exclusive to cancer-classified samples. A small distinction between non-cancer and cancer samples regarding genes related to neck-like cells, proliferative cells, and macrophages were made, however, the cancer-related genes were not uniformly upregulated in cancer samples, which can question the method or the dataset.

An overview of patient-separated clustering showed that many commonalities between histologies were contained within a few patients. All patients were diagnosed with cancer, but not all of them shared the same pattern between histological states. Some differences are expected from the heterogeneity of cancer itself [104] and individual differences [105]. However, the lack of a reliable pattern made it difficult to set a diagnosis by comparing patient samples. For instance, patient no. 2 had all 4 samples classified as cancer but showed no indication of upregulation in cancer-related genes. Patient no. 16 had 2 samples classified as cancer while only one of them showed an upregulation. Note that the sample from patient no. 16 without an upregulation had a biopsy classification of 2 out of 5, not 1 out of 5, where 1 indicate certainly cancer.

The SCT dataset was automatically clustered based on its aggregated version. Clear distinctions were made between normal samples and samples with an upregulation of cancer-related genes, namely, severe intestinal metaplasia and early GC. The intestinal metaplasia showed an upregulation in enterocytes and antral basal gland mucous cell related genes, and downregulation in enteroendocrine cells. The goblet cell related genes are naturally upregulated as they are not present in normal gastric tissue. These findings coincided with the initial report [34]. However, a small upregulation in B-cell related genes seemed to be present, but the significance is impossible to determine with the heatmap visualization method. Similar results indicates that a heatmap with the SCT genes can be a useful visual tool for cellular change in gastric tissue.

5.3 Perspective of sample visualization

Both visualization methods defined subgroups within the histological states but the difference between these was greater than outside the histological states. The importance of the sample location was highlighted in the dimensional reduction plots but was not as clear in the heatmaps.

The lack of consistency with prior histological classification infers an error. The SCT dataset had its results reproduced suggesting that the method works. However, the SCT

dataset were a more ideal dataset to work with. The SCT dataset contained few samples, which were only located in the antrum. A single sample site reduces the biological noise from different cellular compositions in the stomach [101, 102, 103].

The concurrence of the cancer classification was poor, with 12 pathologically classified cancer and 22 biopsy-classified cancers over 25 samples, showing a large discrepancy. Disagreements between biopsy and pathology do occur [106]. A diagnosis of borderline lesions or undifferentiated cancer was more likely to diverge, but later stages of cancer had concurrence rates of 93,6 % and 98,6 % [107]. It is deemed highly unlikely that GC would remain a misdiagnosis at the point of operation. The patients in this study underwent extensive resection with either distal gastrectomy, subtotal gastrectomy, or total gastrectomy during the period 2012–2013. Most patients had a poor survival rate even after gastrectomy. A poor survival rate even after surgery is common due to late diagnosis, presenting them with advanced-stage GC [108].

Due to sampling and biological diversity, an additional factor of error can be tissue composition. Tissue composition is important when accounting for cancerous tissue that interweave with normal tissue, especially diffuse-type GC. The pathological samples used for microarray analysis might contain interweaved healthy tissue. Several patients underwent chemotherapy before the surgery, which likely reduced the number of cancer cells in the tissue before sampling. The survival rate of the patients indicated that cancer in some may have relapsed, even after the extensive surgery encompassing the estimated resection margin of 2-6 cm [3]. The cause of patient deaths was not available. The relapse of cancer may indicate cancer cells in tissue previously evaluated as healthy. Although the number of cancer cells should be small, they may have contributed to upregulated cancer-related genes in some samples.

5.4 IPA results

5.4.1 Signaling pathways

5.4.1.1 Tissue dataset

The top 5 common pathways in the tissue dataset were 4 metabolism pathways and 1 signaling pathway. The metabolism pathways share an upregulation of 4 sulfotransferase 1 (SULT1) family genes, UPD-Glycosyltransferase (UGT)1A1, and 4 UGT2 family genes [109]. The SULT genes are generally known for their metabolism of xenobiotics and are in focus today because of their potential as a drug target for personalized medicine [110, 111]. UGT1 and UGT2 are considered to have a relatively minor contribution to drug metabolism. UGT1A1 is central to the regulation of endogenous and exogenous stimuli. UGT2 family genes are considered important in contributing to interindividual differences in drug disposition. Polymorphisms in UGT related genes are common and some are related to increased cancer risk [112].

The serotonin degradation pathway contained an upregulation of ADH4 and ADH6 compared to the other metabolic pathways. Serotonin is a potent neurotransmitter known to affect the bone, brain [113], and skin [114]. Most circulating serotonin is produced by specialized neuroendocrine enterochoffamine cells in the intestine [113]. Enterochoffamin cells are known to be lining the mucosa [115]. The upregulation of serotonin degradation would decrease the homeostatic level of serotonin.

The thyroid hormone metabolism II did not contain any unique upregulated genes. The main thyroid hormone is produced solely by the thyroid gland [116]. The thyroid hormone receptors are located in the myocardium and vascular tissue [117]. It is unlikely that the thyroid hormone metabolism II was significant in this dataset.

The melatonin degradation I and the superpathway of melatonin degradation contained a unique upregulation of CYP3A4 and did not differ between themselves. Melatonin mainly regulates the sleep/wake cycle and other seasonal rhythms while acting as a cytoprotective agent [118]. While melatonin is related to the pineal gland, it is also

produced in numerous extrapineal sites. The gut contains about 400–500 times more melatonin than the pineal gland and is substantially unmetabolized in the intestinal lumen [119]. The development of intestinal tissue in the stomach may be the reason for upregulation of the melatonin degradation pathway.

The FXR/RXR activation had a significant upregulation of genes, including FXR (NR1H4) itself, but IPA could not predict an up- or downregulation. FXR is a nuclear receptor [120], which involves distinct functional profiles binding to various ligands by the selective use of transcriptional coregulators [121]. The heterocomplex with FXR and RXR is known to be responsible for maintaining many metabolic pathways, including bile acid regulation and glucose and lipid homeostasis [120].

5.4.1.2 Single-cell transcriptome (SCT) dataset

The SCT dataset provided more differentially expressed genes than the tissue dataset. The additional DEGs produced lower p-values and different top pathways. Among the top 5 pathways, with EIF2 signaling being the exception, were correlated with oxidative phosphorylation (OXPHOS) containing several of the same gene families.

OXPHOS is responsible for ATP production in the mitochondria [122] and is highly upregulated in the dataset. The upregulation of some OXPHOS complexes are correlated to GC [123]. Mitochondrial dysfunction is tightly related to OXPHOS and has a significant number of differentially expressed genes, but predicted neither an up- or downregulation. If the mitochondrial dysfunction affects the mitochondria it is through the reactive oxygen species (ROS) superoxide. ROS should be at a low stable expression due to continuous degradation, dysregulation would decrease ATP production and increase ROS generation, leading to several pathological conditions [122]. While mitochondrial dysfunction was not predicted as an up- or downregulation it had more differentially expressed genes than oxidative phosphorylation. However, mitochondrial dysfunction did contain more genes, making the overlap percentage smaller. OXPHOS was not upregulated in the tissue dataset but due to its significance and strong upregulation in the SCT dataset it should be taken into consideration.

Sirtuin is essential to delay cellular senescence and extend the lifespan of a cell through several mechanisms [124]. The effect of sirtuin regarding cancer is conflicting as it has both carcinogenic and cancer inhibitory [125]. The sirtuin signaling pathway was downregulated, however, note that none of the SIRT family genes had a predicted differential expression.

Eukaryotic initiation factor 2 (EIF2) is a family of protein kinase phosphorylates that alleviate cellular injury or induce apoptosis caused by environmental stress [126]. The EIF2 signaling pathway was upregulated and predicted a strong activation of the endoplasmic reticulum (ER) stress response. A state of persistent ER stress sensor activation with its downstream signaling pathway has been correlated with key regulators for tumor growth and metastasis [127].

Glucocorticoids are hormones that regulate inflammation, metabolism, and stress, which are secreted by the adrenal cortex. The glucocorticoid receptor is a ligand-activated transcription factor deriving different functions depending on the glucocorticoid [128]. However, the intestinal mucose epithelial layer has been demonstrated as an extra-adrenal site for glucocorticoid synthesis and aids in immune homeostasis [129]. An increase in expression in the glucocorticoids receptor signaling pathway could be explained by the development of GIM. The up- or downregulation of the glucocorticoid receptor signaling pathway was unpredicted in IPA.

5.4.1.3 Common signaling pathways

None of the top 5 results in either dataset was within the common signaling pathways, except for the sirtuin signaling pathway. Both sucrose degradation V and the sirtuin signaling pathway was not present without background correction in the tissue data. However, the sirtuin signaling pathway will not be further discussed as the datasets contradict each other and neither dataset included a differentially expressed or predicted up- or downregulated SIRT family gene. The xenobiotic metabolism PXR signaling pathway was insignificant in terms of z-score in the SCT dataset but was close enough to be included. Several pathways had a significant p-value in both datasets but were

disregarded without a significant z-score in one or both pathways.

The constitutive androstane receptor (CAR, NR1I3) and the pregnane X receptor (PXR, NR1I2) are part of the nuclear receptor superfamily. Their functions are mainly associated with the transcription xenobiotic enzymes and are involved in physiological and pathological conditions [130, 131]. While CAR and PXR differ, they share several common features and overlap in both xenobiotic activators and target genes [132]. CAR and PXR are mainly expressed in the liver but are also present in the small intestine, colon, and blood cells. CAR has been detected in the duodenum and PXR has been found in the stomach. The PXR expression in the stomach is 1/10 of the expression in the small intestine [130]. Thus, the increase in the xenobiotic metabolism CAR and PXR signaling pathways can be from the development of GIM. The signaling pathways had similar predictions, even with just one shared upregulated gene, except for the upstream cascade in the xenobiotic metabolism PXR signaling pathway. The xenobiotic metabolism PXR signaling pathway had an upregulation of NCOR1 and HSP90B1 in the SCT dataset, predicting an inhibition, however, this inhibition did not affect the downstream result. If the common gene, SULT1A1, was the only common gene by coincidence or if it was an important factor is unknown. SULTA1A is one of the primary sulfotransferases responsible for sulfoconjugation of xenobiotics in phase II metabolism [133, 134], as well as endogenous compounds [133]. The GC risk of SULT1A1 is investigated through its polymorphisms [135, 136, 137]. The SULT1A1 polymorphism 636G>A has been correlated to increased cancer risk in Caucasians [138]. However, only one study differentiated between intestinal-type and diffuse-type GC, and found an increased risk in diffuse-type but not intestinal-type GC [136]. The diffuse-type GC develops independently from GIM [12].

Sucrose degradation V is a small pathway containing only 8 molecules. The enterocytes in the small intestine produce an abundance of the Na(+)-glucose cotransporter isoform SGLT1 (SLC5A1) and the glucose transporter isoforms GLUT2 (SLC2A2) and GLUT5 (SLC2A5) to facilitate the uptake of nutrients [139]. Subsequently, they have produced several glycosidase types, including sucrase isomaltase [140]. Sucrase isomal-

tase is part of sucrose alpha-glucosidase that is the initiating enzyme in the breakdown of sucrose in sucrose degradation V. The increase of enterocytes through GIM has been well-documented [17, 24, 34]. The increase of enterocytes could explain the upregulation of sucrose degradation V.

5.4.2 Upstream regulators

The tissue and SCT contained mostly transcription factors and components that were not screened for in the analysis. The upstream analysis crossreferences the downstream molecules and makes a prediction [83]. The transcription factors can have a minor upregulation, which is not detected in the differential expression analysis, even if it is significant.

HNF1A, HNF4A, and HNF4α dimer are part of the hepatocyte nuclear factor (HNF) family and were predicted to activate in both datasets, except for HNF4α dimer. HNF4α dimer is a homodimer nuclear receptor composed HNF4A [141]. The HNFs were discovered in hepatocytes in the liver [142], but they are expressed in various organs. HNF1A and HNF4A are expressed in the intestine and HNF4A has been observed with a weak expression in the stomach [143]. However, they have both been correlated with cancer [144]. HNF1A has been discovered with significantly increased expression in stomach adenocarcinoma [145]. HNF4A is suggested as a biomarker to differentiate GC from other breast cancer [146] and its inhibition can decrease proliferative cells [36].

CDX2 is a known biomarker for GIM [17] and has been associated with HNF1A in esophagus intestinal metaplasia [147]. However, CDX2 can transactivate itself in gastrointestinal human carcinoma cell lines and binds to its promoter in GIM, thus suggested as a perpetrator in the inevitable progression to GC [148]. CDX2 and HNF4A are tumor suppressors in intestinal cancer [149] but their presence in the stomach is yet to be elucidated. CDX2, HNF4A, and GATA4 are important for the general development of the intestine. CDX2 and HNF4A are considered paramount for developing enterocytes [150]. HNF4A, together with SMAD4, has also been shown to be crucial to

the differentiation and development of enterocytes with a reinforcing feed-forward loop [151]. SMAD4 was slightly downregulated in both datasets, but not statistically. No activation or inhibition prediction was made with SMAD4. GATA4 is necessary for the proper development of the stomach [152], by repressing the forestomach development during hindstomach development [153]. While GATA4 is present in 10% of cancers, it is mainly correlated to the upregulation of HNF4A [36]. The prediction with GATA4 highlighted it as important, but it was not significantly upregulated. The differentiation of enterocytes and increase in CDX2, HFN4A, and GATA4 during the creation of GIM may cause a reinforcing feedback loop in the microbiome making the condition persist.

Elaicid acid is a major trans fatty acid that can induce apoptosis through ROS accumulation and endoplasmic reticulum stress in neuroblastoma cell lines [154] and enhances tumor growth in colorectal cancer [155]. PPARGC1A was not significantly upregulated but is correlated with oxidative damage in epithelial cells [156]. Elaicid acid and PPARGC1A are likely correlated to the mitochondrial dysfunction pathway in the SCT dataset because of their causative relation with ROS [154, 156].

Epidermal growth factor (EGF) was predicted activated in the SCT dataset and is important in the functional development of the stomach [157] and intestine [158] but an overexpression of human epidermal growth factor receptor 2 (HER2) is considered a strong biomarker for GC [159].

L-triiodothyronine was predicted activated in neither datasets and is related to the thyroid hormone degradation pathway. Its pathology is mainly correlated to hypothyroidism [160, 161] and diabetes in relation to the stomach [161]. Due to its connection with the thyroid hormone degradation pathway it is thought to be upregulated in the upstream analysis by coincidence.

D-glucose was predicted activated in the tissue dataset and was likely activated due to the increase in metabolism. Cancers have a high glucose consumption [162], and 2-deoxy-d-glucose is modified d-glucose that attempts to inhibit glycolysis [163].

6 Conclusions

The results of this thesis suggested that the potential biomarkers of GIM encompassed signaling pathways (i.e., xenobiotic metabolism CAR signaling pathway, xenobiotic metabolism PXR signaling pathway, and sucrose degradation V), transcription factors (HNF1A, HNF4A, CDX2, PPARGC1A, and GATA4), endogenous chemicals (elaidic acid and d-glucose), and the growth factor EGF.

Different data preparation methods had negligible differences during visual analysis, but discrepancies were clear during IPA analysis.

The visual analysis of the tissue data yielded figures difficult to interpret due to biological noise. Gastric intestinal metaplasia gave the most distinct expression, while normal and tissue expression interweaved more. The pathological classification of cancer was more distinct from the normal samples and was more conservative in its classification, compared to the biopsy classification.

References

[1] Ferlay J, Colombet M, Soerjomataram I, Mathers C, Parkin DM, Piñeros M, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int J Cancer. 2019;144(8):1941-53.

[2] Larsen IK. Cancer in Norway 2019. Cancer Registry of Norway. 2019.

[3] Johnston FM, Beckman M. Updates on Management of Gastric Cancer. Curr Oncol Rep. 2019;21(8):67.

[4] Tan P, Yeoh KG. Genetics and Molecular Pathogenesis of Gastric Adenocarcinoma. Gastroenterology. 2015;149(5):1153-62.

[5] Uemura N, Okamoto S, Yamamoto S, Matsumura N, Yamaguchi S, Yamakido M, et al. Helicobacter pylori infection and the development of gastric cancer. N Engl J Med. 2001;345(11):784-9.

[6] Plummer M, Franceschi S, Vignat J, Forman D, de Martel C. Global burden of gastric cancer attributable to Helicobacter pylori. Int J Cancer. 2015;136(2):487-90.

[7] Waldum HL, Sagatun L, Mjønes P. Gastrin and Gastric Cancer. Front Endocrinol (Lausanne). 2017;8:1.

[8] Lee L, Ramos-Alvarez I, Ito T, Jensen RT. Insights into Effects/Risks of Chronic Hypergastrinemia and Lifelong PPI Treatment in Man Based on Studies of Patients with Zollinger-Ellison Syndrome. Int J Mol Sci. 2019;20(20).

[9] Hayakawa Y, Fox JG, Wang TC. The Origins of Gastric Cancer From Gastric Stem Cells: Lessons From Mouse Models. Cell Mol Gastroenterol Hepatol. 2017;3(3):331-8.

[10] Westphalen CB, Asfaha S, Hayakawa Y, Takemoto Y, Lukin DJ, Nuber AH, et al. Long-lived intestinal tuft cells serve as colon cancer-initiating cells. J Clin Invest. 2014;124(3):1283-95.

[11] Waldum HL, Hauso Sørdal , Fossmark R. Gastrin May Mediate the Carcinogenic Effect of Helicobacter pylori Infection of the Stomach. Dig Dis Sci. 2015;60(6):1522-7.

[12] McLean MH, El-Omar EM. Genetics of gastric cancer. Nature Reviews Gastroenterology Hepatology. 2014;11:664-74.

[13] LAUREN P. THE TWO HISTOLOGICAL MAIN TYPES OF GASTRIC CARCINOMA: DIFFUSE AND SO-CALLED INTESTINAL-TYPE CARCINOMA. AN ATTEMPT AT A HISTO-CLINICAL CLASSIFICATION. Acta Pathol Microbiol Scand. 1965;64:31-49.

[14] Smyth EC, Nilsson M, Grabsch HI, van Grieken NC, Lordick F. Gastric cancer. Lancet. 2020;396(10251):635-48.

[15] Matsuoka T, Yashiro M. Biomarkers of gastric cancer: Current topics and future perspective. World J Gastroenterol. 2018;24(26):2818-32.

[16] Slack JM, Tosh D. Transdifferentiation and metaplasia–switching cell types. Curr Opin Genet Dev. 2001;11(5):581-6.

[17] Kinoshita H, Hayakawa Y, Koike K. Metaplasia in the Stomach-Precursor of Gastric Cancer? Int J Mol Sci. 2017;18(10).

[18] Correa P. Human gastric carcinogenesis: a multistep and multifactorial process–First American Cancer Society Award Lecture on Cancer Epidemiology and Prevention. Cancer Res. 1992;52(24):6735-40.

[19] Correa P, Shiao YH. Phenotypic and genotypic events in gastric carcinogenesis. Cancer Res. 1994;54(7 Suppl):1941s-1943s.

[20] Shichijo S, Hirata Y, Niikura R, Hayakawa Y, Yamada A, Ushiku T, et al. Histologic intestinal metaplasia and endoscopic atrophy are predictors of gastric cancer development after Helicobacter pylori eradication. Gastrointest Endosc. 2016;84(4):618-24.

[21] Shichijo S, Hirata Y, Sakitani K, Yamamoto S, Serizawa T, Niikura R, et al. Distribution of intestinal metaplasia as a predictor of gastric cancer development. J Gastroenterol Hepatol. 2015;30(8):1260-4.

[22] Sakitani K, Hirata Y, Watabe H, Yamada A, Sugimoto T, Yamaji Y, et al. Gastric cancer risk according to the distribution of intestinal metaplasia and neutrophil infiltration. J Gastroenterol Hepatol. 2011;26(10):1570-5.

[23] Graham DY, Zou WY. Guilt by association: intestinal metaplasia does not progress to gastric cancer. Curr Opin Gastroenterol. 2018;34(6):458-64.

[24] Goldenring JR, Mills JC. Cellular Plasticity, Reprogramming, and Regeneration: Metaplasia in the Stomach and Beyond. Gastroenterology. 2021.

[25] Lam SK, Lau G. Novel treatment for gastric intestinal metaplasia, a precursor to cancer. JGH Open. 2020;4(4):569-73.

[26] Gupta S, Li D, El Serag HB, Davitkov P, Altayar O, Sultan S, et al. AGA Clinical Practice Guidelines on Management of Gastric Intestinal Metaplasia. Gastroenterology. 2020;158(3):693-702.

[27] Pimentel-Nunes P, Libânio D, Marcos-Pinto R, Areia M, Leja M, Esposito G, et al. Management of epithelial precancerous conditions and lesions in the stomach (MAPS II): European Society of Gastrointestinal Endoscopy (ESGE), European Helicobacter and Microbiota Study Group (EHMSG), European Society of Pathology (ESP), and Sociedade Portuguesa de Endoscopia Digestiva (SPED) guideline update 2019. Endoscopy. 2019;51(4):365-88.

[28] Yeoh KG, Tan P. Mapping the genomic diaspora of gastric cancer. Nat Rev Cancer. 2021.

[29] Milward EA, Shahandeh A, Heidari M, Johnstone DM, Daneshi N, Hondermarck H. Transcriptomics. In: Bradshaw RA, Stahl PD, editors. Encyclopedia of Cell Biology. Waltham: Academic Press; 2016. p. 160-5. Available from: `https://www.sciencedirect.com/science/article/pii/B9780123944474400295`.

[30] Supplitt S, Karpinski P, Sasiadek M, Laczmanska I. Current Achievements and Applications of Transcriptomics in Personalized Cancer Medicine. Int J Mol Sci. 2021;22(3).

[31] Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLoS Comput Biol. 2017;13(5):e1005457.

[32] Yong HEJ, Chan SY. Current approaches and developments in transcript profiling of the human placenta. Hum Reprod Update. 2020;26(6):799-840.

[33] Chambers DC, Carew AM, Lukowski SW, Powell JE. Transcriptomics and single-cell RNA-sequencing. Respirology. 2019;24(1):29-36.

[34] Zhang P, Yang M, Zhang Y, Xiao S, Lai X, Tan A, et al. Dissecting the Single-Cell Transcriptome Network Underlying Gastric Premalignant Lesions and Early Gastric Cancer. Cell Rep. 2019;27(6):1934-47.

[35] Companioni O, Sanz-Anquela JM, Pardo ML, Puigdecanet E, Nonell L, García N, et al. Gene expression study and pathway analysis of histological subtypes of intestinal metaplasia that progress to gastric cancer. PLoS One. 2017;12(4):e0176043.

[36] Chia NY, Deng N, Das K, Huang D, Hu L, Zhu Y, et al. Regulatory crosstalk between lineage-survival oncogenes KLF5, GATA4 and GATA6 cooperatively promotes gastric cancer development. Gut. 2015;64(5):707-19.

[37] Zhang M, Hu S, Min M, Ni Y, Lu Z, Sun X, et al. Dissecting transcriptional heterogeneity in primary gastric adenocarcinoma by single cell RNA sequencing. Gut. 2021;70(3):464-75.

[38] Chen M, Xie Y, Story M. An Exponential-Gamma Convolution Model for Background Correction of Illumina BeadArray Data. Commun Stat Theory Methods. 2011;40(17):3055-69.

[39] Ritchie ME, Forrest MS, Dimas AS, Daelemans C, Dermitzakis ET, Deloukas P, et al. Data analysis issues for allele-specific expression using Illumina's GoldenGate assay. BMC Bioinformatics. 2010;11:280.

[40] Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, et al. A comparison of background correction methods for two-colour microarrays. Bioinformatics. 2007;23(20):2700-7.

[41] Zahurak M, Parmigiani G, Yu W, Scharpf RB, Berman D, Schaeffer E, et al. Pre-processing Agilent microarray data. BMC Bioinformatics. 2007;8:142.

[42] Yang YH, Buckley MJ, Speed TP. Analysis of cDNA microarray images. Brief Bioinform. 2001;2(4):341-9.

[43] Tran PH, Peiffer DA, Shin Y, Meek LM, Brody JP, Cho KW. Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. Nucleic Acids Res. 2002;30(12):e54.

[44] Huggett J, Dheda K, Bustin S, Zumla A. Real-time RT-PCR normalisation; strategies and considerations. Genes Immun. 2005;6(4):279-84.

[45] Chapman JR, Waldenström J. With Reference to Reference Genes: A Systematic Review of Endogenous Controls in Gene Expression Studies. PLoS One. 2015;10(11):e0141853.

[46] Qin S, Kim J, Arafat D, Gibson G. Effect of normalization on statistical and biological interpretation of gene expression profiles. Front Genet. 2012;3:160.

[47] de Kok JB, Roelofs RW, Giesendorf BA, Pennings JL, Waas ET, Feuth T, et al. Normalization of gene expression measurements in tumor tissues: comparison of 13 endogenous control genes. Lab Invest. 2005;85(1):154-9.

[48] Adeola F. Normalization of Gene Expression by Quantitative RT-PCR in Human Cell Line: comparison of 12 Endogenous Reference Genes. Ethiop J Health Sci. 2018;28(6):741-8.

[49] Aggarwal J, Sharma A, Kishore A, Mishra BP, Yadav A, Mohanty A, et al. Identification of suitable housekeeping genes for normalization of quantitative real-time PCR data during different physiological stages of mammary gland in riverine buffaloes (Bubalus bubalis). J Anim Physiol Anim Nutr (Berl). 2013;97(6):1132-41.

[50] Do JH, Choi DK. Normalization of microarray data: single-labeled and dual-labeled arrays. Mol Cells. 2006;22(3):254-61.

[51] Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25.

[52] Colantuoni C, Henry G, Zeger S, Pevsner J. SNOMAD (Standardization and NOrmalization of MicroArray Data): web-accessible gene expression data analysis. Bioinformatics. 2002;18(11):1540-1.

[53] Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 2013;14(6):671-83.

[54] Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform. 2015;16(1):59-70.

[55] Park T, Yi SG, Kang SH, Lee S, Lee YS, Simon R. Evaluation of normalization methods for microarray data. BMC Bioinformatics. 2003;4:33.

[56] Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol. 2002;3(7):RESEARCH0034.

[57] Fundel K, Haag J, Gebhard PM, Zimmer R, Aigner T. Normalization strategies for mRNA expression data in cartilage research. Osteoarthritis Cartilage. 2008;16(8):947-55.

[58] Lee MLT. Analysis of Microarray Gene Expression Data. New York, NY: Springer US : Imprint: Springer; 2004.

[59] Thellin O, Zorzi W, Lakaye B, De Borman B, Coumans B, Hennen G, et al. Housekeeping genes as internal standards: use and limits. J Biotechnol. 1999;75(2-3):291-5.

[60] Bustin SA. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. J Mol Endocrinol. 2000;25(2):169-93.

[61] Caradec J, Sirab N, Revaud D, Keumeugni C, Loric S. Is GAPDH a relevant housekeeping gene for normalisation in colorectal cancer experiments? Br J Cancer. 2010;103(9):1475-6.

[62] Guo C, Liu S, Wang J, Sun MZ, Greenaway FT. ACTB in cancer. Clin Chim Acta. 2013;417:39-44.

[63] Jo J, Choi S, Oh J, Lee SG, Choi SY, Kim KK, et al. Conventionally used reference genes are not outstanding for normalization of gene expression in human cancer research. BMC bioinformatics. 2019;20(Suppl 10):245-5.

[64] Kwon MJ, Oh E, Lee S, Roh MR, Kim SE, Lee Y, et al. Identification of novel reference genes using multiplatform expression data and their validation for quantitative gene expression analysis. PLoS One. 2009;4(7):e6162.

[65] Rubie C, Kempf K, Hans J, Su T, Tilton B, Georg T, et al. Housekeeping gene variability in normal and cancerous colorectal, pancreatic, esophageal, gastric and hepatic tissues. Mol Cell Probes. 2005;19(2):101-9.

[66] Rho HW, Lee BC, Choi ES, Choi IJ, Lee YS, Goh SH. Identification of valid reference genes for gene expression studies of human stomach cancer by reverse transcription-qPCR. BMC Cancer. 2010;10:240.

[67] Tunbridge EM, Eastwood SL, Harrison PJ. Changed relative to what? Housekeeping genes and normalization strategies in human brain gene expression studies. Biol Psychiatry. 2011;69(2):173-9.

[68] Ni L. Dimensional Reduction. In: Salkind NJ, editor. Encyclopedia of Measurement and Statistics. Thousand Oaks: SAGE Publications Inc.; 2007. p. 264-7. Available from: https://dx.doi.org/10.4135/9781412952644.n138.

[69] Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411-20.

[70] Sun Z, Wang T, Deng K, Wang XF, Lafyatis R, Ding Y, et al. DIMM-SC: a Dirichlet mixture model for clustering droplet-based single cell transcriptomic data. Bioinformatics. 2018;34(1):139-46.

[71] Yip SH, Sham PC, Wang J. Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. Brief Bioinform. 2019;20(4):1583-9.

[72] Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-87.

[73] Bergenstråhle J, Larsson L, Lundeberg J. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. BMC Genomics. 2020;21(1):482.

[74] Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometrics and Intelligent Laboratory Systems. 1987;2(1):37-52. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.

[75] Lugli E, Pinti M, Nasi M, Troiano L, Ferraresi R, Mussi C, et al. Subject classification obtained by cluster analysis and principal component analysis applied to flow cytometric data. Cytometry A. 2007;71(5):334-44.

[76] Hout MC, Papesh MH, Goldinger SD. Multidimensional scaling. Wiley Interdiscip Rev Cogn Sci. 2013;4(1):93-103.

[77] Nam JH, Yun J, Jin IH, Chung D. hubViz: A Novel Tool for Hub-centric Visualization. Chemometr Intell Lab Syst. 2020;203.

[78] Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. Nat Commun. 2019;10(1):5416.

[79] Li X, Gao Y, Xu Z, Zhang Z, Zheng Y, Qi F. Identification of prognostic genes in adrenocortical carcinoma microenvironment based on bioinformatic methods. Cancer Med. 2020;9(3):1161-72.

[80] Maravi ME, Snyder LE, McEwen LD, DeYoung K, Davidson AJ. Using Spatial Analysis to Inform Community Immunization Strategies. Biomed Inform Insights. 2017;9:1178222617700626.

[81] Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32(18):2847-9.

[82] Qiagen. Ingenuity® Pathway Analysis (IPA®); 2015. Accessed: 03.09.2021. Available from: `https://digitalinsights.qiagen.com/files/flyers/IPA_data_sheet_web.pdf`.

[83] Krämer A, Green J, Pollard J, Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics. 2014;30(4):523-30.

[84] Chen Z, Soutto M, Rahman B, Fazili MW, Peng D, Blanca Piazuelo M, et al. Integrated expression analysis identifies transcription networks in mouse and human gastric neoplasia. Genes Chromosomes Cancer. 2017;56(7):535-47.

[85] Companioni O, Bonet C, García N, Ramírez-Lázaro MJ, Lario S, Mendoza J, et al. Genetic variation analysis in a follow-up study of gastric cancer precursor lesions confirms the association of MUC2 variants with the evolution of the lesions and identifies a significant association with NFKB1 and CD14. Int J Cancer. 2018;143(11):2777-86.

[86] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2021. Available from: `https://www.R-project.org/`.

[87] Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the tidyverse. Journal of Open Source Software. 2019;4(43):1686.

[88] Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics. 2008;24(13):1547-8.

[89] Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, Ahmad S, et al. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49(D1):D480-9.

[90] Zhong S. ctrlGene: Assess the Stability of Candidate Housekeeping Genes; 2019. R package version 1.0.1. Available from: https://CRAN.R-project.org/package=ctrlGene.

[91] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15:550.

[92] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015;43(7):e47-7. Available from: https://doi.org/10.1093/nar/gkv007.

[93] Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016.

[94] Ooms J. writexl: Export Data Frames to Excel 'xlsx' Format; 2021. R package version 1.4.0. Available from: https://CRAN.R-project.org/package=writexl.

[95] Morgan M. BiocManager: Access the Bioconductor Project Package Repository; 2021. R package version 1.30.16. Available from: https://CRAN.R-project.org/package=BiocManager.

[96] Du P, Feng G, Kibbe W, Lin S. lumiHumanIDMapping: Illumina Identifier mapping for Human; 2016. R package version 1.10.1.

[97] Wisnieski F, Calcagno DQ, Leal MF, dos Santos LC, Gigek CdeO, Chen ES, et al. Reference genes for quantitative RT-PCR data in gastric tissues and cell lines. World J Gastroenterol. 2013;19(41):7121-8.

[98] Zhang JY, Zhang F, Hong CQ, Giuliano AE, Cui XJ, Zhou GJ, et al. Critical protein GAPDH and its regulatory mechanisms in cancer cells. Cancer Biol Med. 2015;12(1):10-22.

[99] Yang J. Identification of novel biomarkers, MUC5AC, MUC1, KRT7, GAPDH, CD44 for gastric cancer. Med Oncol. 2020;37(5):34.

[100] Xie Y, Wang X, Story M. Statistical methods of background correction for Illumina BeadArray data. Bioinformatics. 2009;25(6):751-7.

[101] Ramsay PT, Carr A. Gastric acid and digestive physiology. Surg Clin North Am. 2011;91(5):977-82.

[102] Bauer M, Morales-Orcajo E, Klemm L, Seydewitz R, Fiebach V, Siebert T, et al. Biomechanical and microstructural characterisation of the porcine stomach wall: Location- and layer-dependent investigations. Acta Biomater. 2020;102:83-99.

[103] Choi E, Roland JT, Barlow BJ, O'Neal R, Rich AE, Nam KT, et al. Cell lineage distribution atlas of the human stomach reveals heterogeneous gland populations in the gastric antrum. Gut. 2014;63(11):1711-20.

[104] Yan HHN, Siu HC, Law S, Ho SL, Yue SSK, Tsui WY, et al. A Comprehensive Human Gastric Cancer Organoid Biobank Captures Tumor Subtype Heterogeneity and Enables Therapeutic Screening. Cell Stem Cell. 2018;23(6):882-97.

[105] Goetz LH, Schork NJ. Personalized medicine: motivation, challenges, and progress. Fertil Steril. 2018;109(6):952-63.

[106] Tae CH, Lee JH, Min BH, Kim KM, Rhee PL, Kim JJ. Negative Biopsy after Referral for Biopsy-Proven Gastric Cancer. Gut Liver. 2016;10(1):63-8.

[107] Takao M, Kakushima N, Takizawa K, Tanaka M, Yamaguchi Y, Matsubayashi H, et al. Discrepancies in histologic diagnoses of early gastric cancer between biopsy and endoscopic mucosal resection specimens. Gastric Cancer. 2012;15(1):91-6.

[108] Tan Z. Recent Advances in the Surgical Treatment of Advanced Gastric Cancer: A Review. Med Sci Monit. 2019;25:3537-41.

[109] Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, et al. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. 2020;48(D1):D265-8.

[110] Suiko M, Kurogi K, Hashiguchi T, Sakakibara Y, Liu MC. Updated perspectives on the cytosolic sulfotransferases (SULTs) and SULT-mediated sulfation. Biosci Biotechnol Biochem. 2017;81(1):63-72.

[111] Hebbring SJ, Moyer AM, Weinshilboum RM. Sulfotransferase gene copy number variation: pharmacogenetics and function. Cytogenet Genome Res. 2008;123(1-4):205-10.

[112] Meech R, Hu DG, McKinnon RA, Mubarokah SN, Haines AZ, Nair PC, et al. The UDP-Glycosyltransferase (UGT) Superfamily: New Members, New Functions, and Novel Paradigms. Physiol Rev. 2019;99(2):1153-222.

[113] Rosen CJ. Serotonin rising–the bone, brain, bowel connection. N Engl J Med. 2009;360(10):957-9.

[114] Semak I, Korik E, Naumova M, Wortsman J, Slominski A. Serotonin metabolism in rat skin: characterization by liquid chromatography-mass spectrometry. Arch Biochem Biophys. 2004;421(1):61-6.

[115] Raghupathi R, Duffield MD, Zelkas L, Meedeniya A, Brookes SJ, Sia TC, et al. Identification of unique release kinetics of serotonin from guinea-pig and human enterochromaffin cells. J Physiol. 2013;591(23):5959-75.

[116] Köhrle J. Thyroid Hormones and Derivatives: Endogenous Thyroid Hormones and Their Targets. Methods Mol Biol. 2018;1801:85-104.

[117] Razvi S, Jabbar A, Pingitore A, Danzi S, Biondi B, Klein I, et al. Thyroid Hormones and Cardiovascular Function and Diseases. J Am Coll Cardiol. 2018;71(16):1781-96.

[118] Hardeland R, Pandi-Perumal SR, Cardinali DP. Melatonin. Int J Biochem Cell Biol. 2006;38(3):313-6.

[119] Hardeland R. Melatonin, hormone of darkness and more: occurrence, control mechanisms, actions and bioactive metabolites. Cell Mol Life Sci. 2008;65(13):2001-18.

[120] Zheng W, Lu Y, Tian S, Ma F, Wei Y, Xu S, et al. Structural insights into the heterodimeric complex of the nuclear receptors FXR and RXR. J Biol Chem. 2018;293(32):12535-41.

[121] Rocchi S, Picard F, Vamecq J, Gelman L, Potier N, Zeyer D, et al. A unique PPARgamma ligand with potent insulin-sensitizing yet weak adipogenic activity. Mol Cell. 2001;8(4):737-47.

[122] Nolfi-Donegan D, Braganza A, Shiva S. Mitochondrial electron transport chain: Oxidative phosphorylation, oxidant production, and methods of measurement. Redox Biol. 2020;37:101674.

[123] Feichtinger RG, Neureiter D, Skaria T, Wessler S, Cover TL, Mayr JA, et al. Oxidative Phosphorylation System in Gastric Carcinomas and Gastritis. Oxid Med Cell Longev. 2017;2017:1320241.

[124] Lee SH, Lee JH, Lee HY, Min KJ. Sirtuin signaling in cellular senescence and aging. BMB Rep. 2019;52(1):24-34.

[125] Costa-Machado LF, Fernandez-Marcos PJ. The sirtuin family in cancer. Cell Cycle. 2019;18(18):2164-96.

[126] Wek RC, Jiang HY, Anthony TG. Coping with stress: eIF2 kinases and translational control. Biochem Soc Trans. 2006;34(Pt 1):7-11.

[127] Chen X, Cubillos-Ruiz JR. Endoplasmic reticulum stress signals in the tumour and its microenvironment. Nat Rev Cancer. 2021;21(2):71-88.

[128] Dinarello A, Licciardello G, Fontana CM, Tiso N, Argenton F, Dalla Valle L. Glucocorticoid receptor activities in the zebrafish model: a review. J Endocrinol. 2020;247(3):R63-82.

[129] Noti M, Sidler D, Brunner T. Extra-adrenal glucocorticoid synthesis in the intestinal epithelium: more than a drop in the ocean? Semin Immunopathol. 2009;31(2):237-48.

[130] Daujat-Chavanieu M, Gerbal-Chaloin S. Regulation of CAR and PXR Expression in Health and Disease. Cells. 2020;9(11).

[131] Xu C, Li CY, Kong AN. Induction of phase I, II and III drug metabolism/transport by xenobiotics. Arch Pharm Res. 2005;28(3):249-68.

[132] Yang H, Wang H. Signaling control of the constitutive androstane receptor (CAR). Protein Cell. 2014;5(2):113-23.

[133] Hempel N, Gamage N, Martin JL, McManus ME. Human cytosolic sulfotransferase SULT1A1. Int J Biochem Cell Biol. 2007;39(4):685-9.

[134] Glatt H, Boeing H, Engelke CE, Ma L, Kuhlow A, Pabel U, et al. Human cytosolic sulphotransferases: genetics, characteristics, toxicological aspects. Mutat Res. 2001;482(1-2):27-40.

[135] Boccia S, Persiani R, La Torre G, Rausei S, Arzani D, Gianfagna F, et al. Sulfotransferase 1A1 polymorphism and gastric cancer risk: a pilot case-control study. Cancer Lett. 2005;229(2):235-43.

[136] Boccia S, Sayed-Tabatabaei FA, Persiani R, Gianfagna F, Rausei S, Arzani D, et al. Polymorphisms in metabolic genes, their combination and interaction with tobacco smoke and alcohol consumption and risk of gastric cancer: a case-control study in an Italian population. BMC Cancer. 2007;7:206.

[137] Du L, Lei L, Zhao X, He H, Chen E, Dong J, et al. The Interaction of Smoking with Gene Polymorphisms on Four Digestive Cancers: A Systematic Review and Meta-Analysis. J Cancer. 2018;9(8):1506-17.

[138] Loh M, Koh KX, Yeo BH, Song CM, Chia KS, Zhu F, et al. Meta-analysis of genetic polymorphisms and gastric cancer risk: variability in associations according to race. Eur J Cancer. 2009;45(14):2562-8.

[139] Davidson NO, Hausman AM, Ifkovits CA, Buse JB, Gould GW, Burant CF, et al. Human intestinal glucose transporter expression and localization of GLUT5. Am J Physiol. 1992;262(3 Pt 1):795-800.

[140] Galand G. Brush border membrane sucrase-isomaltase, maltase-glucoamylase and trehalase in mammals. Comparative development, effects of glucocorticoids, molecular mechanisms, and phylogenetic implications. Comp Biochem Physiol B. 1989;94(1):1-11.

[141] Jiang G, Sladek FM. The DNA binding domain of hepatocyte nuclear factor 4 mediates cooperative, specific binding to DNA and heterodimerization with the retinoid X receptor alpha. J Biol Chem. 1997;272(2):1218-25.

[142] Cereghini S. Liver-enriched transcription factors and hepatocyte differentiation. FASEB J. 1996;10(2):267-82.

[143] Lau HH, Ng NHJ, Loo LSW, Jasmen JB, Teo AKK. The molecular functions of hepatocyte nuclear factors - In and beyond the liver. J Hepatol. 2018;68(5):1033-48.

[144] Jonckheere N, Vincent A, Franquet-Ansart H, Witte-Bouma J, Korteland-van Male A, Leteurtre E, et al. GATA-4/-6 and HNF-1/-4 families of transcription factors control the transcriptional regulation of the murine Muc5ac mucin during stomach development and in epithelial cancer cells. Biochim Biophys Acta. 2012;1819(8):869-76.

[145] Zhang E, Huang X, He J. Integrated bioinformatic analysis of HNF1A in human cancers. J Int Med Res. 2021;49(3):300060521997326.

[146] Saad DZ, Sidhom KF, Gadallah MF, Samir NA, Shakweer MM. Diagnostic utility of the combined use of HNF4A and GATA3 in distinction between primary and metastatic breast and gastric carcinomas. APMIS. 2021;129(9):548-55.

[147] Ma L, Jüttner M, Kullak-Ublick GA, Eloranta JJ. Regulation of the gene encoding the intestinal bile acid transporter ASBT by the caudal-type homeobox proteins CDX1 and CDX2. Am J Physiol Gastrointest Liver Physiol. 2012;302(1):G123-33.

[148] Barros R, da Costa LT, Pinto-de Sousa J, Duluc I, Freund JN, David L, et al. CDX2 autoregulation in human intestinal metaplasia of the stomach: impact on the stability of the phenotype. Gut. 2011;60(3):290-8.

[149] Saandi T, Baraille F, Derbal-Wolfrom L, Cattin AL, Benahmed F, Martin E, et al. Regulation of the tumor suppressor homeogene Cdx2 by HNF4α in intestinal cancer. Oncogene. 2013;32(32):3782-8.

[150] San Roman AK, Aronson BE, Krasinski SD, Shivdasani RA, Verzi MP. Transcription factors GATA4 and HNF4A control distinct aspects of intestinal homeostasis in conjunction with transcription factor CDX2. J Biol Chem. 2015;290(3):1850-60.

[151] Chen L, Toke NH, Luo S, Vasoya RP, Fullem RL, Parthasarathy A, et al. A reinforcing HNF4-SMAD4 feed-forward module stabilizes enterocyte identity. Nat Genet. 2019;51(5):777-85.

[152] Rodríguez-Seguel E, Villamayor L, Arroyo N, De Andrés MP, Real FX, Martín F, et al. Loss of GATA4 causes ectopic pancreas in the stomach. J Pathol. 2020;250(4):362-73.

[153] DeLaForest A, Kohlnhofer BM, Franklin OD, Stavniichuk R, Thompson CA, Pulakanti K, et al. GATA4 Controls Epithelial Morphogenesis in the Developing Stomach to Promote Establishment of Glandular Columnar Epithelium. Cell Mol Gastroenterol Hepatol. 2021;12(4):1391-413.

[154] Ma WW, Zhao L, Yuan LH, Yu HL, Wang H, Gong XY, et al. Elaidic acid induces cell apoptosis through induction of ROS accumulation and endoplasmic reticulum stress in SH-SY5Y cells. Mol Med Rep. 2017;16(6):9337-46.

[155] Ohmori H, Fujii K, Kadochi Y, Mori S, Nishiguchi Y, Fujiwara R, et al. Elaidic Acid, a Trans-Fatty Acid, Enhances the Metastasis of Colorectal Cancer Cells. Pathobiology. 2017;84(3):144-51.

[156] Liang D, Zhuo Y, Guo Z, He L, Wang X, He Y, et al. SIRT1/PGC-1 pathway activation triggers autophagy/mitophagy and attenuates oxidative damage in intestinal epithelial cells. Biochimie. 2020;170:10-20.

[157] Johnson LR. Functional development of the stomach. Annu Rev Physiol. 1985;47:199-215.

[158] Date S, Sato T. Mini-gut organoids: reconstitution of the stem cell niche. Annu Rev Cell Dev Biol. 2015;31:269-89.

[159] Abrahao-Machado LF, Scapulatempo-Neto C. HER2 testing in gastric cancer: An update. World J Gastroenterol. 2016;22(19):4619-25.

[160] Escobar-Morreale HF, Botella-Carretero JI, Morreale de Escobar G. Treatment of hypothyroidism with levothyroxine or a combination of levothyroxine plus L-triiodothyronine. Best Pract Res Clin Endocrinol Metab. 2015;29(1):57-75.

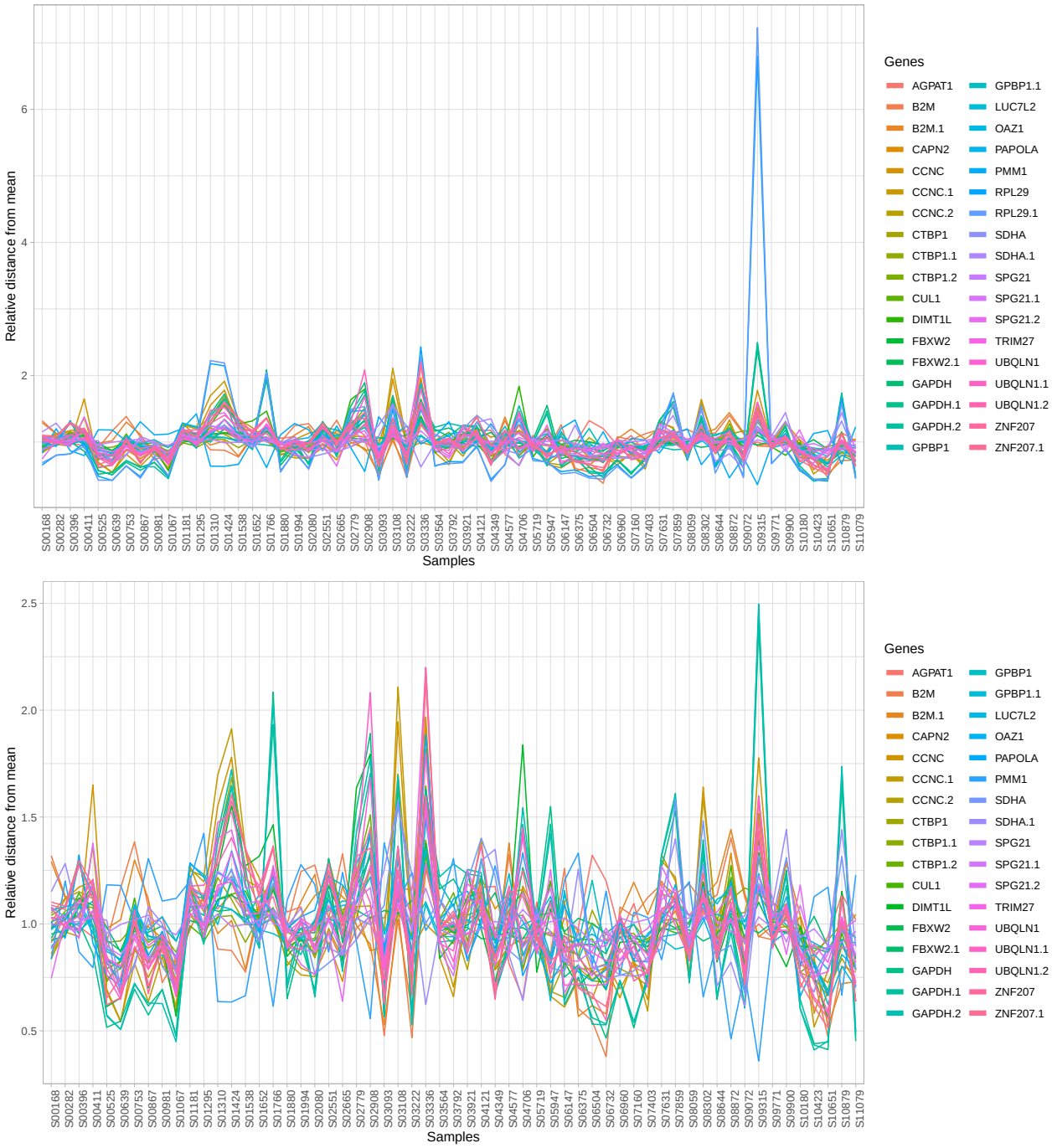
[161] Goes LG, da Luz Eltchechem C, Wouk J, Malfatti CRM, da Silva LA. Relationship Between Hormonal Mechanisms of Diabetes Mellitus and Hypothyroidism Post-Bariatric Surgery. Curr Diabetes Rev. 2020;16(3):200-3.

[162] Bose S, Le A. Glucose Metabolism in Cancer. Adv Exp Med Biol. 2018;1063:3-12.

[163] Pajak B, Siwiak E, Sołtyka M, Priebe A, Zieliński R, Fokt I, et al. 2-Deoxy-d-Glucose and Its Analogs: From Diagnostic to Therapeutic Agents. Int J Mol Sci. 2019;21(1).

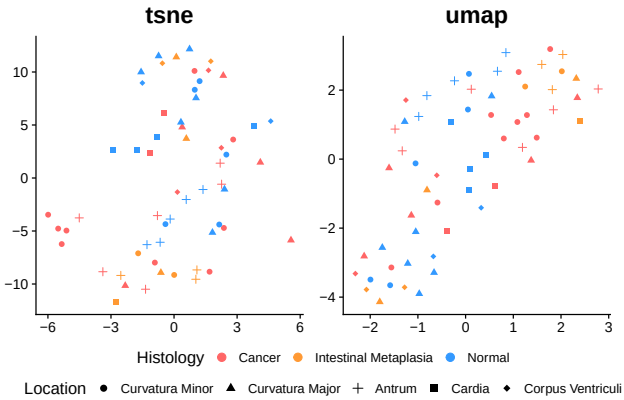
Appendix

A Relative expression of housekeeping gene from gene mean

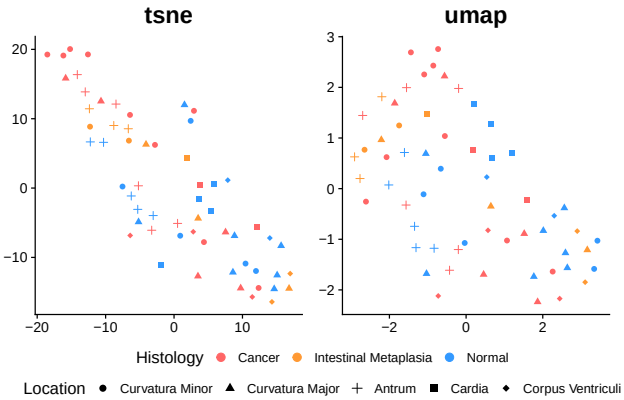


B Dimensional reduction of tissue transcriptome

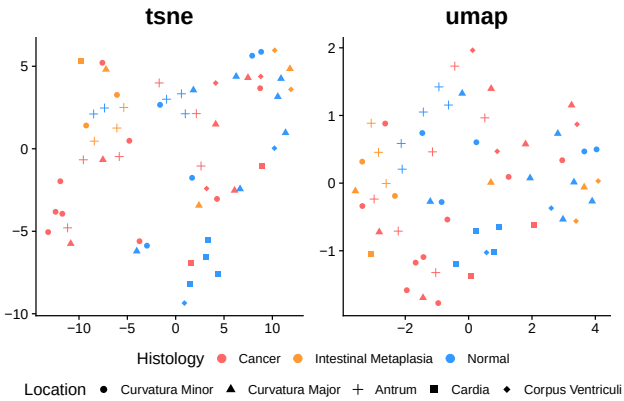
Tissue; Housekeeping gene normalization



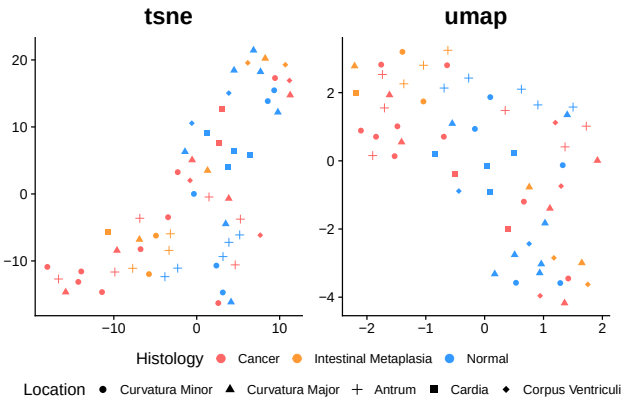
Tissue; Housekeeping gene norm. and log2 trans.



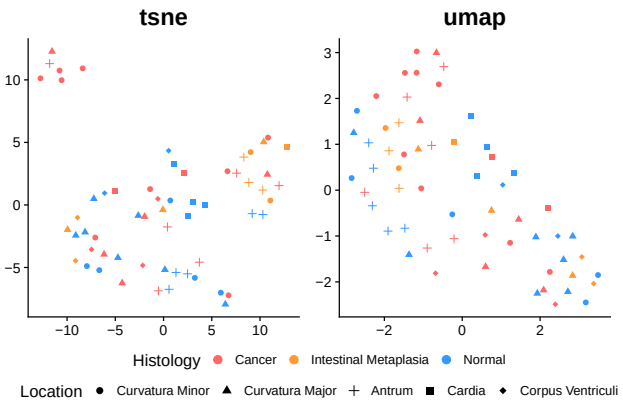
Tissue; Quantile normalization and log2 trans.



Tissue; Geometric mean normalization

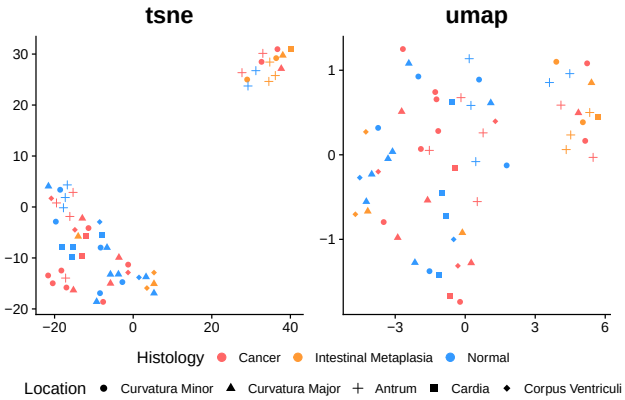


Tissue; Geometric mean normalization and log2 trans.

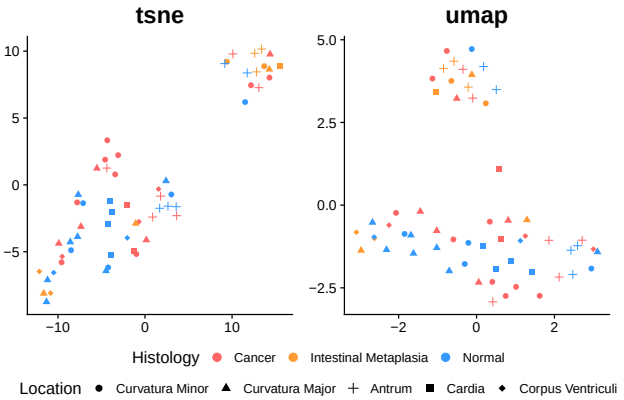


C Dimensional reduction of enterocyte genes

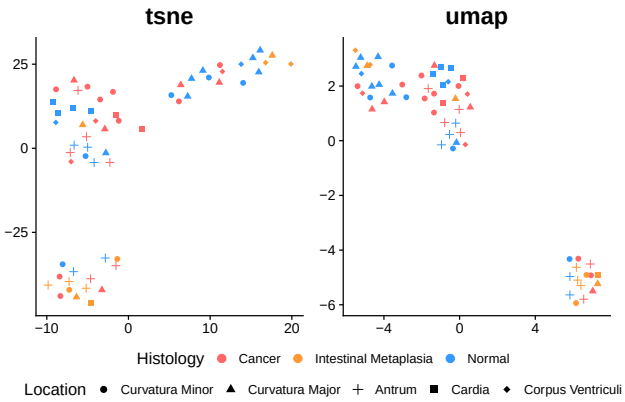
Enterocyte; Housekeeping gene normalization



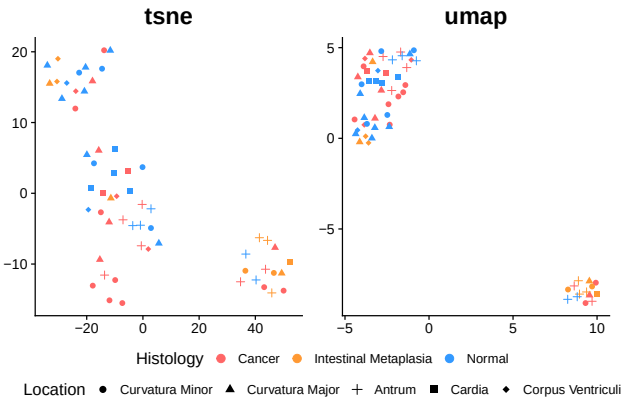
Enterocyte; Housekeeping gene norm. and log2 trans.



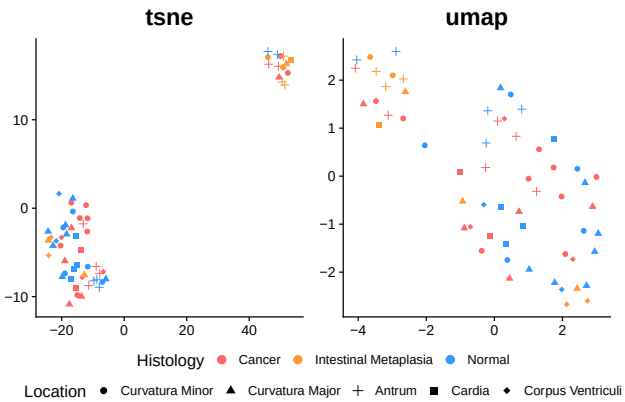
Enterocyte; Quantile normalization and log2 trans.



Enterocyte; Geometric mean normalization

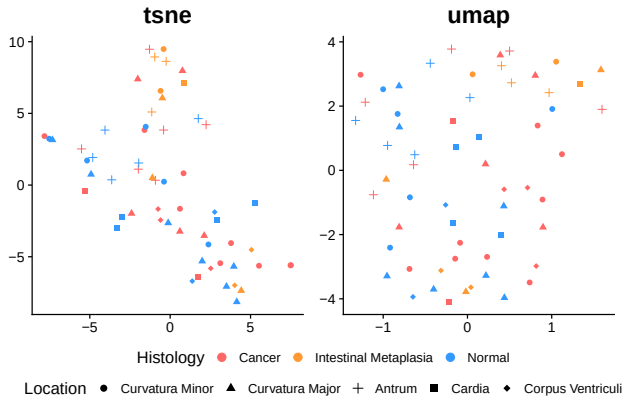


Enterocyte; Geometric mean normalization and log2 trans.

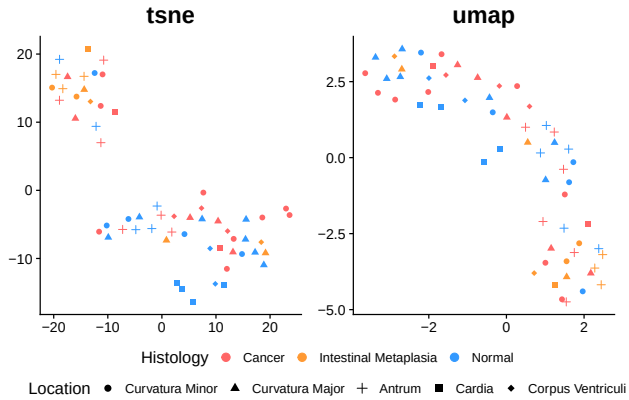


D Dimensional reduction of goblet cell genes

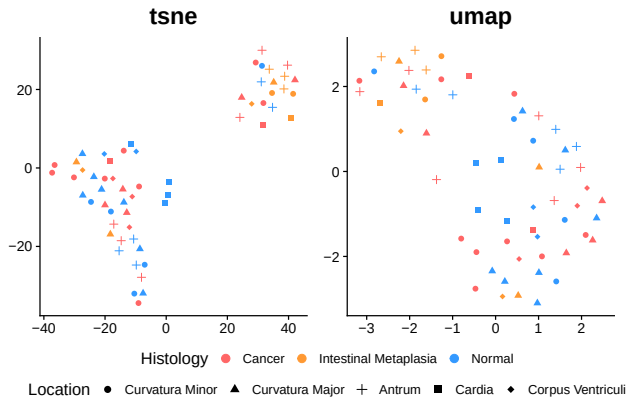
Goblet_cell; Housekeeping gene normalization



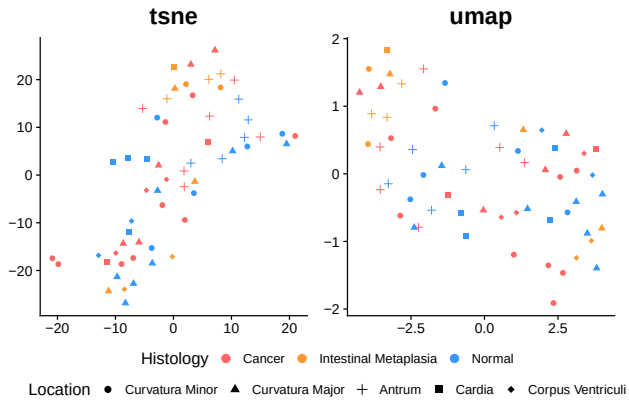
Goblet_cell; Housekeeping gene norm. and log2 trans.



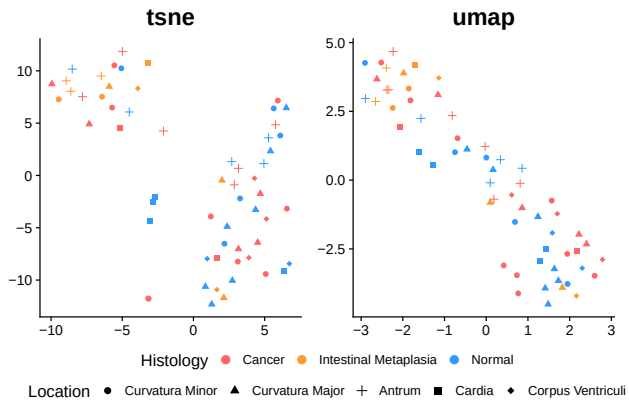
Goblet_cell; Quantile normalization and log2 trans.



Goblet_cell; Geometric mean normalization

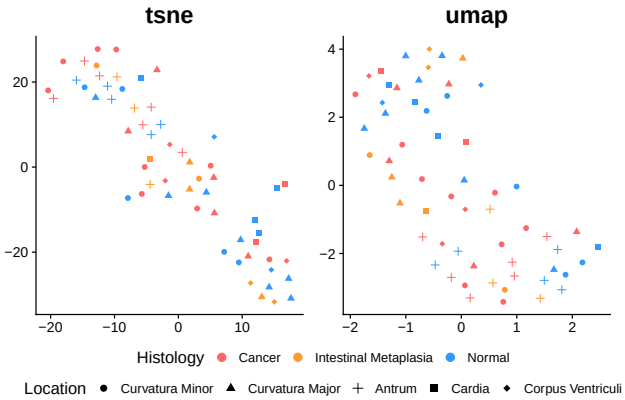


Goblet_cell; Geometric mean normalization and log2 trans.

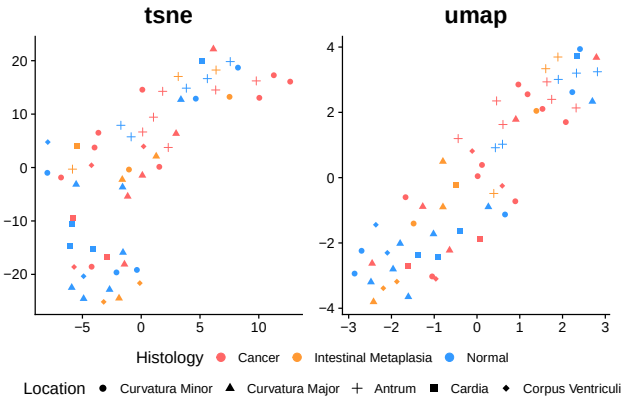


E Dimensional reduction of chief cell genes

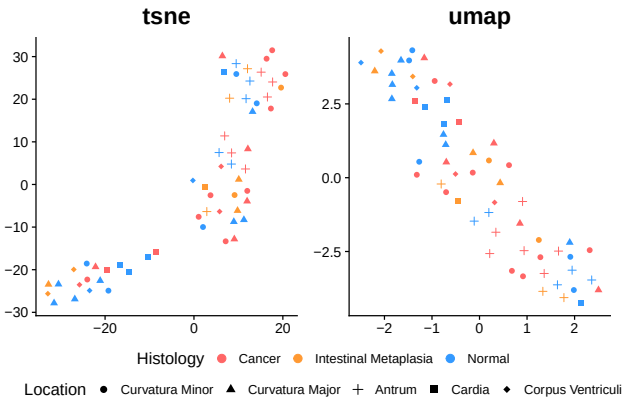
Chief_cell; Housekeeping gene normalization



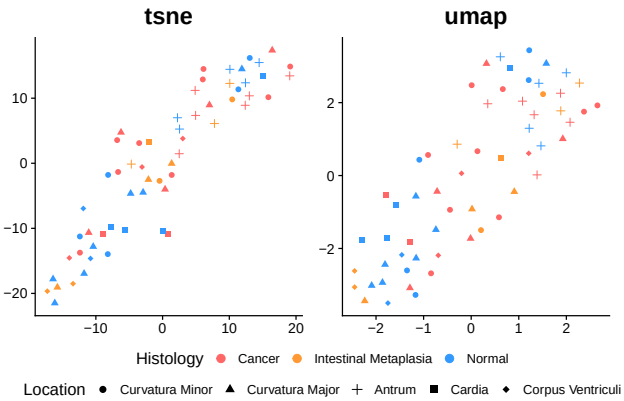
Chief_cell; Housekeeping gene norm. and log2 trans.



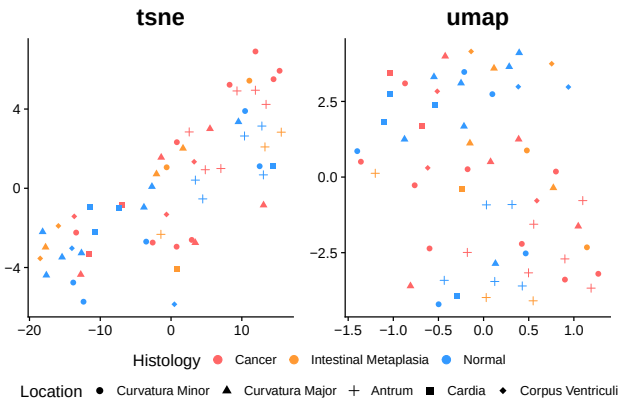
Chief_cell; Quantile normalization and log2 trans.



Chief_cell; Geometric mean normalization

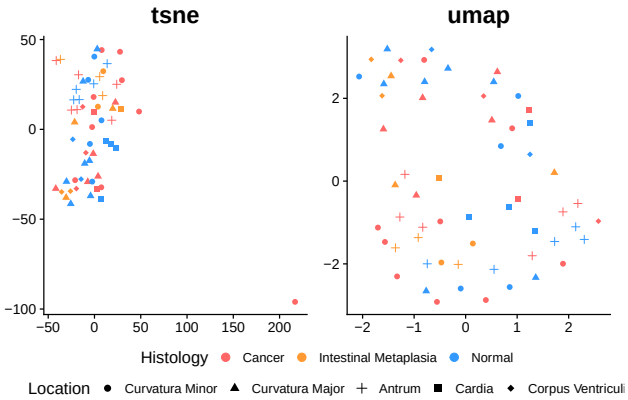


Chief_cell; Geometric mean normalization and log2 trans.

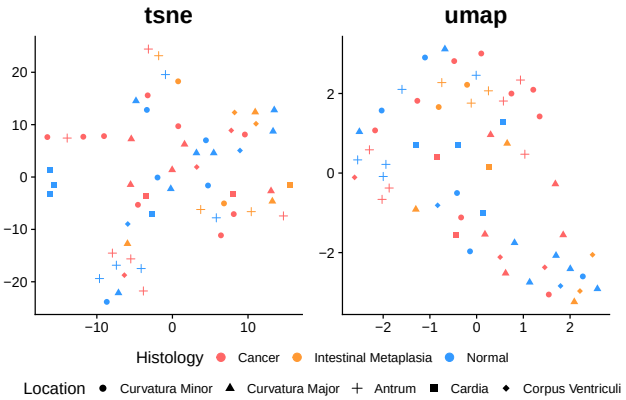


F Dimensional reduction of cancer cell genes

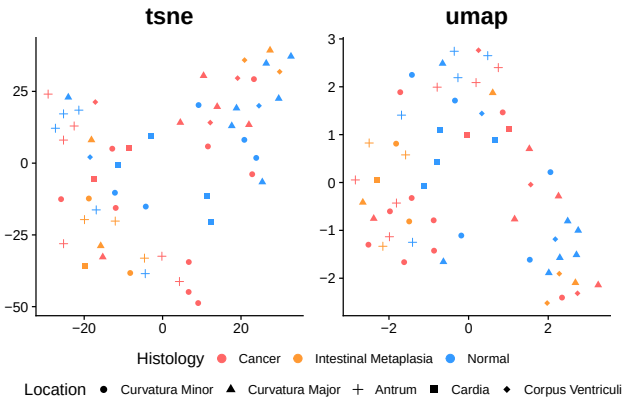
Cancer_cell; Housekeeping gene normalization



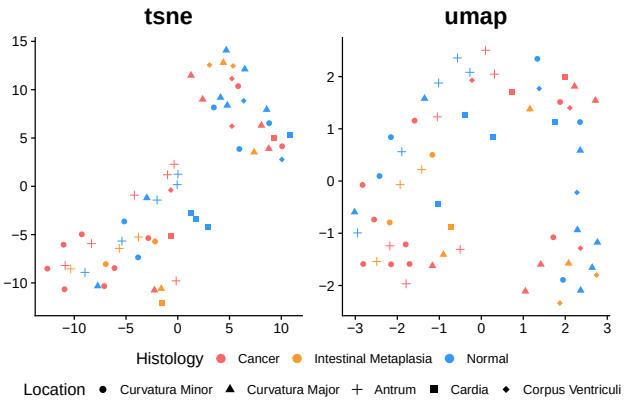
Cancer_cell; Housekeeping gene norm. and log2 trans.



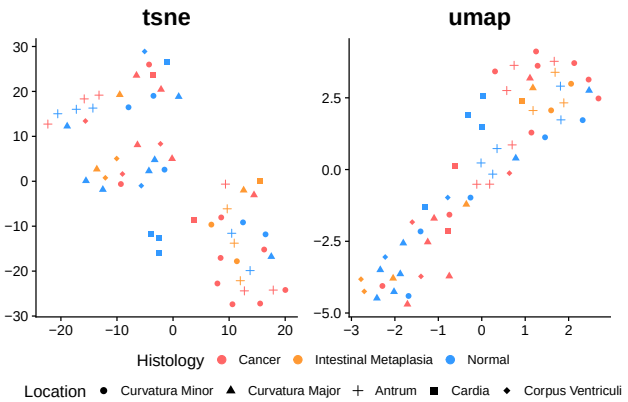
Cancer_cell; Quantile normalization and log2 trans.



Cancer_cell; Geometric mean normalization



Cancer_cell; Geometric mean normalization and log2 trans.

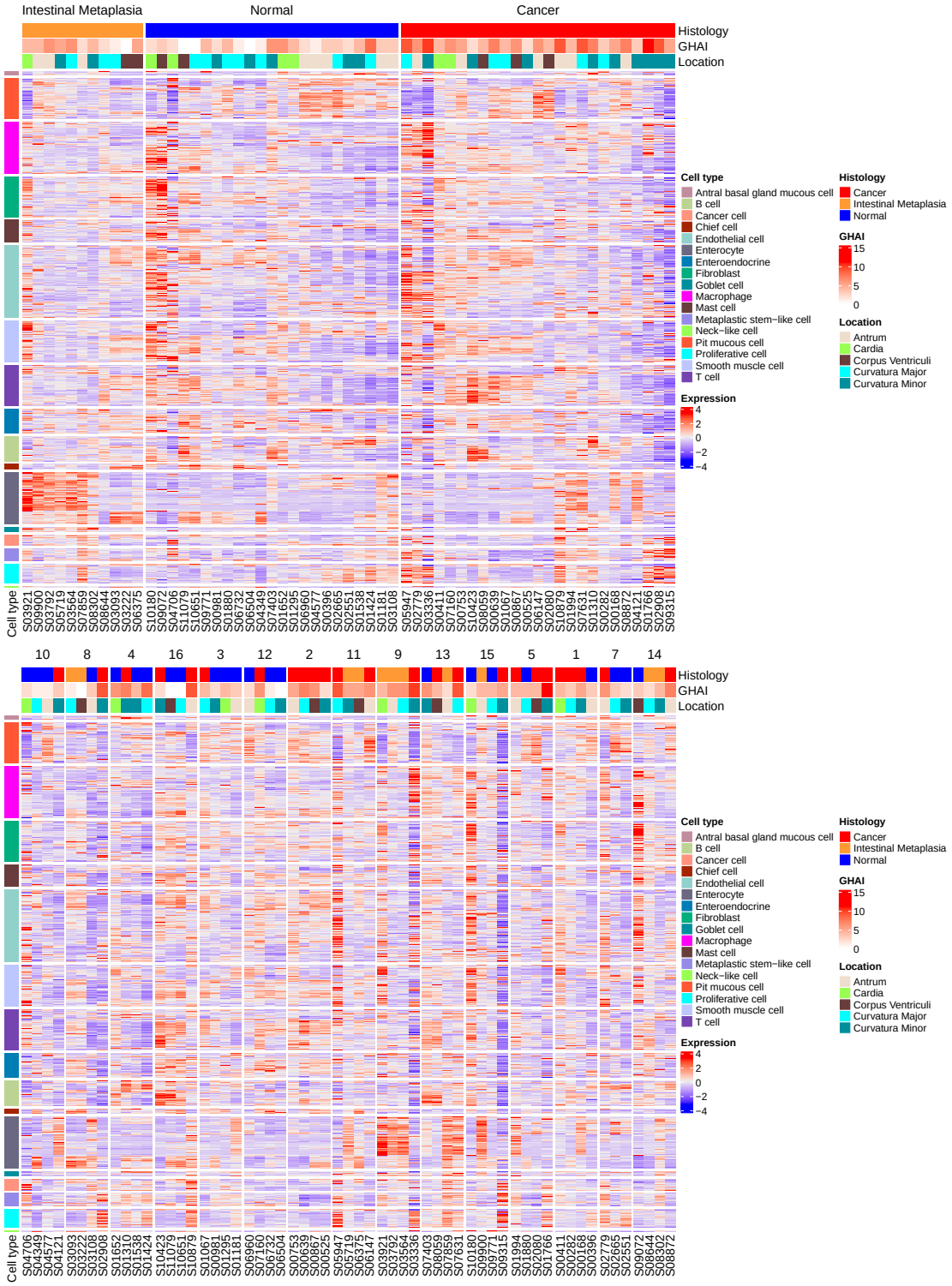


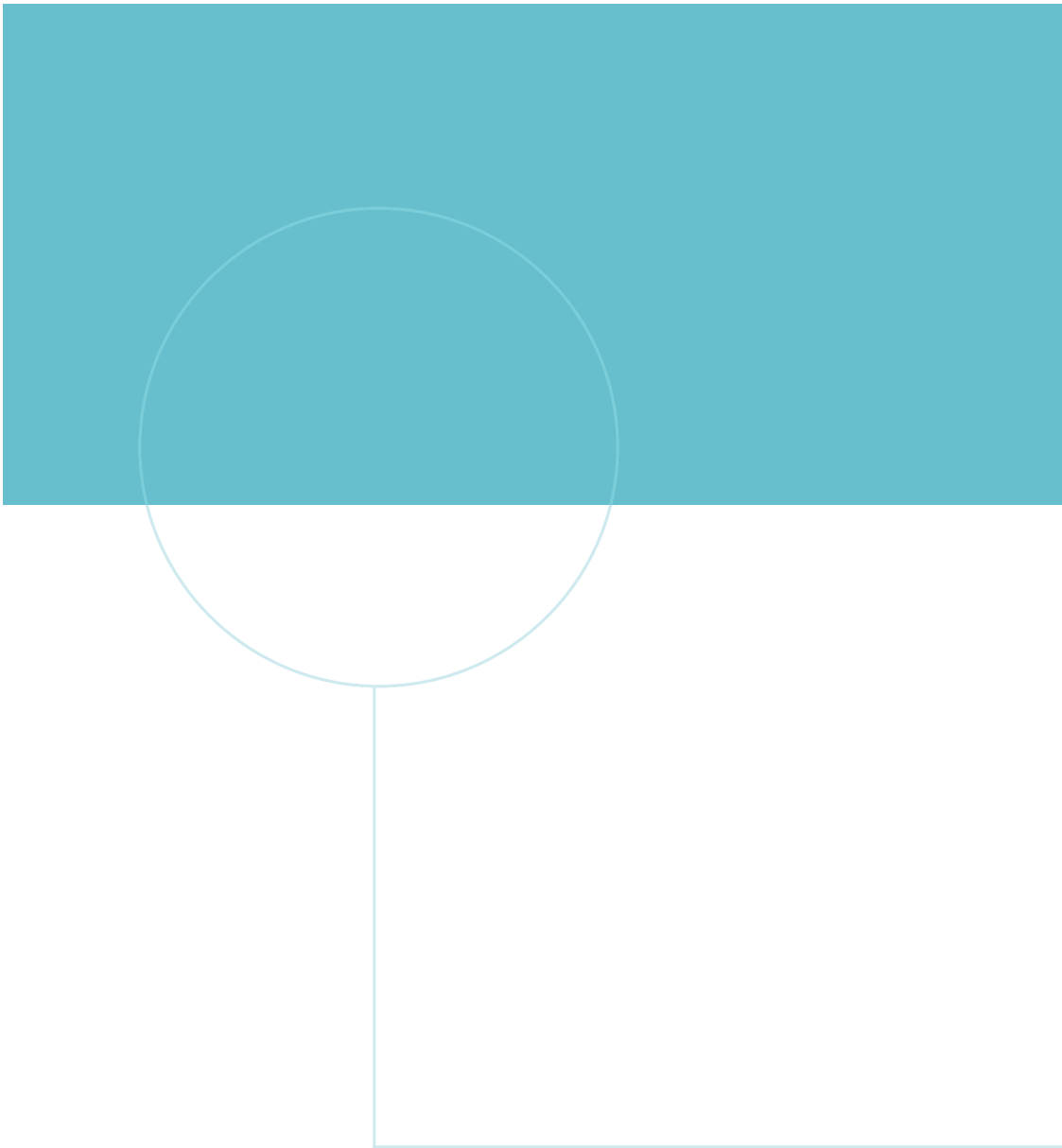
G Classification of gastric samples

Patient	Sample location	Sample ID	Inflammation (0-4)	Epithelial defects (0-4)	Oxynic atrophy (0-4)	Epithelial hyperplasia (0-4)	Pseudopyloric metaplasia (0-4)	GHAI	Intestinal metaplasia	Tumor sample	Biopsy evaluation	Pathological evaluation
1	Curvatura Minor	S00168	2	1	2	0	0	5	2	1 Cancer		
1	Curvatura Major	S00282	2	1	0	0	0	3	0	2 Cancer?		
1	Antrum	S00396	1	1	0	1	0	3	0	4 Cancer?		
1	Cardia	S00411	2	1	0	1	0	4	1	1 Cancer		
2	Curvatura Minor	S00525	1	0	0	1	0	2	0	1 Cancer	Cancer	
2	Curvatura Major	S00639	2	1	3	1	0	7	0	1 Cancer		
2	Antrum	S00753	1	2	0	2	0	5	0	1 Cancer		
2	Corpus Ventriculi	S00867	3	1	1	1	0	6	0	1 Cancer		
3	Curvatura Minor	S00981	0	1	1	0	0	2	0	5 No Tumor		
3	Curvatura Major	S01067	1	0	2	1	0	4	0	1 Cancer	Cancer	
3	Antrum	S01181	1	0	1	1	0	3	0	5 No Tumor		
3	Cardia	S01295	1	0	1	2	0	4	0	5 No Tumor		
4	Curvatura Minor	S01310	2	1	2	2	0	7	0	1 Cancer		
4	Curvatura Major	S01424	2	1	2	2	1	8	0	5 No Tumor		
4	**Curvitura major***	S01538	1	0	0	2	1	4	0	5 No Tumor		
4	Cardia	S01652	2	0	2	1	1	6	0	5 No Tumor		
5	Curvatura Minor	S01766	3	3	2	1	2	11	0	1 Cancer	Cancer	
5	Curvatura Major	S01880	1	0	1	1	1	4	0	3 Cancer?		
5	Antrum	S01994	2	2	0	0	0	4	0	1 Cancer	Cancer	
5	Corpus Ventriculi	S02080	1	0	2	1	0	4	0	2 Cancer?		
7	Curvatura Minor	S02551	1	0	0	1	0	2	0	5 No Tumor		
7	Curvatura Major	S02665	1	0	0	2	0	3	0	5 No Tumor		
7	Antrum	S02779	2	2	0	2	0	6	0	1 Cancer	Cancer	
8	Curvatura Minor	S02908	3	3	0	3	0	9	0	1 Cancer	Cancer	
8	Curvatura Major	S03093	1	0	0	0	0	1	2	5 No Tumor		
8	Antrum	S03108	1	0	0	2	0	3	0	5 No Tumor		
8	Corpus Ventriculi	S03222	0	0	0	0	0	0	3	5 No Tumor		
9	Curvatura Minor	S03336	3	2	3	2	0	10	3	1 Cancer	Cancer	
9	Curvatura Major	S03564	2	0	3	1	0	6	3	5 No Tumor		
9	Antrum	S03792	2	0	3	1	0	6	2	5 No Tumor		
9	Cardia	S03921	1	0	2	1	0	4	1	5 No Tumor		
10	Curvatura Minor	S04121	1	0	1	1	0	3	0	1 Cancer		
10	Curvatura Major	S04349	1	0	0	0	0	1	0	5 No Tumor		
10	Antrum	S04577	1	0	0	0	0	1	0	5 No Tumor		
10	Cardia	S04706	1	0	1	0	0	2	0			
11	Curvatura Minor	S05719	2	1	1	1	0	5	1	5 No Tumor		
11	Curvatura Major	S05947	3	3	3	0	0	9	0	1 Cancer	Cancer	
11	Antrum	S06147	2	1	2	1	0	6	0	5 No Tumor	Cancer	
11	Corpus Ventriculi	S06375	1	1	2	1	0	5	2	5 No Tumor		
12	Curvatura Minor	S06504	0	0	0	0	0	0	0	5 No Tumor		
12	Curvatura Major	S06732	0	1	0	0	0	1	0	5 No Tumor		
12	Antrum	S06960	0	1	0	1	0	2	0	5 No Tumor		
12	Cardia	S07160	2	1	1	2	0	6	0	1 Cancer		
13	Curvatura Minor	S07403	2	1	2	1	0	6	0	5 No Tumor		
13	Curvatura Major	S07631	3	3	3	0	0	9	0	1 Cancer		
13	Antrum	S07859	1	0	0	1	0	2	0	5 No Tumor		
13	Corpus Ventriculi	S08059	2	0	2	0	0	4	0	5 No Tumor	Cancer	
14	Curvatura Minor	S08302	1	1	0	2	0	4	3	5 No Tumor		
14	Curvatura Major	S08644	1	0	0	1	0	2	1	5 No Tumor		
14	Antrum	S08872	2	1	0	4	0	7	1	1 Cancer		
14	Corpus Ventriculi	S09072	1	0	0	2	0	3	0	5 No Tumor		
15	Curvatura Minor	S09315	1	4	0	0	0	5	0		Cancer	
15	Curvatura Major	S09771	2	1	1	0	0	4	0	5 No Tumor		
15	Antrum	S09900	2	0	0	1	0	4	1	5 No Tumor		
15	Cardia	S10180	1	0	1	0	0	2	0	5 No Tumor		
16	Curvatura Minor	S10423	1	0	1	0	0	2	0	2 Cancer?		
16	Curvatura Major	S10651	0	0	0	0	0	0	0	5 No Tumor		
16	Antrum	S10879	2	2	2	3	0	9	2	1 Cancer	Cancer	
16	Corpus Ventriculi	S11079	0	0	0	0	0	0	0			

*Prior Bil distal gastrectomy

H Heatmaps normalized through housekeeping genes





NTNU

Kunnskap for en bedre verden