

# RNA-SEQ DATA ANALYSIS

Example project report



# CONTENTS

Introduction
Methods5
Data5
Quality Control5
Alignment5
Differentially Expressed Genes5
Enrichment Analysis6
Results7
Quality Control7
Alignment8
Principal Component Analysis and Pearson's Correlation Heatmap11
Differentially Expressed Genes12
Enrichment Analysis
Deliverables
References



# INTRODUCTION

The customer's research group is studying transcriptomic features of prostate cancer (PC). The team has produced RNA-sequencing data of prostate tissues (45 samples) presented in **Table 1**. There are measurements from eight non-cancerous benign prostate hyperplasia specimens (BPH), 16 localised PC specimens, nine advanced PC specimens and 12 castration-resistant prostate cancer (CRPC) specimens.

In this project, the gene expression patterns of each progression step of the PC was compared against a preceding progression step, leading into three comparisons: 1) Localised PC samples vs. BPH samples, 2) Advanced PC samples vs. Localised PC samples and 3) CRPC samples vs. Advanced PC samples. After finding differentially expressed genes, functional analysis of the gene groups was performed by assessing the enrichment of biological processes and signaling pathways.

Sample number	Tissue type	Progression step	Replicate	Sample name
1	normal	BPH	1	BPH 1
2	normal	BPH	2	BPH 2
3	normal	BPH	3	BPH 3
4	normal	BPH	4	BPH 4
5	normal	BPH	5	BPH 5
6	normal	BPH	6	BPH 6
7	normal	BPH	7	BPH 7
8	normal	BPH	8	BPH 8
9	tumor	Localized PC	1	Localized PC 1
10	tumor	Localized PC	2	Localized PC 2
11	tumor	Localized PC	3	Localized PC 3
12	tumor	Localized PC	4	Localized PC 4
13	tumor	Localized PC	5	Localized PC 5
14	tumor	Localized PC	6	Localized PC 6
15	tumor	Localized PC	7	Localized PC 7
16	tumor	Localized PC	8	Localized PC 8
17	tumor	Localized PC	9	Localized PC 9
18	tumor	Localized PC	10	Localized PC 10

#### Table 1: Samples used in the analyses





Sample number	Tissue type	Progression step	Replicate	Sample name
19	tumor	Localized PC	11	Localized PC 11
20	tumor	Localized PC	12	Localized PC 12
21	tumor	Localized PC	13	Localized PC 13
22	tumor	Localized PC	14	Localized PC 14
23	tumor	Localized PC	15	Localized PC 15
24	tumor	Localized PC	16	Localized PC 16
25	tumor	Advanced PC	1	Advanced PC 1
26	tumor	Advanced PC	2	Advanced PC 2
27	tumor	Advanced PC	3	Advanced PC 3
28	tumor	Advanced PC	4	Advanced PC 4
29	tumor	Advanced PC	5	Advanced PC 5
30	tumor	Advanced PC	6	Advanced PC 6
31	tumor	Advanced PC	7	Advanced PC 7
32	tumor	Advanced PC	8	Advanced PC 8
33	tumor	Advanced PC	9	Advanced PC 9
34	tumor	CRPC	1	CRPC 1
35	tumor	CRPC	2	CRPC 2
36	tumor	CRPC	3	CRPC 3
37	tumor	CRPC	4	CRPC 4
38	tumor	CRPC	5	CRPC 5
39	tumor	CRPC	6	CRPC 6
40	tumor	CRPC	7	CRPC 7
41	tumor	CRPC	8	CRPC 8
42	tumor	CRPC	9	CRPC 9
43	tumor	CRPC	10	CRPC 10
44	tumor	CRPC	11	CRPC 11
45	tumor	CRPC	12	CRPC 12



# METHODS

### DATA

The RNA-seq data had been deposited into the Gene Expression Omnibus (GEO) [Edgar et al., 2002], accession ID GSE80609. The data was delivered in SRA format and was converted into paired-end fastq read files using SRA-toolkit, v. 2.9.0 [SRA Knowledge Base, 2011].

### QUALITY CONTROL

Quality of the RNA-seq reads was inspected using FastQC software, v. 0.11.7 [Andrews, 2010]. General statistics table of the FastQC analysis results from all samples were generated using R package fastqcr [Kassambara, 2017]. TrimGalore!, v. 0.4.5 [Krueger, 2017] was run to all samples with default settings to remove low quality bases with Phred quality score less than 20 as well as reads whose length was less than 20 bp after the trimming. The quality analysis with FASTQC and fastqcr was repeated for the trimmed fastq files.

### ALIGNMENT

RNA-seq reads were aligned to the GRCh37.75 reference genome using STAR aligner, v. 2.6.0c [Dobin et al., 2013]. Gene-level read counts were obtained simultaneously with the alignment process. Expression matrix generation and all further analyses were performed with R version 3.5.0 [R Core Team, 2018].

For visual exploration of the data, the obtained read counts were normalized using regularized log transformation function of DESeq2 R package, v. 1.20.0 [Love et al., 2014], which transforms the count data to the log2 scale in a way that minimizes differences between samples for rows with small counts and also normalizes the data with respect to library size. Visual inspection of the samples was done using principal component analysis (PCA) and a Pearson's correlation heat map using DESeq2 functions [Love et al., 2014] and R package pheatmap [Kolde, 2015], respectively.

### DIFFERENTIALLY EXPRESSED GENES

Data normalisation and differential expression analysis were performed using R package DESeq2 [Love et al., 2014] by making the following contrasts: 1) Localised PC samples vs. all BPH samples, 2) Advanced PC samples vs. Localized PC samples, and 3) CRPC samples vs. Advanced PC samples. Thresholds for statistical and biological significance were set to adjusted p-value < 0.05 and at least 2-fold up- or down-

regulation in expression, respectively, and a gene was considered significantly differentially expressed if both of these conditions were met. The resulting gene lists were visualised using R-package VennDiagram [Chen, 2018.]

### ENRICHMENT ANALYSIS

Biological Process Gene Ontology (GO BP) [Ashburner et al., 2000; GO Consortium, 2017] term and Kyoto Encyclopedia of Genes and Genomes (KEGG) [Kanehisa and Goto, 2000] pathway term overrepresentation analysis was performed using R package clusterProfiler, v. 3.8.1 [Yu et al., 2012]. The analysis determined whether any terms are annotated to a list of specified genes, in this case a list of differentially expressed (DE) genes, at a frequency greater than what would be expected by chance, and calculated a p-value using the hypergeometric distribution. The minimum number of DE genes required to be annotated by a given ontology term was set to 2. The p-values of enrichment analysis were corrected for multiple testing using Benjamini-Hochberg multiple testing adjustment procedure [Benjamini & Hochberg, 1995]. Thresholds for adjusted p-value and q-value of the enrichment were set to default values 0.05 and 0.2, respectively.

REACTOME Pathway over-representation analysis was performed using R-package ReactomePA, v. 1.24.0 [Yu & He, 2016]. The same thresholds for statistical significance were used as described above.

Enriched GO BP terms and REACTOME Pathway terms were visualized using clusterProfiler functions. KEGG enrichment results were visualised using R-package pathview, v. 1.20.0 [Luo & Brouwer, 2013].



# RESULTS

### QUALITY CONTROL

The quality of the raw read data was inspected using FastQC software [Andrews, 2010]. A few samples contained some low-quality reads (**Figure 1 A**) marked by FastQC-modules "Per base sequence quality" and "Per base sequence quality scores". After quality trimming, the quality of all files was sufficient for further analysis (**Figure 1 B**). The summary of quality statistics for each paired fastq file is provided as a supplementary table. **Table 2** describes the contents of these tables. From the reports it could be seen that some of the samples had relatively high duplication rate and GC-content, but this is not unusual for R N A - s e q data. Sequencing adapter contamination was not seen in the samples.



*Figure 1: FASTQC sequence quality score graphs before (A) and after (B) trimming. X-axis shows the base position in read (bp) and y-axis the base quality score. The central red line is the median value in each position and the yellow box represents the interquartile range (25%-75%). The blue line shows the mean quality across the positions. The scores in green area (Phred score > 28) are considered as good quality scores. After trimming, the mean quality of the bases in all positions is corrected.* 



#### Table 2: Description of the summary table of quality statistics

Column	Description
sample	Name of the fastq file: "SAMPLENAME_X" where X refers to the paired-read file (1 for pair 1 and 2 for pair 2)
pct.dup	Duplication rate (%) of the reads before and after trimming
pct.qc	GC-content (%) of the reads before and after trimming
tot.seq	Total amount of sequences I.e. library size before and after trimming
seq.length	Sequence length before and after trimming
pct.seq.removed	Rate of the reads (%) removed by trimming

### ALIGNMENT

The alignment statistics are presented in **Table 3**. The alignment rate of the samples varied between 85,1–94,7 %, being over 90 % in almost all samples, which can be considered a good alignment result. Thus there was no need for further preprocessing or removal of samples before the downstream analysis steps.

Table 3: Alignment statistics of the samples showing the total number or read pairs in the samples, the average length of the mapped reads and percentages of reads mapping to unique positions, reads mapping to multiple positions and reads that could not be mapped to the reference.

Sample name	No of read pairs	Mapped length	% Uniquely mapped	% Multi-mapped	% Unmapped
BPH 1	10471280	186	93.50%	4.30%	2.10%
BPH 2	9982546	187	93.00%	4.60%	2.30%
BPH 3	17164077	185	93.10%	4.50%	2.40%
BPH 4	14129951	186	93.90%	3.80%	2.20%
BPH 5	24554330	185	93.10%	4.40%	2.40%
BPH 6	31547235	184	93.30%	4.40%	2.30%
BPH 7	13605180	185	92.50%	5.20%	2.30%



Sample name	No of read pairs	Mapped length	% Uniquely mapped	% Multi-mapped	% Unmapped
BPH 8	11890374	184	92.60%	4.80%	2.50%
Localized PC 1	11111986	183	93.50%	3.90%	2.60%
Localized PC 2	12467958	183	93.20%	4.40%	2.30%
Localized PC 3	12267792	184	93.60%	4.30%	2.10%
Localized PC 4	14683939	183	93.20%	4.10%	2.70%
Localized PC 5	22117528	183	93.70%	3.50%	2.70%
Localized PC 6	12359696	184	93.30%	4.20%	2.50%
Localized PC 7	16340233	185	92.80%	5.10%	2.00%
Localized PC 8	15610549	184	93.20%	4.70%	2.00%
Localized PC 9	18155683	184	94.00%	3.60%	2.30%
Localized PC 10	20623725	185	93.10%	4.60%	2.20%
Localized PC 11	18649331	185	93.20%	4.50%	2.20%
Localized PC 12	25650530	184	93.80%	3.90%	2.30%
Localized PC 13	25135889	185	93.30%	4.30%	2.40%
Localized PC 14	18588276	185	93.00%	4.60%	2.30%
Localized PC 15	16358873	184	92.60%	4.60%	2.70%
Localized PC 16	11664637	183	93.60%	3.80%	2.50%
Advanced PC 1	11010408	192	93.20%	5.40%	1.40%
Advanced PC 2	7329094	190	92.80%	4.20%	2.90%
Advanced PC 3	6342924	190	91.20%	5.90%	2.90%
Advanced PC 4	10739863	192	92.90%	5.00%	2.00%
Advanced PC 5	11324808	190	92.80%	5.10%	2.10%
Advanced PC 6	6435489	190	91.90%	4.60%	3.40%
Advanced PC 7	6094843	190	92.50%	4.50%	3.00%
Advanced PC 8	11130052	191	94.70%	3.30%	2.00%
Advanced PC 9	8269301	190	92.00%	5.00%	3.00%
CRPC 1	11004263	192	94.00%	4.30%	1.70%
CRPC 2	16470827	198	87.00%	5.00%	7.90%
CRPC 3	10419554	191	93.60%	3.90%	2.40%
CRPC 4	18593262	198	86.30%	5.20%	8.40%



Sample name	No of read pairs	Mapped length	% Uniquely mapped	% Multi-mapped	% Unmapped
CRPC 5	21346122	198	85.10%	6.20%	8.70%
CRPC 6	7663241	190	91.80%	5.30%	2.80%
CRPC 7	25713631	199	85.80%	5.20%	8.90%
CRPC 8	12752718	191	92.90%	5.70%	1.40%
CRPC 9	6080233	190	91.80%	5.50%	2.70%
CRPC 10	8497661	191	93.30%	5.20%	1.40%
CRPC 11	12580694	191	92.90%	4.90%	2.10%
CRPC 12	17613955	199	87.60%	3.80%	8.50%



# PRINCIPAL COMPONENT ANALYSIS AND PEARSON'S CORRELATION HEATMAP

**Figures 2** and **3** present the visualisation of the principal component analysis result and the Pearson's correlation coefficients as a heatmap, respectively. These analyses were both performed to data as a final method to ensure data quality. In the principal component analysis, the two principal components explained 29 % and 9 % of the variance between samples. According to the visualisation of the principal component analysis, the samples mainly separated into clusters based on the progression step, except for the advanced PC and CRPC samples that were not clearly distinguishable from each other. As was seen also in Pearson's correlation heatmap (**Figure 3**), the samples were highly variable within all the three cancer sample groups, which is not unusual due to for example different stages of the tumor samples and gene copy number variations and chromosomal aberrations often occurring in the advanced tumor stages.



*Figure 2: Presentation of the results of principal component analysis (PCA).* The data is regularised log2-transformed. 5000 genes with the highest variance were used in the analysis. The samples are coloured based on the progression step of prostate cancer.





*Figure 3: Heatmap presenting Pearson's correlation coefficients calculated pair-wisely for all samples.* The data used for the calculation was regularised log2-transformed, and 5000 genes with the highest variance was used.

### DIFFERENTIALLY EXPRESSED GENES

The analysis of differential gene expression resulted in finding genes with statistically significant difference in expression for each of the three comparisons (**Table 4**). Significant differential expression (DE) was regarded if expression fold change > 2 and adjusted p-value < 0.05. To compare the overlap of the gene lists of the three comparisons, a Venn diagram was constructed (**Figure 4**). Moreover, the results of the DE analysis of each comparison were visualised as volcano plots, as seen in **Figure 5**, presenting the DE analysis results of the comparison "CRPC vs. Advanced PC".





CRPC vs Advanced PC

*Figure 4: Venn diagram showing the overlap of the differentially expressed gene lists.* Gene lists of each of the three comparisons (Localized PC vs. BPH, Advanced PC vs. Localised PC and CRPC vs. Advanced PC) were obtained by thresholding adjusted p-value < 0.05 and absolute log2 fold change > 1.

#### Table 4: Counts of statistically significantly differentially expressed genes in each comparison.

Comparison	Number of DE genes (adj. p-value < 0.05 and abs. log2 fold change > 1)
Localized PC vs BPH	4877
Advanced PC vs Localized PC	734
CRPC vs Advanced PC	148





CRPC vs. Advanced PC

*Figure 5: Volcano plot illustrating the results of the differential expression analysis of the CRPC vs. Advanced PC comparison.* The blue dots represent the genes that are considered significantly differentially expressed (adjusted p-value <0,05 and absolute log2 fold change > 1).



The full lists of differentially expressed genes are reported in attached tables (see section *Deliverables*). The columns of the tables are explained in **Table 5.** All differential expression tables attached to this report follow similar format. The expression patterns of the most significantly changed genes in each comparison were visualised using boxplots, showing the gene-level variations of the expression inside the sample groups. **Figure 6** shows an example of such plots, visualising the expression profiles of the top 4 DE genes from the comparison "CRPC vs Advanced PC".



*Figure 6: Box plots of regularised log transformed expression levels of top DE-genes of comparison "CRPC vs Advanced PC".* Black vertical line of each box shows the median expression level inside the sample group, while the bottom and the top of the box represent 25th and 75th percentile, respectively. Whiskers are placed into the nearest measured expression value being at the distance of abs(1,5\*IQR), where IQR is the range between 75th and 25th percentile.



#### Table 5: The columns in tables of differentially expressed genes

Title	Description
Gene Symbol	HGNC symbol of the gene
Ensembl ID	Ensembl gene ID
Gene Description	Description of the gene
Average Expression	Average gene expression across all samples
Log2 Fold Change	Log2-transformed fold change of expression between contrast groups
P-value	P-value from a Wald test
Adjusted p-value	P-value adjusted for multiple testing by Benjamini-Hochberg procedure

Finally, the expression of the most significantly differentially expressed genes were visualised as a heatmap that was clustered column-wise and row-wise (**Figure 7**). The heatmap was in line with the PCA-plot (**Figure 2**), showing that all normal samples grouped together and a majority of the tumor samples also formed their own clusters. However, there is no complete distinction between the three cancer types due to great variance within each sample group. Top 50 most interesting DE genes (smallest p-values) in all three gene lists were chosen for plotting. Since the gene lists are overlapping (see **Figure 4**), the visualised data is actually the unique set of these genes.





Figure 7: Heatmap of regularised log-transformed expression levels of the most significantly differentially expressed genes from all three comparisons. Intersection of the top 50 statistically significant DE genes from each comparison (based on ascending p-values from DESeq2 analysis) was used for generating the heatmap. Samples and genes have been hierarchically clustered, and the gene expression values have been scaled in row direction.

### ENRICHMENT ANALYSIS

The genes found differentially expressed were further analysed by enrichment analysis. The enrichment analysis of Gene Ontology Biological Process terms, KEGG pathway terms and REACTOME pathway terms resulted in finding terms that were statistically significantly overrepresented (adjusted p-value < 0.05) among the differentially expressed genes. However, a majority of the significant enrichments were



obtained from comparison "Localized PC vs. BPH" that had the largest amount of DE genes, while the other two comparisons produced only a few enriched terms or no enrichments at all. The enriched GO BP terms, KEGG pathway terms and REACTOME terms are reported in attached tables (see section *Deliverables*). The format of the attached tables is presented in **Table 6.** All enrichment analysis tables attached to this report follow similar format.

Title	Description
Term ID	GO Biological Process / KEGG / REACTOME term ID
Term description	GO Biological Process / KEGG / REACTOME term description
Gene ratio	The ratio between the number of DE genes associated with the term in question and the number of DE genes mapped to any term.
Background ratio	The ratio between the number of all genes associated with the term in question and the number of all genes mapped to any term.
P-value	P-value of the enrichment calculated using the hypergeometric distribution
Adjusted p-value	P-value adjusted for multiple testing using Benjamini-Hochberg method
DEGs associated with term	Names (as gene symbols or Entrez IDs) of the differentially expressed genes annotated to the term in question

#### Table 6: Columns in the result tables of enrichment analysis.

Pathway enrichment results were visualised using network plots or pathway diagrams. **Figure 8** shows an exemplary network visualisation of the top 5 most significantly enriched GO Biological Process terms associated with the DE genes from the comparison between localised PC samples and BPH samples. **Table 7** presents the top 10 enriched terms for the same comparison with further details. Among the enriched terms there were several terms related to morphogenesis, cellular metabolism and cell-cell adhesion, which are all processes that are very likely to be affected by tumor morphogenesis.





**Figure 8: Visualisation of the top 5 enriched GO BP terms associated with the DE genes obtained from the comparison of localized PC samples versus BPH samples**. Enrichment was done using DE-genes with adjusted p-value < 0.05 and absolute log<sub>2</sub> fold change > 1. Adjusted p-value threshold of 0.05 was used to find terms that were statistically significantly overrepresented. The larger, brown nodes represent the enriched GO BP terms, and the smaller nodes represent the genes that are annotated to the terms. Gene nodes have been coloured according to the expression fold change between the sample groups, with red colour indicating up-regulation and green colour indicating down-regulation.





Term ID	Term description	Gene ratio	Background ratio	P-value	Adjusted p- value	DEGs annotated to term
GO:0042472	inner ear morphogenesis	41/2893	94/15878	1.054E-08	8.675E-05	HPN/ATP8A2/ MYO7A/
GO:0001823	mesonephros development	40/2893	99/15878	2.018E-07	8.303E-04	PGF/MYC/ EPCAM/
GO:0034109	homotypic cell- cell adhesion	31/2893	73/15878	1.301E-06	1.946E-03	HBB/SLC7A11/ TLN1/
GO:0006801	superoxide metabolic process	27/2893	60/15878	1.595E-06	1.946E-03	NOX4/ITGAM/ PREX1/
GO:0051952	regulation of amine transport	29/2893	68/15878	2.556E-06	2.384E-03	SYT4/TACR2/ SV2A/
GO:0015837	amine transport	31/2893	75/15878	2.608E-06	2.384E-03	SYT4/TACR2/ SV2A/
GO:1903779	regulation of cardiac conduction	29/2893	69/15878	3.646E-06	3.000E-03	TRPM4/NPR2/ SLC8A3/
GO:1900274	regulation of phospholipase C activity	18/2893	34/15878	5.143E-06	3.526E-03	RASGRP4/ BICD1/
GO:0010863	positive regulation of phospholipase C activity	17/2893	32/15878	8.864E-06	4.401E-03	RASGRP4/ FGFR1/KIT/
GO:0001766	membrane raft polarization	8/2893	9/15878	9.092E-06	4.401E-03	MAL2/MAL/ GSN/

 Table 7: Top 10 significantly enriched GO BP terms for comparison "Localized PC vs. BPH" ordered by ascending adjusted p-value.

KEGG pathway term enrichment analysis also resulted in finding statistically significantly enriched pathways for comparison "Localized PC vs. BPH" that are likely to be related to cancer. Enriched terms presented in **Table 8** included terms such as "Cytokine-cytokine receptor interaction", "Rap1-signaling pathway" and "Proteoglycans in cancer". As an example, **Figure 9** shows a KEGG pathway diagram of "Rap1-signaling pathway", displaying the expression changes of the significantly changed genes in localised PC samples when compared to BPH samples by colouring up-regulated genes in red and down-regulated genes in blue. Several genes involved in Rap1-signaling have been significantly changed, suggesting that changes in this signaling pathway are likely to be associated with tumorigenesis.





*Figure 9: Expression changes on KEGG Rap1-signaling pathway were found significantly enriched in comparison "Localized PC vs. BPH".* Log2-transformed gene expression fold changes are re-scaled into the range between -1 to 1, showing the down-regulated genes in blue and up-regulated genes in red.

Term ID	Term description	Gene ratio	Background ratio	P-value	Adjusted p-value
hsa04060	Cytokine-cytokine receptor interaction	67/1258	253/6581	2.145E-03	2.478E-02
hsa04080	Neuroactive ligand-receptor interaction	65/1258	242/6581	1.702E-03	2.110E-02
hsa04020	Calcium signaling pathway	56/1258	176/6581	3.160E-05	1.959E-03
hsa04015	Rap1 signaling pathway	56/1258	200/6581	1.246E-03	1.756E-02
hsa04024	cAMP signaling pathway	54/1258	193/6581	1.539E-03	1.994E-02
hsa05205	Proteoglycans in cancer	53/1258	197/6581	4.214E-03	3.959E-02
hsa04022	cGMP-PKG signaling pathway	48/1258	158/6581	3.887E-04	8.033E-03
hsa04062	Chemokine signaling pathway	48/1258	170/6581	2.238E-03	2.478E-02
hsa04360	Axon guidance	48/1258	173/6581	3.288E-03	3.398E-02
hsa04514	Cell adhesion molecules (CAMs)	44/1258	135/6581	1.177E-04	5.011E-03

Table 8: Top 10 significantly enriched KEGG pathway terms from comparison "Localized PC vs. BPH" ordered by the count of DE genes annotated to the pathway term.



# DELIVERABLES

The figures and files produced in the project are presented in **Table 9**.

File/Folder name	Description
1_QC/fastqc_stats_before_and_after_trimming.xlsx	FastQC quality control summary statistics table showing the number of reads and their lengths, the percentage of duplicate reads and the percentage of GC content.
1_QC/STAR_alignment_statistics.xlsx	Aligment statistics of the samples showing total number of reads in samples, average length of mapped reads and percentages of reads mapping to unique positions, reads mapping to multiple positions and reads that could not be mapped to the reference.
1_QC/FASTQC_examples_raw, 1_QC/FASTQC_examples_trimmed	Subfolders containing an exemplary FASTQC reports (html) for two samples before and after trimming together with histograms (pdf) visualising library size, percentage of duplicated reads and percentage of QC-content experiment- widely.
2_Upstream_visualisations/top5000_PCAplot.pdf, 2_Upstream_visualisations/top5000_PCAplot.png	Presentation of the results of principal component analysis. Both pdf-version and png-version of the same figure are included.
2_Upstream_visualisations/top5000_correlation_heatmap.pdf, 2_Upstream_visualisations/top5000_correlation_heatmap.png	Heatmap presenting Pearson's correlation coefficient for all samples pair-wise, when 5000 genes with the highest variance was used. Both pdf-version and png-version of the same figure are included.
3_DE_tables/DE_gene_counts.xlsx	Statistics for the sizes of DE-gene lists with different absolute fold change thresholds used.
3_DE_tables/All_annotated DEs	Tables of annotated differentially expressed genes before any filtering. There is an xlsx-file for each of the three comparisons (Localized PC vs. BPH, Advanced PC vs. Localized PC and CRPC vs. Advanced PC). Files are named with pattern [COMPARISON]_FULL.xls.
3_DE_tables/Filtered_DEs	Tables of filtered annotated differentially expressed genes for the three comparisons (Localized PC vs. BPH, Advanced PC vs. Localized PC and CRPC vs. Advanced PC). There are two xlsx files for each comparison: one containing DE genes with P-value <0,05 and at least 2-fold up- or downregulation and other with P-value 0.05 and at least 3- fold up- or downregulation. Files are named with pattern [COMPARISON]_adj.p_[P-VALUE]_intFC_[ABSOLUTE FOLD CHANGE]xls.
4_Visualisations/Boxplots/rlog_boxplots.pdf	Gene-level visualisations showing the median expression level (regularized log transformed) and the deviation of the expression inside each of the four sample groups. There is a pdf file for each comparison, showing plots for top-genes in each comparisons DE-table. Moreover, there is a pdf file of manually selected genes of interest.
4_Visualisations/DE_heatmaps	Heatmaps of r rlog-transformed expression values.

#### Table 9. The figures and files produced in the project are presented and explained.



File/Folder name	Description
4_Visualisations/Venn_diagrams	Venn_FC2.tiff containing visualisations of DE gene lists with P-value <0,05 and at least 2-fold up- or downregulation and Venn_FC3.tiff visualising gene lists with P-value 0.05 and at least 3-fold up- or downregulation. There is also a png file used in the report (edited font sizes).
4_Visualisations/Volcano_plots	Visualisation of differentially expressed genes for the three comparisons (Localized PC vs. BPH, Advanced PC vs. Localized PC and CRPC vs. Advanced PC). There are two png-files for each comparison: one colouring the DE genes with P-value <0,05 and at least 2-fold up- or downregulation and other colouring the genes with P-value 0.05 and at least 3-fold up- or downregulation. Files are named with pattern volcano_[COMPARISON]_adj.p_[P- VALUE]_intFC_[ABSOLUTE FOLD CHANGE].png
5_Enrichment_analyses/GOBP	Results for over representation analysis of Enriched GO Biological Processes for differentially expressed genes with p-value cutoff 0.05 and q-value cutoff 0.2. For each comparison, there is an xls file named with pattern GOBP_[COMPARISON].xls. There are also network visualisations of top terms , named with the pattern cnetplot_GOBP_[COMPARISON][_topX].pdf. In these visualizations, enriched terms and differentially expressed genes are presented together with their relationships.
5_Enrichment_analyses/KEGG	Results for over representation analysis of KEGG Pathway terms for differentially expressed genes with p-value cutoff 0.05 and q-value cutoff 0.2. There is an xls file per comparison for which the statistically significant results were found,with pattern KEGG_[COMPARISON].xls.
5_Enrichment_analyses/ReactomePA	Results for REACTOME Pathway enrichment analysis. Significant results were found only for comparison "Localized PC vs BPH. There is an xls-file presenting the result of analysis with p-value cutoff 0.05 and q-value cutoff 0.2. Visualization of the top 5 terms is attached as a pdf and as an png-file where the font sizes were edited for the report. Enriched terms and differentially expressed genes are presented together with their relationships.



# REFERENCES

Andrews, S. 2010. FastQC: a quality control tool for high throughput sequence data. http:// www.bioinformatics.babraham.ac.uk/projects/fastqc

Ashburner, M., Ball, C.A., Blake, J.A. et al. 2000. Gene ontology: tool for the unification of biology. Nat. Genet. 25: 25-29

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57: 289–300

Chen, H. 2018. VennDiagram: Generate High-Resolution Venn and Euler Plots. R package version 1.6.20. https://cran.r-project.org/web/packages/VennDiagram/

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21

Edgar, R., Domrachev, M. and Lash, A.E. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30: 207-210

GO Consortium. 2017. Expansion of the Gene Ontology knowledgebase and resources. Nucleid Acids Res. 45: D331-D338

Kanehisa, M. and Goto, S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28: 27-30

Kassambara, A. 2017. fastqcr: Quality Control of Sequencing Data. R package version 0.1.0. URL: https:// CRAN.R-project.org/package=fastqcr

Kolde, R. 2015. pheatmap: Pretty Heatmaps. R package version 1.0.8. URL: https://CRAN.R-project.org/package=pheatmap

Krueger, F. 2017. TrimGalore!: a wrapper script to automate quality and adapter trimming. URL: https://www.bioinformatics.babraham.ac.uk/projects/trim\_galore/

Love, M.I., Huber, W. and Anders, S. 2014. Moderated estimation of fold change and dispersion for RNAseq data with DESeq2. Genome Biology 15:550.

Luo, W. and Brouwer, C. 2013. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics 29: 1830-1831.

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.





SRA Knowledge Base. 2011. Bethesda (MD): National Center for Biotechnology Information (US). URL: https://www.ncbi.nlm.nih.gov/books/NBK56551/

Yu, G., Wang, L.G., Han, Y. and He, Q.Y. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology 16: 284–287.

Yu, G. and He, Q. 2016. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Molecular BioSystems, 12: 477–479.

Yun, S. J., Kim, S.-K., Kim, J., Cha, E.-J., Kim, J.-S., Kim, S.-J., ... Kim, W.-J. 2017. Transcriptomic features of primary prostate cancer and their prognostic relevance to castration-resistant prostate cancer. *Oncotarget*, *8*(70), 114845–114855.



