

Otso Niiranen

# RNA-sequencing reveals stromal cell induced changes in cancer cell gene expression profiles

---

Helsinki Metropolia University of Applied Sciences

Bachelor of Engineering

Biotechnology and Food Engineering

Bachelor's Thesis

25 May 2015

Author(s) Title	Otso Niiranen RNA-sequencing reveals stromal cell induced changes in cancer cell gene expression profiles
Number of Pages Date	41 pages + 2 appendices 25 May 2015
Degree	Bachelor of Engineering
Degree Programme	Biotechnology and Food Engineering
Specialisation option	Biomedicine
Instructor(s)	Juha E A Knuutila, M.Sc., lecturer Päivi Ojala, Ph.D. Professor Pirita Pekkonen, M.D., Ph.D. student
<p>Biology and medicine changed forever in June 2000 when the first complete draft sequence of human genome was announced. This was quickly followed by expansion of bioinformatic data management possibilities due to remarkable increase in computing capacity. In recent years, RNA-sequencing (RNA-seq) has become superior choice in revealing the transcriptome of a cell and when studying global gene expression profiling. It offers better sensitivity, range of expression, more sequencing depth and allows to sequence novel genomes with low background noise signal compared to other existing conventional array technologies.</p> <p>The aim of this thesis was to analyse RNA-sequencing data of untreated and treated samples of two melanoma cancer cell lines. This thesis is part of the research conducted at the laboratory of Professor Päivi Ojala, Translational Cancer Biology Research Program Unit, Faculty of Medicine, University of Helsinki. Laboratory's research goal is to find and produce novel knowledge and deepen the current understanding of the mechanisms of metastasis in melanoma cancer.</p> <p>The melanoma cancer cell lines WM852 and Bowes were co-cultured according to experimental setup together with lymphatic endothelial cells. After culturing two cell types were separated with a method utilizing magnetic nanoparticles, and the RNA was extracted from the cancer cell samples. The RNAs were sent to DNA Sequencing and Genomics laboratory of Institute of Biotechnology to be sequenced. The sequencing results were then quality tested, aligned to a human reference genome GRCh38.76, and aligned reads were counted for the differential expression analysis. Generally Applicable Gene-set Enrichment analysis (GAGE) and Pathway analysis were done to get an impression of the functional and mechanistic changes of the treatment.</p> <p>The interesting gene expression changes found by this experiment are currently being validated using qRT-PCR. These results will contribute towards expanding the understanding of malignant melanoma and metastasis, further help us to develop better treatments for cancer centric diseases.</p>	
Keywords	RNA-sequencing, cancer, melanoma, metastasis

Tekijä(t) Otsikko  Sivumäärä Aika	Otso Niiranen Strooman aiheuttamien geeniekspressiomuutosten selvittäminen melanoomasoluissa RNA-sekvenssointi tekniikalla 41 sivua + 2 liitettä 25.05.2015
Tutkinto	Insinööri (AMK)
Koulutusohjelma	Bio- ja elintarviketekniikka
Suuntautumisvaihtoehto	Biolääketiede
Ohjaaja(t)	Juha E A Knuutila, M.Sc., lehtori Päivi Ojala, Ph.D. Professor Pirita Pekkonen, M.D., Ph.D. opiskelija
<p>Biologian ja biolääketieteellinen tutkimus muuttuivat vuonna 2000 ratkaisevasti, kun ensimmäinen kokonainen vedos ihmisen genomista julkaistiin kaikkien saataville. Bioinformaattisen tietojen käsittelyn mahdollisuudet lisääntyivät samanaikaisesti valtavasti voimakkaan tietokoneiden laskentakapasiteetin kasvun seurauksena. RNA-sekvenssoinnin suorittaminen on viime vuosina yleistynyt, halutessa selvittää solun transkriptomin sisältö. Se tarjoaa suuremman herkkyuden ja pienemmän taustamelusignaalin ja mittaa tehokkaammin geeni-ilmentymisen absoluuttista määrää kuin tähän asti muilla menetelmillä on saavutettu. RNA-sekvenssointi on myös sekvensoimattoman genomien tai mutatoituneet syövän selvittämiseksi paras saatavissa oleva vaihtoehto.</p> <p>Tämän tutkimuksen tavoitteena on selvittää imusuonten seinämäsolujen interaktion vaikutusta melanoomasolujen geeniekspressioon. Tämä tehtiin vertaamalla kahden erityyppisen melanooma ihosyöpäsolulinjan geenien ekspressioita ennen ja jälkeen käsittelyn. Syöpäsolujen RNA-sekvenssoinnin ja tulosten analysoinnin kautta voidaan ymmärtää paremmin melanoomaa ja metastasoimiseen liittyviä biologisia mekanismeja. Tutkimus toteutettiin professori Päivi Ojalan tutkimusryhmässä, joka on osa Helsingin yliopiston lääketieteellisen tiedekunnan Translationaalisen syöpäbiologian tutkimusohjelmaa..</p> <p>Melanooma WM852- sekä Bowes-syöpäsolulinjat kasvatettiin koesuunnitelman mukaisessa yhteiskasvatusmallissa yhdessä imusuonten seinämäsolujen (LEC) kanssa, jonka jälkeen solut erotettiin toisistaan ja syöpäsolujen RNAt eristettiin. Sekvenssointi toteutettiin ”DNA Sequencing and Genomics” laboratoriossa Biotekniikan Instituutissa. Sekvenssointituloksien laatu tarkastettiin ja rinnastettiin ihmisen referenssigenomia GRCh38.76 vasten, jonka jälkeen genomiin rinnastetut sekvenssifragmentit laskettiin. Fragmenttiluvuista suoritettiin differentiaalinen ekspressioanalyysi, jonka tuloksena pääteltiin, kuinka paljon geenien ilmentyminen on muuttunut käsittelyn aikana. Geenien ekspressiomuutosten kokonaiskuvan saamiseksi suoritettiin lisäksi geenien rikastumis- sekä signaalintireittianalyysit.</p> <p>Tällä hetkellä mielenkiintoisimpia geenilöydöksiä ollaan validoimassa käyttämällä qRT-PCR menetelmää sekä funktionaalisia kokeita soluilla. Nämä tulokset lisäävät ymmärrystämme melanoomasta ja sen leviämisestä, sekä auttavat kehittämään uusia hoitomuotoja syöpätaudeille.</p>	
Avainsanat	RNA-sekvenssointi, syöpä, melanooma, metastasia, bioinformatiikka

## Contents

### Abbreviations

1	Introduction	1
	THEORETICAL PART	2
2	Features of genes and gene expression	2
2.1	Introduction to genetics	2
2.2	Genes and chromosomes	3
2.3	Regulation of gene expression	3
3	Origin of cancer	5
3.1	Carcinogenesis	5
3.2	Melanoma overview	7
3.3	Melanoma progression	9
4	Next generation sequencing	11
4.1	RNA-sequencing	11
4.2	Illumina sequencing technology	12
	PRACTICAL PART	14
5	Materials and methods	14
5.1	RNA sequencing (RNA-seq)	14
5.2	Chipster	14
5.3	FastQC and PRINSEQ	15
5.4	Tophat2 and Bowtie2	16
5.5	HTSeq for counting reads	17
5.6	The R Project	18
5.7	DESeq2	19
5.8	KEGG and GO ontologies	19
5.9	Generally applicable gene enrichment analysis (GAGE)	20
5.10	Pathview	20
6	Quality control of data and alignment	21
6.1	Experimental design	21

6.2	FastQC	23
6.3	Tophat 2 alignment	24
6.4	HTSeq count reading	25
6.5	Quality control of count data	26
7	Results of RNA-seq data analysis	29
7.1	Differential expression analysis	29
7.2	Methods of result verification	31
7.3	Enrichment and pathway analysis results	33
8	Conclusions	36
9	References	37

## Appendices

Appendix 1. Chipster run parameters

Appendix 2. R run parameters and script

## Abbreviations

cDNA	copy deoxyribonucleic acid
DNA	Deoxyribonucleic acid
dsDNA	double stranded DNA
ER	Endoplasmic reticulum
ES	Enrichment scores
FDR	False discovery rate
GAGE	Generally applicable gene-set enrichment analysis
GLM	Generalized linear model
GO	Gene Ontology
HTS	High-throughput sequencing
HUGO	Human Gene Nomenclature Committee
KEGG	Kyoto encyclopaedia of genes and genomes
LEC	lymphatic endothelial cell
LFC	Logarithmic fold change
MAPK	mitose activated protein kinase
mRNA	messenger RNA
miRNA	micro RNA
RLOG	regularized logarithmic transformation

RNA	Ribonucleic acid
RNA-seq	RNA-sequencing
rRNA	ribosomal RNA
siRNA	silencing RNA
tRNA	transfer RNA
VEGF	vascular endothelial growth factor
VST	Variance-stabilizing transformation

## 1 Introduction

This thesis was done for the laboratory of Prof. Päivi Ojala, Translational Cancer Biology Program, Research Programs Unit, Faculty of Medicine, University of Helsinki. The aim of this thesis was to analyse RNA-sequencing data with the current bioinformatic methods. Main methods for this study were differential gene expression analysis, generally applicable gene enrichment analysis and pathway analysis. The cells that were RNA-sequenced were produced in a cell culturing experiment where the goal was to study lymphatic endothelial cell and melanoma cell interaction.

Melanoma refers to a condition in melanocytes when the cells begin to proliferate abnormally. Melanocytes localize to the skin, between the epidermis and the dermis, where they produce and store melanin pigments. Melanoma is clinically described to appear as pigmented macules, which can change in symmetry, size or colour and appear as new lesions or in pre-existing lesions (Muller, 2008). Melanoma pathogenesis is driven by both genetic and environmental risk factors. The melanoma incidence is influenced by skin pigmentation, sun-exposure history as well as extensive UVB/A radiation and geographical location (Chin, 2003).

Although melanoma is responsible for the majority of deaths in skin related malignancies, it only accounts for about 5 % of all skin cancers (Kim, et al., 2015). Most of the diagnosed melanoma that ends up fatal are caused by metastasised nodes which have spread around the body (Ferlay, et al., 2014). Therefore, it is important to understand the mechanisms of the melanoma metastatic spread to develop better treatments for these patients and for this reason, the research done to understand the metastasis progression is exceptionally rewarding.

## THEORETICAL PART

### 2 Features of genes and gene expression

#### 2.1 Introduction to genetics

The mechanisms that make simple and advanced life possible rely on the structure of the double-stranded DNA molecule. DNA is made from simple subunits called nucleotides, each consisting of sugar-phosphate molecule with a nitrogen side group, base, attached to it. The deoxynucleotides in DNA are adenine, guanine, cytosine and thymine, labelled A, G, C and T. According to strict rule defined by the complementary structures of the bases A binds to T and C binds to G. DNA cannot form by itself, it has to be synthesised from a pre-existing strand. DNA replication occurs with different controls and helper auxiliary molecules that affect within the process by allowing it to continue with different rates or by inhibiting it completely. The fundamentals of the DNA are universal: DNA is the information store for heredity and it's the way how life is copied throughout the living world (Alberts, et al., 2015).

To transmit its information out, DNA must do more than replicate itself. It must express its information by guiding the synthesis of other molecules in the cell. Expression mechanism is nearly identical in all living organisms and leads to formation of two key classes of polymers: RNA and proteins. The first phase of the process is called transcription, where DNA sequence is used as a template for the synthesis of RNA molecules. Many of the RNA molecules are directing the synthesis of proteins in an process called translation. RNA molecule is formed from a different sugar, ribose, instead of deoxyribose and has a different base uracil (U) in place of thymine (T). Newly produced transcripts function most notably as messenger RNA (mRNA) molecules which guide the synthesis of proteins according to the genetic instructions stored in the DNA. Therefore, knowing how to read the mRNA molecules, it is possible to gain major insights how individual or groups of cells work. Three sequential nucleobases form a codon which tells the cell to produce a specific amino acid, the building block of proteins. In transcription, the DNA monomers are written as RNA monomers which are mass produced and disposable, whereas DNA in the cell is left intact and untouched. New proteins are synthesized from mRNA by using it as a template in an event called

translation. Ribosome proteins carry out instructions written in the mRNA by adding amino acids together and forming polypeptide chains (Alberts, et al., 2015).

## 2.2 Genes and chromosomes

In their cells, humans have 46 chromosomes, made up of 23 pairs, which are constructed from the DNA. These include 44 chromosomes called autosomes; they are numbered from 1 to 22 according to size from the smallest to the largest, as well as the two sex chromosomes: X and Y. Each chromosome consist of two very long thin strands of DNA chains twisted into the shape of a double helix and are located in the nucleus (the 'control centre') of the cells. Since the chromosomes in the nucleus come in 23 pairs, the genes that are located as part of the DNA in chromosomes, also appear as pairs. It is approximated that only 1% of total genome is made up of genes. These areas are called coding DNA regions whereas the areas between genes are called a non-coding DNA. Although it is called non-coding DNA it has a very important role to play in regulating genes for example switching them on and off. Alleles are forms of the same gene with small differences in their sequence of DNA nucleobases. These small differences in the DNA contribute to each unique physical feature phenotype in animalia. In total, human genome contains approximately over 20 000 genes. Some of the genes are also located at the mitochondria, which are scattered in cytoplasm (Alberts, et al., 2015).

## 2.3 Regulation of gene expression

A cell normally expresses only a fraction of its genes, and differentiation and development of a cell to multicellular organism is the inevitable outcome of this behaviour. Different and specialized cells in multicellular organisms are able to alter the pattern of gene expression in response to extracellular signals. Specialized cells express different gene profiles at different times. For example, if liver cell is exposed to a hormone called glucocorticoid, the production of number of proteins that increases energy production from amino acids increases exponentially. When the glucocorticoid is not present the protein production returns to its basal unstimulated levels in liver cells. If exposed to fat cells, glucocorticoid reduces the production of tyrosine aminotransferase, whereas some cell types do not respond to hormone at all (Alberts, et al., 2015). Thus, different specialized cell types regularly respond very differently to the same extracellular cues. Due to nature of general biological variance in gene expressions, not always the same signal causes the same outcome. Consequently, every cell has its genetically

inherited expression level which is not normally subjected to major changes, therefore giving each specialized cell type its distinguished characteristics (Alberts, et al., 2015).

Regulation of gene expression can take place in at many different stages from DNA to RNA to protein (Figure 1). The cell can control the proteins it produces by (1) transcriptional control by controlling how often and when the gene is transcribed, (2) RNA processing control by controlling the processing and splicing of the RNA transcripts, (3) RNA transport and localization control by selecting the cytoplasmic localization of mRNAs to be exported from the nucleus to the cytoplasm for translation, (4) translational control by selecting which mRNA is translated by ribosomes, (5) mRNA degradation control by degrading the unnecessary mRNA molecules in the cytoplasm, and (6) protein activity control by activating, inactivating, relocalizing or degrading certain proteins.

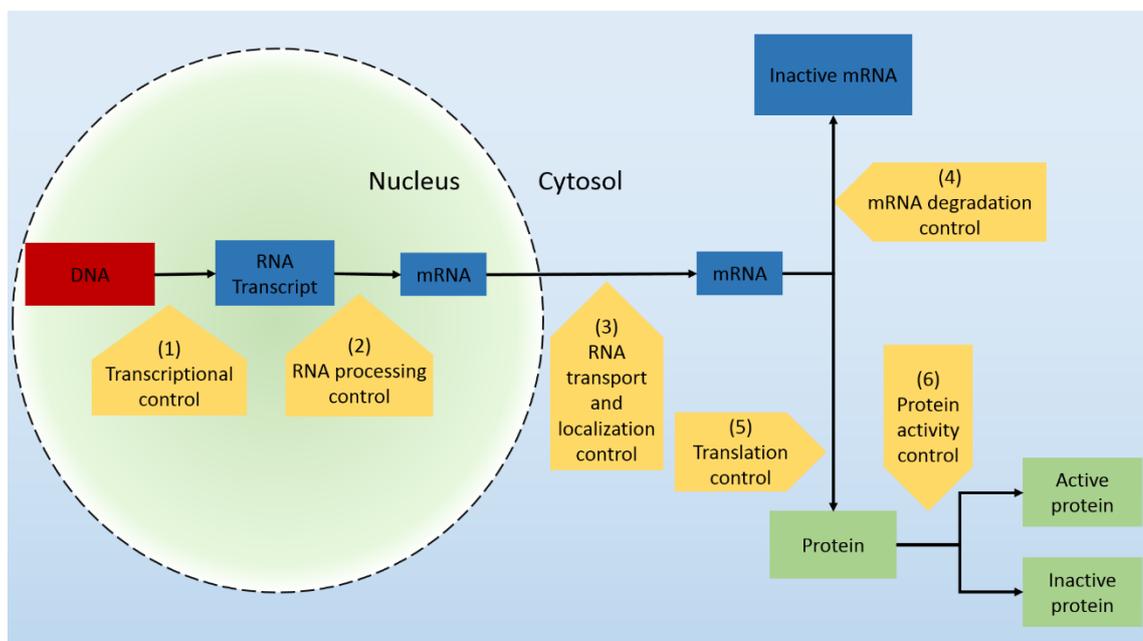


Figure 1. Steps of eukaryotic gene expression control (Alberts, et al., 2015).

Although every regulatory step can and will participate in expressing a gene, transcriptional control is the most important point of regulation. Transcriptional control switches the process of transcription on and off for individual genes in cells. Every eukaryotic gene is typically controlled by many different transcriptional regulators. Some of the regulators, also called transcription factors, are bind in the close proximity of the gene start site and are intimately linked to the expression of the gene. Some regulators or promoter areas which initiates the gene transcription might be far away from the target

transcriptional start location and they activate through different regulators. The major reason why changes in gene expression might be difficult to track is because of the vast complexity of the regulatory network inside and outside of the cell.

### 3 Origin of cancer

#### 3.1 Carcinogenesis

Cancer cells are notorious by for two intrinsic properties: they reproduce in defiance of the natural limitations on cell growth and division, and they invade and colonize areas normally reserved for other cells. It is the combination of these two properties that makes the cancers generally dangerous for living organisms. A cell that grows abnormally in mass and proliferates uncontrollably will result in a neoplasm, or a *tumour*. Neoplasm is called benign if the tumour cells have not yet become invasive. Benign tumour is usually easy to cure by removing the tumour mass from the tissue. If tumour acquires an ability to invade into the surrounding tissues, it is considered malignant in other words cancerous. Increased invasiveness might allow the cancer cells to enter surrounding tissues and form secondary tumours called metastases. In general, the more dispersed cancer the harder it is to cure and usually the metastases are the reason for the death of the cancer patient (Alberts, et al., 2015).

Cancer is caused by mutations that occur mainly in somatic cells and therefore most cancers are not inheritable. Cancer development requires generally multiple mutations in growth regulatory genes. Additional drivers of cancer formation are epigenetic changes which modify the chromatin structure of the cell without altering the DNA sequence. Factors that play a role in changing the genetic material and in the development of cancer are called carcinogens. The factors that change the DNA sequence are called mutagens. Different kinds of chemical and physical factors and some viruses induce mutations and cancer. Local changes in the nucleotide sequence are typically caused by chemical carcinogens like asbestos, tobacco smoke, dioxins or natural aflatoxin B<sub>1</sub>. Radiation, such as x-rays or ultraviolet light, causes typically DNA base alterations, double strand breaks and sometimes chromosome breaks and chromosomal translocations, where parts of nonhomologous chromosomes rearrange. Viruses can also cause cancer by transforming the infected cells by interfering with the

expression of regular proliferation and growth control. Typically the human tumour viruses carry and express e.g. human papillomavirus (HPV), Epstein Barr virus (EBV), Kaposi's sarcoma herpesvirus (KSHV)) with it or activate cellular genes through integration into the host DNA next to proto-oncogene leading to its activation e.g. hepatitis B virus (HBV)). Infection by these viruses can increase the chance of a normal cell to transform into a cancer cell but does not necessarily lead to the development of the disease. Examples of cancers caused by these viruses are cervical carcinoma by HPV, Burkitt's lymphoma (EBV), human herpesvirus 8 (HHV-8) causing Kaposi's sarcoma, hepatitis B and C –viruses which cause liver cancer by HBV (Solunetti, 2015).

Although carcinogens can be accounted for some of the most common mutations in cancers, mutations occur spontaneously even with no known mutagens present. Rate of spontaneous gene mutations per gene per cell division is around  $10^{-6}$  (Alberts, et al., 2015), which means that in a lifetime single gene might undergo about  $10^{10}$  separate mutations. Some of these mutations occurs in genes that regulate division and growth, causing the cells to defy the normal limitations of cell proliferation. However, not all of these cells survive the mutations, and apoptosis eradicates the mutated cell clones. DNA repairing machinery also plays a vital role in the minimizing the harm mutations cause to DNA. Several percentages of genomes coding capacity is reserved for DNA repair functions, which is essential for the preservation of genome integrity. (Alberts, et al., 2015). Less than 0.02% of tens of thousands of daily mutations generated by heat, metabolic accidents, radiation or substances in environment accumulate as permanent mutations in the DNA sequence. Variations in the genetic information that do not make any difference to the way the message is read or to the protein that is produced by the cell are quite common (Alberts, et al., 2015).

Proto-oncogene is an initially normal cellular gene that can transform into an oncogene at least two different ways: i) through a point mutation such as replacement of one base pair to an alternative one, ii) through larger genomic rearrangement such as translocation, duplication, inversion or deletion of DNA from one chromosome to another (Alberts, et al., 2015). Proto-oncogene encoded proteins act like accelerators of growth and in protein translation they are participating in many processes like cell cycle progression, apoptosis and differentiation (Chi, 2012) (Chial, 2008). An example of a proto-oncogene is a transcriptional regulator MYC, which regulates approximately 10-15% of human genes. MYC expression is documented to be elevated in approximately 70% of human cancers (Gurell, et al., 2008).

In addition to activation of oncogenes, cancer initiation depends on inactivation of genes that restrict growth and proliferation. These genes produce cellular proteins that regulate cell cycle progression and will stop cell cycle if the DNA is damaged, hormone receptors that restrict the cell from dividing, proteins that promote apoptosis, and enzymes that contribute to DNA repair. For example, p53 is a gene that takes regulates apoptosis and takes part in regulation of cell cycle by stopping the cell cycle to the G1 phase if the DNA is damaged which means that DNA replication in the S phase will not start (Figure 2).

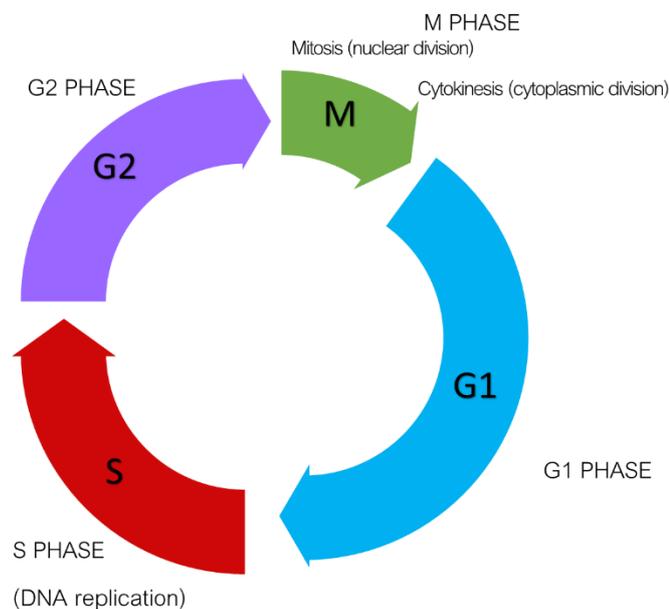


Figure 2. The division cycle of most eukaryotic cells is divided into four discrete phases: M, G1, S, and G2. M phase (mitosis) is usually followed by cytokinesis. S phase is the period during which DNA replication occurs. The cell grows throughout the interphase, which includes G1, S, and G2. (Alberts, et al., 2015)

### 3.2 Melanoma overview

It is estimated that globally there is over 200,000 new melanoma cases per year, with an estimation of 46,000 deaths per year (Erdman, et al., 2012). In Finland, nearly 1200 new cases of melanoma are found diagnosed per year (Suomen syöpärekisteri, 2015). That is 4.1% of all cancers recorded in Finland. The combined 5-year survival rates of all melanoma are 84% for males and 89% for females in Finland (Engholm, et al., 2014). Mortality rates around continents range from 1.8 deaths in Africa to 11.2 in Europe per

100,000 where the difference can be explained mostly by ethnic backgrounds (Ferlay, et al., 2014).

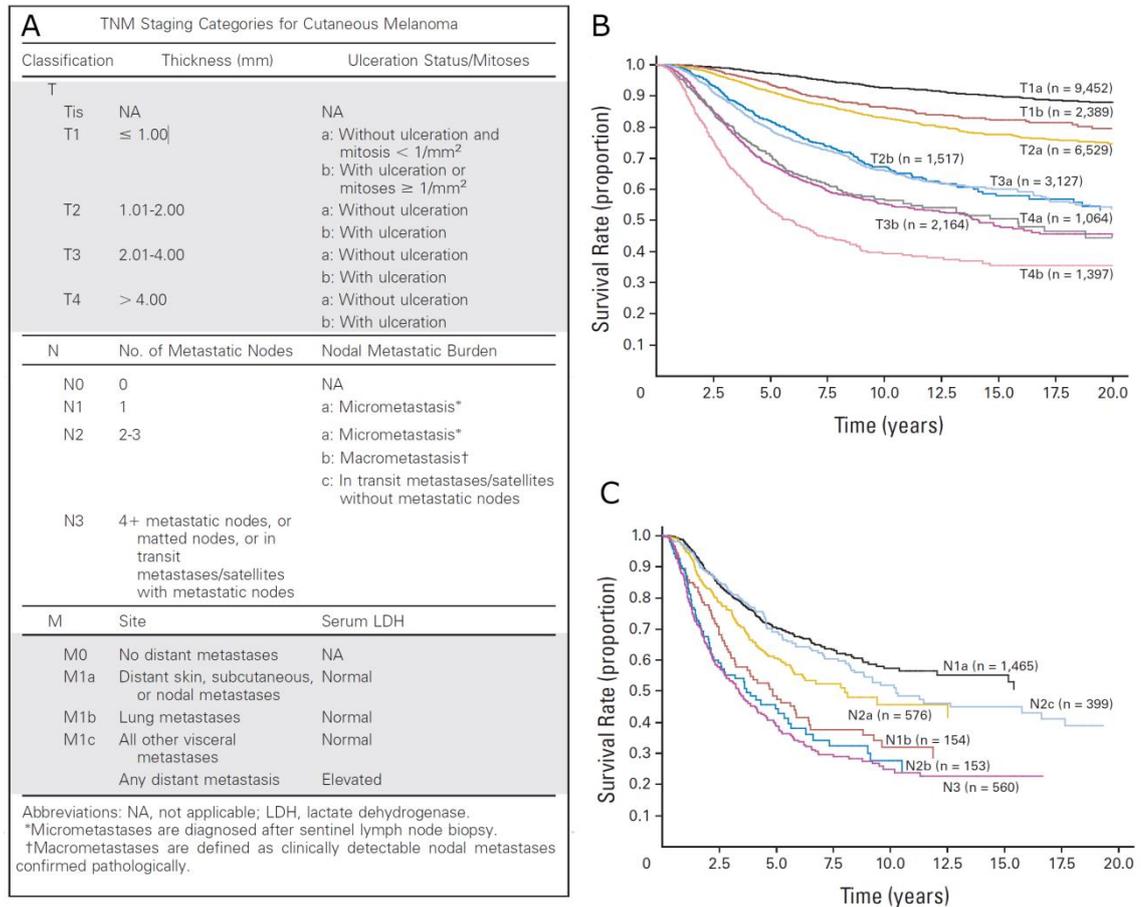


Figure 3. The TNM (Tumour Node Metastasis) categories for the seventh edition of the American Joint Committee of Cancer Staging and Classification Manual (3A), Survival curves from the American Joint Committee on Cancer Melanoma Staging Database comparing the different T categories and (3B), the different N categories (3C).

Although the melanoma survival rates are good in cases where the tumor has not spread, however once the melanoma has spread to the lymph nodes (Stage III or N1-3), the 5-year survival rate drops to 40% - 70% depending on the N category as seen on Figure 5. Despite of recent progresses in the treatment regimens targeted towards the main driver mutations of melanoma, most notably the BRAF mutant which is found in approximately 40 % of cutaneous melanomas (Davies et al., 2002, Curtin et al., 2005) (Kim, et al., 2015), the treatments are hardly ever curative in patients with the M-category

distant metastatic disease (stage IV) where survival rate drops to 15% - 20 % (Larkin et al., 2014).

### 3.3 Melanoma progression

Thickness of the primary tumour is the best characterized property when predicting the melanoma metastasis (Balch et al., 2009). The most significant event in the melanoma primary tumour progression is the shift from a radial growth phase, where the melanoma cells are growing along the epidermis, to a vertical growth phase, which is characterized by tumour cell invasion into the deeper layers of the skin (Figure 3) (Chin, 2003). Adhesive properties and increased stimulation of the tissue degrading proteases are shown to increase tumour progression and increases invasion of cancer cells (Villanueva and Herlyn, 2008, Moro et al., 2014). Moreover, the change from radial to vertical growth phase is related to an increase of vascular endothelial growth factor (VEGF) quantity leading to vascularization and capillary leakage are both predictive factors of the malignant melanoma metastatic development as well (Figure 5) (Braeuer et al., 2011).

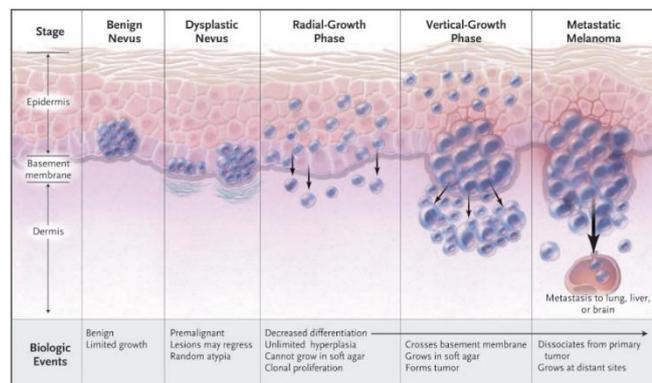


Figure 4. Primary malignant melanoma progression. If the melanoma tumour is detected in radial-growth phase, it can usually be removed quite easily with surgical actions. In vertical-growth phase tumour gains the potential to invade tissue and surgical removal becomes much harder (CancerLink - Neoplastic Diseases Reviews, 2010).

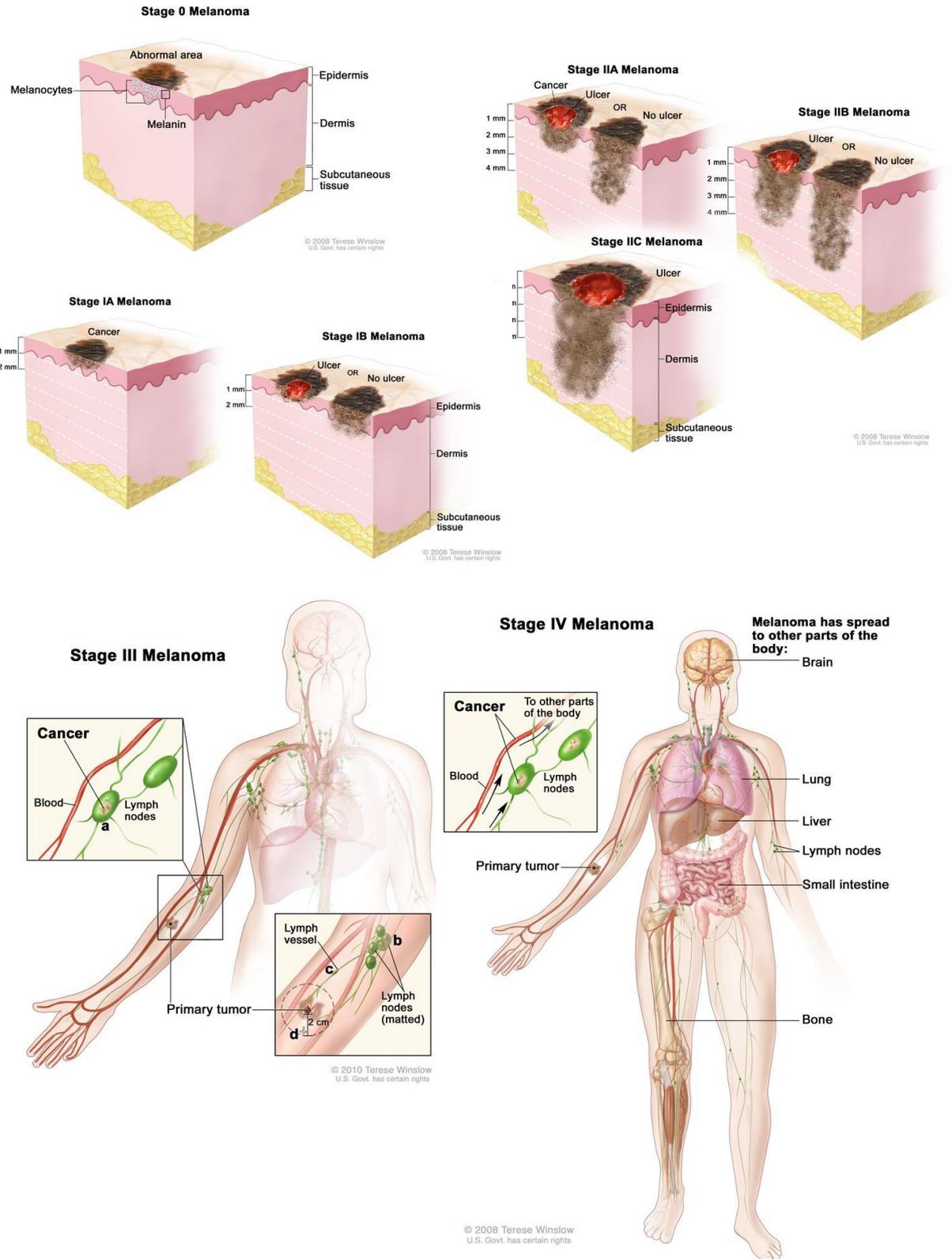


Figure 5. Invasion staging of melanoma (PDQ Melanoma Treatment—for health professionals , 2015).

The prognosis of melanoma gets worse as with the progression of stages and time (Figure 4 and Figure 5A). From stage III onwards the melanoma has been linked to poor survival in the time of five years (Balch, et al., 2009). During recent years it has become increasingly clear that the lymphatic endothelial cells (LECs) might interact with their surrounding cells and tissues and regulating both physical and pathological processes (Emmett, et al., 2011) (Nagano, et al., 2012) (Pekkonen, 2015). In this study, we set to investigate if the LEC – melanoma cell interactions are directly contributing to the metastatic phenotype of melanoma, and would thus facilitate the tumour progression. To this end, Prof. Päivi Ojala's laboratory developed several LEC-melanoma co-culturing systems, from which RNA-seq analysis was performed to decipher global gene expression changes in the melanoma cells co-cultured with LECs.

## **4 Next generation sequencing**

### **4.1 RNA-sequencing**

In May 2008, astonishingly within three weeks, Science, Cell, Nature and Nature Methods published online the five world first articles introducing the new novel technique that has been upsetting microarray industry since then. The method now recognized as RNA-sequencing provides higher resolution picture of transcription than what was the standard before. The main goal of experiments using RNA-sequencing is to assay the transcriptome and RNA isoforms to accurately quantify gene expression levels. RNA-seq technology does not require species- or transcript-specific probes like microarrays do. The technology behind RNA-seq relies heavily on deep sequencing which means that every genomic region in the samples is sequenced hundreds or thousands of times to make sure that the possible errors while sequencing are singled out (Meyerson, et al., 2010) (Illumina, 2015). A population of RNA is converted to a library of cDNA fragments with nucleotide adaptors, commonly poly-T tails, connected to one or both ends of the cDNA fragment (single or paired end sequencing). Each polynucleotide molecule is then sequenced in a high-throughput manner, to obtain short sequences. Reads are typically

30-400 base pairs long, depending on the sequencing technology used (Illumina, 2015). Various sequencing platforms are supported including Illumina, Life Sciences, Roche 454 and Applied Biosystems. After sequencing the resulting reads are assembled de novo without the genomic sequence or aligned to a reference genome or reference transcripts to produce a genome scale transcription map.

Over the years, RNA-seq has been gradually overtaking microarrays to become the tool of choice for genome wide analysis of the transcriptome (Wang, et al., 2009). It can detect novel transcripts, gene fusions, single nucleotide variants, small insertions and deletions, and other previously unknown nucleotide changes that arrays cannot detect, which makes it extremely versatile for studying complex transcriptomes (Wang, et al., 2009).

#### 4.2 Illumina sequencing technology

Sequencing technology used for the RNA-seq data in this thesis was from Illumina. Illumina sequencing relies on standard dideoxy method (Lasken & McLean, 2014) with few innovations incorporated (Alberts, et al., 2015). Assay in which the sample cDNA fragments are immobilized receives a DNA polymerase cocktail with different fluorescently labelled nucleotides A, T, G and C. Different fluorescent molecules are attached to each four nucleobases that thus emit four different wavelengths of light for nucleobase identification purposes. Bases also carry a chemical group that is attached to the base (3'-end block), while inhibiting the elongation by DNA polymerase. The cycles of sequencing are regulated by this block so fluorescent signal can be read correctly. Only complementary nucleotides in the template binds to the clusters and the rest is washed away alongside with blocking group and fluorescent label (Figure 6). Every time the process repeats high-resolution camera takes image or continuous video of which four nucleotides was added to the chain of each cluster.

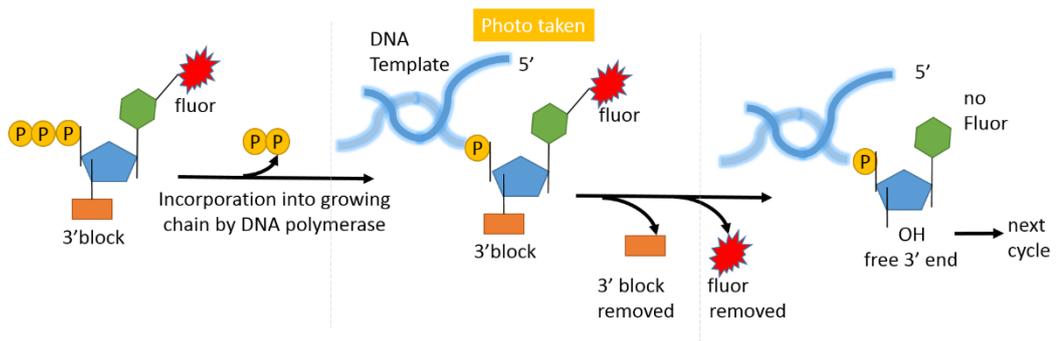


Figure 6. Each cycle completed by RNA-sequencing takes approximately one hour in which modified nucleotide incorporation, image acquisition and removal of the 3'block and fluorescent group happens billions of times simultaneously (Alberts, et al., 2015) (Illumina, 2015).

Illumina platform makes it possible to sequence billions of reactions at the same time. Each individual sequence read is approximately 200 nucleotides long, which is shorter than with the other sequencing instruments but the vast amount of massive parallel sequencing is able to produce several human genomes worth of a replica sequence during 24h meaning additional increase in the sequencing depth (Illumina, 2015).

## PRACTICAL PART

### 5 Materials and methods

#### 5.1 RNA sequencing (RNA-seq)

RNA extraction for the RNA sequencing analysis was done from three independent experiments with a TRI reagent (Sigma) protocol supplemented with phenol-chloroform precipitation step or with NucleoSpin RNA II kit (Macherey Nagel) (Pekkonen, 2015). The extracted RNA concentrations were measured with NanoDrop, and Bioanalyzer (Agilent Technologies) analysis was performed to check the RNA quality. RNA sequencing was performed with NextSeq500 sequencer (Illumina) as quadruplicates following standard protocol. The sequence data was aligned to human GRCh38.76 reference genome with using Chipster's Tophat2 alignment and the differentially expressed genes were obtained by using DESeq2 Bioconductor package. Genes with adjusted p-values less than 0.05 were considered significant. The cutoff for the FoldChange was set from 1 to -1 for further studies. Individual gene/transcript expressions are shown as counts.

#### 5.2 Chipster

Chipster is a multipurpose bioinformatics data analysis platform with interactive visualizations and possibility for pipeline workflows developed by CSC - IT Center for Science Ltd (Figure 7). It is released under General Public Licence version 3 (GPLv3) which means it is free to use and share for any purpose. Chipster has an extensive collection of analysis tools for next generation sequencing (NGS), microarray and proteomics data (Chipster, 2015). The NGS options cover quality control analysis from raw sequence data to alignment, differential expression analysis and extensive applications such as pathway analysis. In this thesis Chipster was used for quality control, alignment and reading the count matrix which can be used for downstream analysis. Chipster client software uses Java Web Start that connects to CSC's servers where the actual computing is done. The server runs variety of programs which can be selected and executed from the users Java interface. Functions used in this thesis are FastQC, Tophat2, HTSeq and Define NGS Experiment utility (Kallio, et al., 2011).

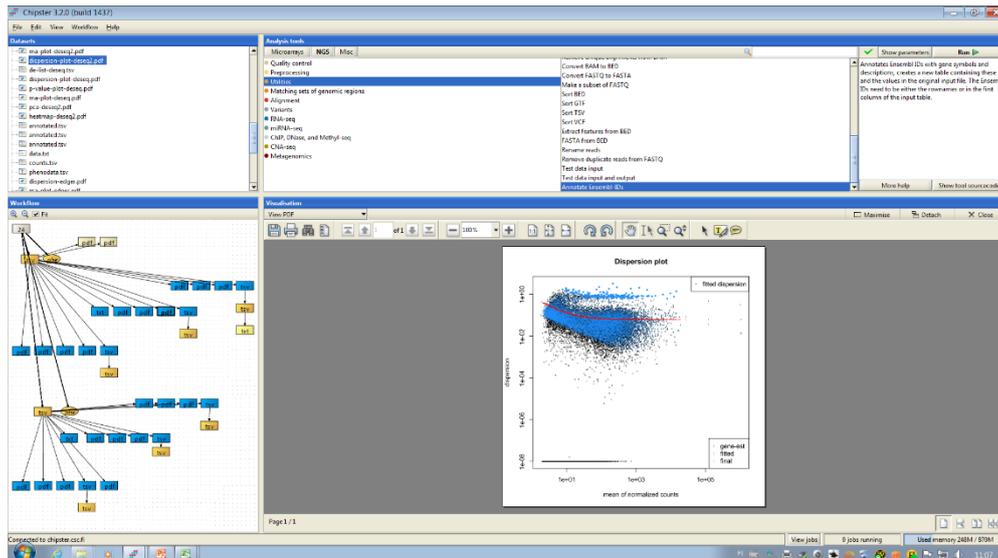


Figure 7. The Chipster program focuses on graphical interface and the representation of functions is informal therefore it is quite easy to start using (Kallio, et al., 2011).

### 5.3 FastQC and PRINSEQ

FastQC produces several quality control plots which are important when evaluating the condition of raw sequence files before doing any further analysis (Figure 8). If the quality is not optimal, trimming and filtering of sequence reads must be done, otherwise the downstream analysis will not provide statistically relevant results (Kallio, et al., 2011). The FastQC in this thesis was used as in Chipster. PRINSEQ can be used for quality checks as well, but here we used it to trim the ends of the poly N or A/T tails which were used in the sequencing.

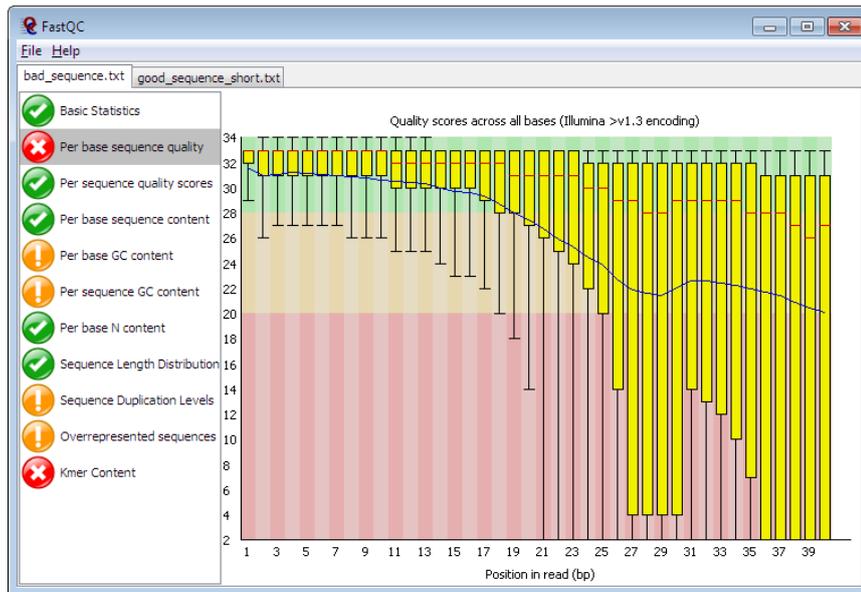


Figure 8. FastQC standalone program example. The quality control is essential in discovering how successful the RNA-seq data run was (Anders, et al., 2015).

#### 5.4 Tophat2 and Bowtie2

Tophat2 is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to genomes using the short read aligner Bowtie2, and then analyses the mapping results to identify splice junctions between exons. Bowtie2 is fast and memory-efficient tools for aligning sequencing reads to long known reference sequences. It is relatively good at aligning either in the range of 50 – 100 or up to thousands of sequence character worth of reads and excels in aligning to long genomes. By default, Tophat 2 performs end-to-end read alignment which can be also called by the name “untrimmed” or “unclipped” alignment. Basically, it compares reference genome characters to read characters and calculates alignment scores.

The following example of how the Tophat 2 alignment scores are calculated is described in the manual at <http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml> and it clarifies the alignment process.

```

Read:      GACTGGGCGATCTCGACTTCG
Reference: GACTGCGATCTCGACATCG

Alignment:
Read:      GACTGGGCGATCTCGACTTTCG
          | | | | | | | | | | | | | |
Reference: GACTG--CGATCTCGACATCG

```

A mismatched base at a high-quality position in the read receives a penalty of -6 by default. A length-2 read gap receives a penalty of -11 by default (-5 for the gap open, -3 for the first extension, -3 for the second extension). Thus, in end-to-end alignment mode, if the read is 50 bp long and it matches the reference exactly except for one mismatch at a high-quality position and one length-2 read gap, then the overall score is  $-(6 + 11) = -17$ .

The best possible alignment score in end-to-end mode is 0, which happens when there are no differences between the read and the reference. (<http://bowtie-bio.sourceforge.net/>)

## 5.5 HTSeq for counting reads

A very typical part when using the HTSeq library is to give it a list of genomic features (such as genes or exons) and it calculates how many sequencing reads overlap each of the feature (Anders, et al., 2015). These reads are transcribed to counts and used for main analysis measurement in downstream analyses with methods like DESeq2 (Love, et al., 2014), GAGE (Luo, et al., 2009) and Pathview (Luo & Brouwer, 2013). The script htseq-count is designed for differential expression analysis, only reads mapping unmistakably to a single gene are counted, while reads aligned to multiple locations or overlapping with more than one gene are discarded. This sort of mapping is desirable, as one of two genes which have some similarity between sequences are differentially expressed. A read that maps to both genes similarly should be discarded, because if it were counted for both genes, the additional reads from the differentially expressed gene could also cause the other gene to be accidentally called differentially expressed (Nature America, Inc., 2013) (Anders, et al., 2015).

## 5.6 The R Project

R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering etc.) and graphical techniques, and is highly extensible (Figure 9) (R Core Team, 2015). The platform of R can be extended through community created packages as the functions allow for specialized statistical techniques or graphical devices. This thesis used R project Bioconductor which is a software developed for computational biology and bioinformatics. In total there is over 6600 additional packages as of May 2015 which are available at the Comprehensive R Archive Network (CRAN), Bioconductor and other repositories. The main packages used in this thesis through Bioconductor were: DESeq2, GAGE, pathview, ggplot2.

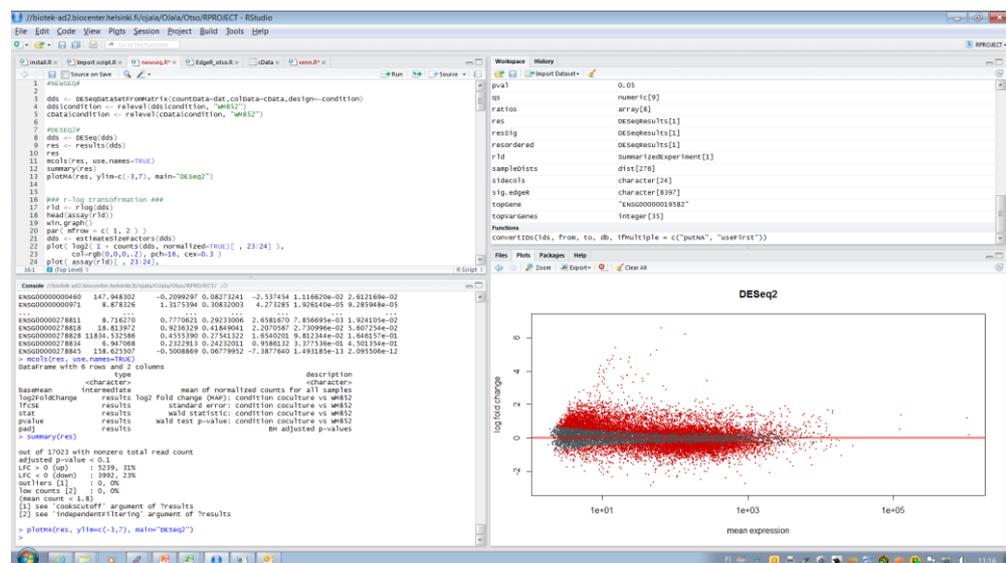


Figure 9. R code can be highly specified to meet the needs of the changing are of bioinformatics and statistics (R Core Team, 2015).

## 5.7 DESeq2

An essential task in analysing the high-throughput sequencing data such as RNA-seq count data is to find indication of systematic change across experimental treatment conditions (Lesk, 2014). DESeq2, which was used here as a bioconductor package, uses shrinkage estimation of dispersion and fold changes to analyse differential expression of count data (Zhang, et al., 2014). Analyse method focuses more on quantitative strength rather than mere existence of differential expression (Love, et al., 2014). The null hypothesis in these kinds of tests is commonly the same: fold change of gene expression between untreated and treated samples is precisely zero. This means that if the null hypothesis is true, we will not find genes significantly affected by the treatment. Statistical p-value represents the probability for the null hypothesis. The aim of the analysis is to acquire list of genes that pass several test criteria. These genes are called statistically significantly. The most interesting genes in this list are significantly expressed, usually meaning under -2 or over 2 Fold Change (FC) ratio, but the FC bounds are situational case by case. Although gene might not change its expression at all it still might be significant to know that the gene does not change its behaviour according to treatment. (Love, et al., 2014)

DESeq2 operates by using shrinkage estimators for dispersion and test for differential expression by using model based on negative binomial distribution which is a discrete probability (Love, et al., 2014). The shrinkage of fold changes per sample basis is calculated by the regularized-logarithm transformation (rlog) or by variance stabilized transformation (VST). Both rlog and VST produce data on the log<sub>2</sub> scale which has been normalized with respect to RNA-seq library size (Love, et al., 2014). For genes with higher counts, the rlog or VST transformation does not differ much from a common log<sub>2</sub> transformation, however genes with lower counts, the values are shrunken towards the genes average across all samples which improves sensitivity of the method (Zhang, et al., 2014).

## 5.8 KEGG and GO ontologies

One of the major challenges in a current and future genomic era is to build a complete computer model of the cell or a multi cellular organism, in which we could be able to predict higher complex cellular processes and behaviour from the genomic information.

KEGG (Kyoto Encyclopedia of Genes and Genomes) tries to answer this problem by integrating the top current understanding on molecular high-level functions such as pathways, genes and proteins and biochemical compounds and reactions that happen in the cells (Kaneisha & Goto, 2000). Basically KEGG is a collection of manually drawn pathway maps from literature. GO (Gene Ontology Consortium) on the other hand focuses more on describing genes and gene products and their association with biological processes, molecular functions and cellular components. GO aims to address problems that KEGG's manual curation has by being responsive to automatic annotation possibilities. This has led to almost 99% of the GO data being curated automatically (Plessis, et al., 2011). In the thesis we use these two knowledge based systems as a platform to visualize the enrichment of our RNA-sequencing experiment genes.

### 5.9 Generally applicable gene enrichment analysis (GAGE)

If a researcher does not want to study only down or up regulation of gene expression, GAGE is a simple and well published method for performing a gene set enrichment analysis (Luo, et al., 2009). It gives insight to wider scale changes in a cell culture. In this study GAGE is used as R package and enrichment analysis with both GO and KEGG databases is performed at chapter 7.3.

### 5.10 Pathview

The Pathview R package is a tool set for pathway based data integration and visualization (Luo & Brouwer, 2013). It maps and renders user data on relevant pathway graphs. User needs is to supply their gene or compound data and specify the target pathway and Pathview automatically downloads the pathway graph data, parses the data file, maps user data to the pathway, and renders pathway graph with the mapped data. Although built as a stand-alone program, Pathview seamlessly integrates with pathway (and functional) analysis tools for a large-scale and fully automated analysis pipeline.

## 6 Quality control of data and alignment

### 6.1 Experimental design

The melanoma cell lines used for the RNA-sequencing were WM852 and Bowes. WM852 is isolated from a melanoma skin metastasis and exhibits more aggressive features whereas Bowes is isolated from superficially spreading melanoma and is more indolent. The experiment setup (Figure 10.) for the WM852 and Bowes melanoma cell lines were the same. Both cancer cell lines had 3 samples with consisting of only cancer cells, and 3 samples with of LEC primed cancer cells. WM852 was cultured for 48h and Bowes for 24h. Both cell lines were treated with the same procedure and the melanoma cells were separated from LECs and harvested. Cells were pelleted and frozen to  $-80^{\circ}\text{C}$  to wait for mRNA isolation.

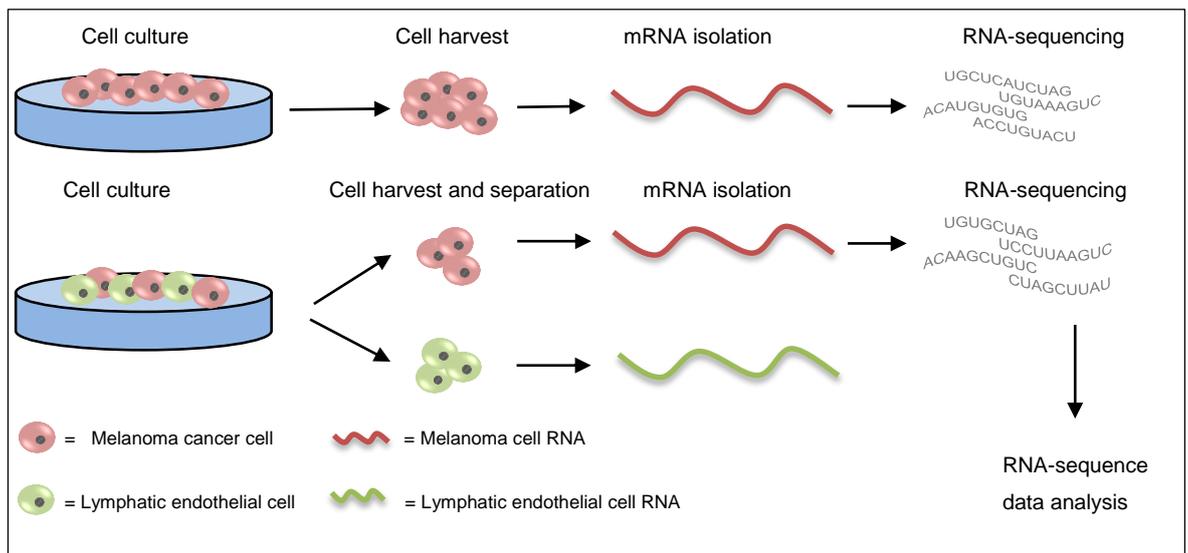


Figure 10. Overview of experimental setup for cell line based events. All cells were pelleted and also LECs could be subjected to RNA-seq as for later date if findings are interesting.

Cancer cells were subjected to RNA-isolation with NucleoSpin RNA II kit (Macherey Nagel) or with Trizol isolation protocol (Sigma) supplemented with acid phenol-chloroform precipitation step. The integrity of the RNA was analyzed with Bioanalyzer and concentrations were measured with Nanodrop (Thermo Fisher Scientific) and the nanodrop results can be found in Table 1.

Table 1. Experimental RNA extracting sheets of the two cell lines WM852 and Bowes. Pellet date tells the date when the culturing was done and the pellet was formed after the harvesting and separation of the cells.

Sample	Description	RNA isolation	Pellet date	RNA c(ng/ul)	260/280	260/230	Kit
1	WM852 1.exp	12.11.2014	7.11.2014	266,8	2,12	1,74	MN
2	LEC primed WM852 1.exp	12.11.2014	7.11.2014	235,27	2,1	2,05	MN
3	WM852 2.exp	28.11.2014	27.11.2014	133,18	1,83	0,77	Sigma
4	LEC primed WM852 2.exp	28.11.2014	27.11.2014	101,71	1,79	0,63	Sigma
5	WM852 3.exp	2.12.2014	27.11.2014	172,8	1,75	0,55	Sigma
6	LEC primed WM852 3.exp	2.12.2014	27.11.2014	121,68	1,75	0,41	Sigma

Sample	Description	RNA isolation	Pellet date	RNA c(ng/ul)	260/280	260/230	Kit
1	Bowes 1.exp	16.2.2015	22.1.2015	179,75	1,88	1,08	Sigma
2	LEC primed Bowes 1.exp	16.2.2015	22.1.2015	221,84	1,86	0,83	Sigma
3	Bowes 2.exp	16.2.2015	30.1.2015	126,34	1,86	1,38	Sigma
4	LEC primed Bowes 2.exp	16.2.2015	30.1.2015	136,68	1,78	0,55	Sigma
5	Bowes 3.exp	16.2.2015	3.2.2015	87,63	1,74	0,48	Sigma
6	LEC primed Bowes 3.exp	16.2.2015	3.2.2015	124,95	1,77	0,52	Sigma

Extracted melanoma mRNA samples were sent to DNA Sequencing and Genomic's laboratory at Institute of Biotechnology where it was sequenced in January 2015 (WM852) and in March 2015 (Bowes).

Both sequencing runs were performed with single end protocol using Illumina NextSeq 500 sequencer. RNA sequencing raw results are in a format .fastq that is default format for most of the sequencing programs. In total we had six biological samples derived from three cancer cell only cultures and three LEC primed cancer cell cultures. Each sample was divided to four technical replicates for the RNA-sequencing step. In the data analysis phase we had 24 samples (.fastq files) for further studies. In this project we used data analysis platform Chipster to perform a quality check, alignment and count extraction for both of the different data sets WM852 and Bowes. Both experiments are analysed separately

## 6.2 FastQC

The FastQC analysis was done to each of the 24 sample files we got from the sequencing laboratory. Quality of the reads were evaluated by Phred quality scores. For example, Phred 30 score means to tell us that 99,9% of the reads are accurately sequenced (Table 2).

Table 2. Phred Quality Score is linked logarithmically to error probabilities (Ewing & Green, 1998).

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%

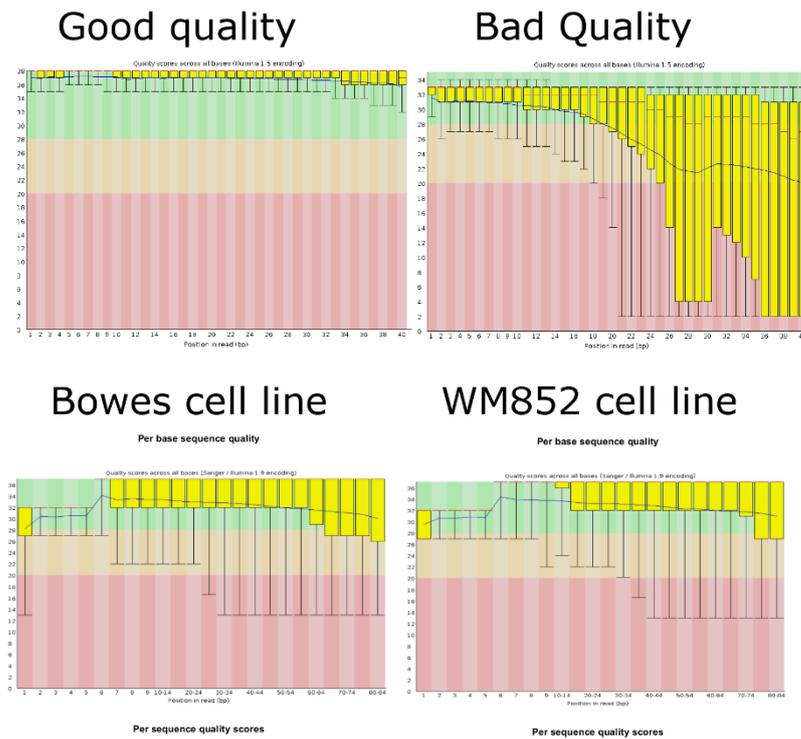


Figure 11. FastQC for the cancer cell lines. The two quality examples in the top row are done with different encoding Illumina 1.5 encoding versus our samples which were done with Sanger / Illumina 1.9 encoding so the two are not possibly likewise comparable. In the Y-axis is phred33 quality code, in the X-axis base pair position in the read. Green area means good, yellow area acceptable and red is bad quality.

In our experiment, the quality check of the sequencing results looked good, as all 24 samples in both experiments showed Phred scores over 26 (see examples, Figure 11 bottom row), meaning our base call accuracy was calculated accordingly equation (1), where Q is phred score, P is base calling error probability. Our base call accuracy minimum was 99,7488%.

$$(1) \quad Q = -10 \log_{10} P$$

As the sequencing quality was particularly good, no trimming or filtering of the reads was necessary except for the poly A/T tails which was done with PRINSEQ (Figure 11). If sequencing quality is poor, i.e. yellow bars in the Figure 11 drop to red area, the trimming and filtering is necessary for accurate downstream analysis.

### 6.3 Tophat 2 alignment

From the conducted steps, the alignment step needs the most of raw computational power. To accelerate this process, Chipster is built so that it does the calculations in a server at the CSC and not from the computer it is used from. However, Chipster can only queue 10 commands, so the commands had to be added to the queue after the first few queues were done and the alignment process itself took approximately 72 hours. The genome used for the alignment was Homo sapiens GRCh38.76.

The alignment was conducted with parameters described under:

```
Dataset name: tophat.bam
Created with operation: Alignment / TopHat2 for single end reads
Parameter Genome: Homo_sapiens.GRCh38
Parameter Use annotation GTF: yes
Parameter When GTF file is used, ignore novel junctions: no
Parameter Base quality encoding used: Sanger - Phred+33
Parameter How many hits is a read allowed to have: 20
Parameter Number of mismatches allowed in final alignment: 2
Parameter Minimum anchor length: 8
Parameter Maximum number of mismatches allowed in the anchor: 0
Parameter Minimum intron length: 70
Parameter Maximum intron length: 500000
Parameter Library type: fr-unstranded
```

### 6.4 HTSeq count reading

Next the count data was obtained by using Chipster function HTSeq (Figure 12). For each gene, it is counted separately how many aligned reads from the genome overlap its exons. The count table generated by HTSeq is in an easily accessible .csv format and can be opened with most of the text or table editors.

Ensembl ID	S1_R1	S1_R2	S1_R3	S1_R4	S2_R1	S2_R2	S2_R3	S2_R4	S3_R1	S3_R2	S3_R3	S3_R4	S4_R1	S4_R2	S4_R3	S4_R4	S5_R1	S5_R2	S5_R3	S5_R4	S6_R1	S6_R2	S6_R3	S6_R4
ENSG00000000003	133	111	121	134	198	207	221	197	202	218	208	211	463	407	474	459	385	347	391	367	404	374	386	380
ENSG00000000005	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG000000000419	268	240	270	210	478	477	457	418	474	512	480	483	1377	1303	1358	1354	993	918	950	943	875	884	947	828
ENSG000000000457	31	28	29	26	23	23	26	25	19	19	27	19	60	56	69	47	43	47	51	35	33	36	40	34
ENSG000000000460	141	128	147	126	166	144	176	163	81	73	90	90	277	299	298	288	205	189	213	224	199	210	251	228
ENSG000000000938	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
ENSG000000000971	0	0	0	0	0	0	1	1	0	0	0	0	0	2	0	0	0	0	0	0	0	2	1	1
ENSG00000001036	166	150	181	166	206	197	176	207	124	110	141	130	297	277	326	317	263	233	268	278	298	291	335	310
ENSG00000001084	3	2	5	5	16	13	16	19	1	2	3	4	20	30	30	28	10	8	6	8	34	30	38	26
ENSG00000001167	107	127	133	102	118	123	118	130	107	119	113	114	281	241	288	264	214	201	233	235	213	177	195	178
ENSG00000001460	58	60	56	61	64	51	65	52	15	22	21	24	78	63	80	85	77	71	62	64	54	63	79	66
ENSG00000001461	714	697	741	755	497	463	543	536	293	328	335	351	711	623	633	689	734	767	848	786	715	712	746	753
ENSG00000001497	649	592	661	623	479	490	521	528	463	436	467	479	865	804	977	882	845	798	879	835	887	935	1030	1008
ENSG00000001561	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000001617	2	6	8	7	19	15	21	18	3	1	4	2	13	14	16	16	1	3	3	3	37	46	45	47
ENSG00000001626	2	1	0	2	2	4	4	1	0	0	0	0	0	1	3	2	0	0	0	2	0	0	2	0

Figure 12. The count table of WM852 opened in Microsoft Excel 2013 with color coded columns for clarity. Red is the Ensembl gene ID, green columns represent the cancer only samples and blue ones the LEC primed samples. Column name S1\_R1 can be interpreted as Sample1\_Replicate1.

The counts for each gene were used for differential expression in chapter 7.1. Approximately 62 000 gene IDs were retrieved in the raw count data file. One third of them were protein coding genes and the rest two thirds is categorised as noncoding DNA sequences. The noncoding sequence include genes that are automatically annotated as well as pseudogenes, introns, rRNA, tRNA, microRNA, and telomeres.

Table 3. Top hits in WM852 by abundance for reference where the counts are sorted by descending row count average. Ensembl ID and Gene Symbol is just an alternative representation for each other and the description is pulled from HUGO.

	Ensembl ID	Gene Symbol	Description	Average counts per sample
W	ENSG00000210082	MT-RNR2	mitochondrially encoded 16S RNA	1182006
M	ENSG00000202198	RN7SK	RNA, 7SK small nuclear	782998
8	ENSG00000211459	MT-RNR1	mitochondrially encoded 12S RNA	126018
5	ENSG00000274012	RN7SL2	RNA, 7SL, cytoplasmic 2	86365
2	ENSG00000111640	GAPDH	glyceraldehyde-3-phosphate dehydrogenase	30640

	Ensembl ID	Gene Symbol	Description	Average counts per sample
B	ENSG00000210082	MT-RNR2	mitochondrially encoded 16S RNA	2165487
O	ENSG00000202198	RN7SK	RNA, 7SK small nuclear	385973
W	ENSG00000211459	MT-RNR1	mitochondrially encoded 12S RNA	177427
E	ENSG00000198938	MT-CO3	mitochondrially encoded cytochrome c oxidase III	43054
S	ENSG00000198804	MT-CO1	mitochondrially encoded cytochrome c oxidase I	37085

Mitochondrial and ribosomal RNA is the most abundant mRNA in the samples. The top genes by abundance (Table 3) are usually household genes both non-coding and coding, that are required to maintain the cell viable and operational and are therefore produced in great numbers. MT-RNR 1 and MT-RNR 2 are genes that encode rRNA, RN7SK controls transcription and RN7SL2 is a part of signal recognition particle. GAPDH works as catalysing enzyme in the sixth step of glycolysis breaking down glucose for energy and carbon molecules for the cell. Cytochrome oxidases MT-CO1 and MT-CO2 are part of the last enzymes in the electron transport chain (Cunningham, et al., 2015).

#### 6.5 Quality control of count data

Quality of the count data needs to be analysed within R where the processing of the differential expression analysis is done. Relationships between samples can be observed with different statistical methods including principal component analysis (Figure 14). Most common method is to normalise the count data for the downstream analysis. The normalisation method (Figure 13) used for the Deseq2 quality analysis can be regularized logarithmic transformation (RLOG) or variance stabilised transformation (VST). The normalisation used in this thesis was VST. Basic rule for the transformation is to pick a one with visually straight red line (Figure 13) or just choose between either RLOG or VST (Love, et al., 2015) (Icay, 2015).

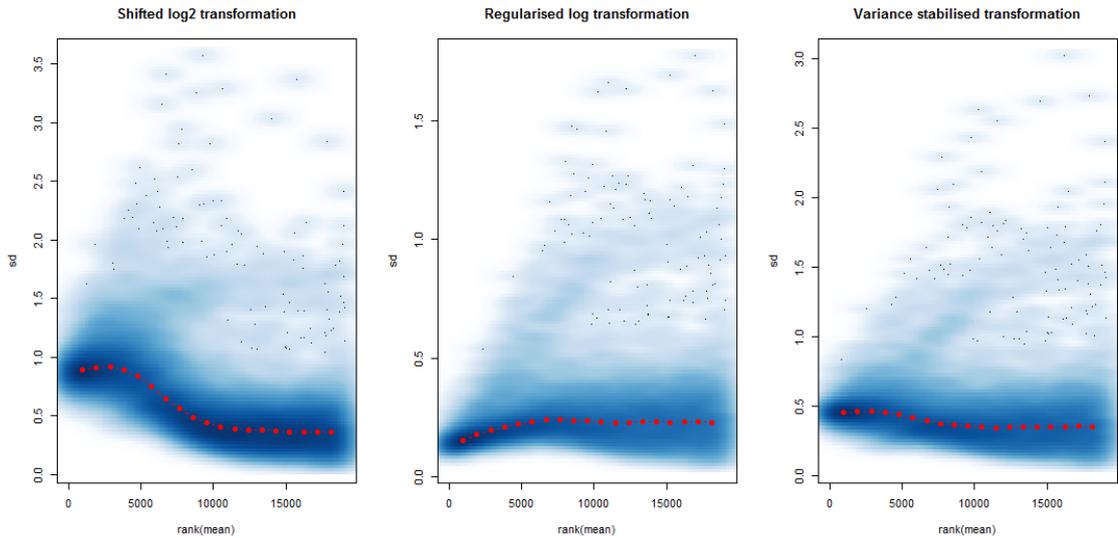


Figure 13. Effect of different transformations visualised. The per-gene standard deviation taken across samples, against the rank of the mean ( $\log_2(n+1)$ ). Variance stabilised transformation takes genes with small count numbers better in to account. Y-axis presents standard deviation and x-axis the gene rank sorted by counts in ascending order.

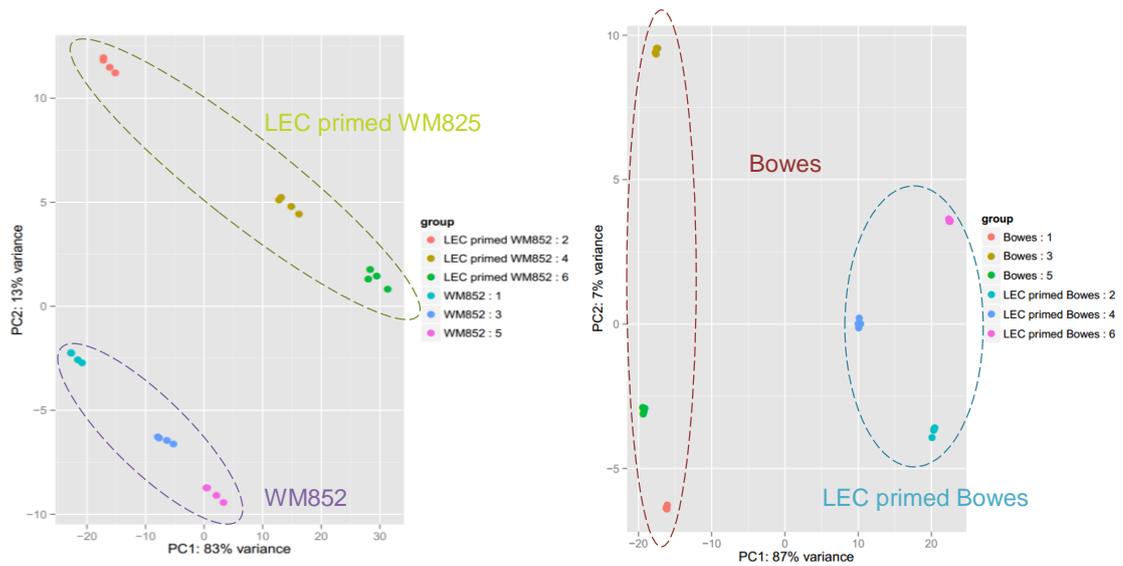


Figure 14. Principal component analysis of the top 500 most variance gene rows from count data of WM852 and Bowes after VST. Samples are divided into Principal component 1 and Principal component 2 according to their variance. Accuracy of the four replicates in each samples is high. The degree of distribution is clear because the sample conditions, visualized by circles, do not overlap.

Quality of the data and similarities between samples was good overall. In Figure 14, some disparity between WM852 samples one and two can be seen. These samples

represent the first pair of the whole experimental set and were isolated with different RNA isolation kit (Table 1). WM852 samples are clustered well in PC2 and slightly divided in PC1 because of the disparity of samples one and two. In Bowes sample 3 was distanced from its condition average in PC2, but remains clustered in PC1. Overall the technical replicate data was excellent, but the biological variance between the samples was quite high. Although the cell line grows and multiplies while the phenotype of the cell stays the same, gene expression levels might have significant variation and act differently depending on the situation. In conclusion, biological variance has to be accounted as a major variable.

The analysis of significant gene expressions changes was done in R – Bioconductor package DESeq2, which calculates significant expression of genes from raw counts. Research group made a decision to filter out genes that had under one count in more than 12 samples before their count table inclusion to the DESeq2 analysis. The reason for the filtering is that low or zero count genes do not have same statistical power in the analysis because of the variance difference. There might be also uncertain counts because of the read errors in both sequencing and alignment. DESeq2 automatically applies VST normalisation and fits the dispersion as seen in Figure 15 and explained at 5.7.

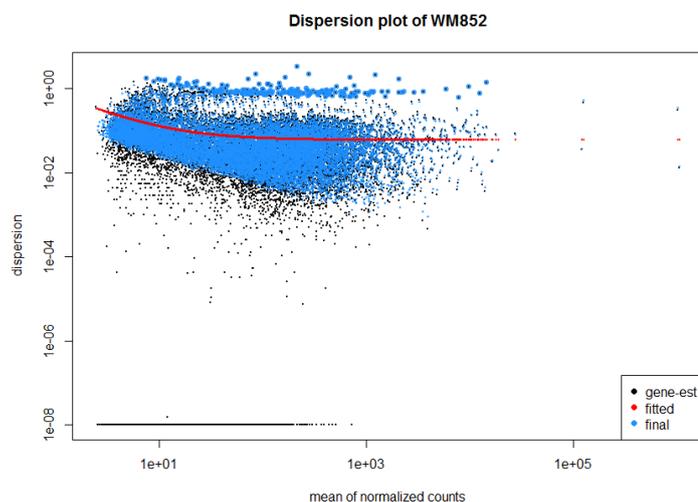


Figure 15. Variance stabilized transformation effect shown in dispersion estimate plot. The gene-wise estimates (black), the fitted values (red), and the final maximum estimates used in testing (blue).

## 7 Results of RNA-seq data analysis

### 7.1 Differential expression analysis

In standard statistical analysis the different condition values are usually compared to one another. In this analysis the experimental conditions are untreated cancer cells (WM852 or Bowes) versus treated cells (LEC Primed WM852 or Bowes).

Easiest way to visualize the complex expression patterns that genes form is to draw a heatmap. In Figure 16, graphical representations of expression data matrix from DESeq2. With a quick look similarities and dissimilarities can be observed.

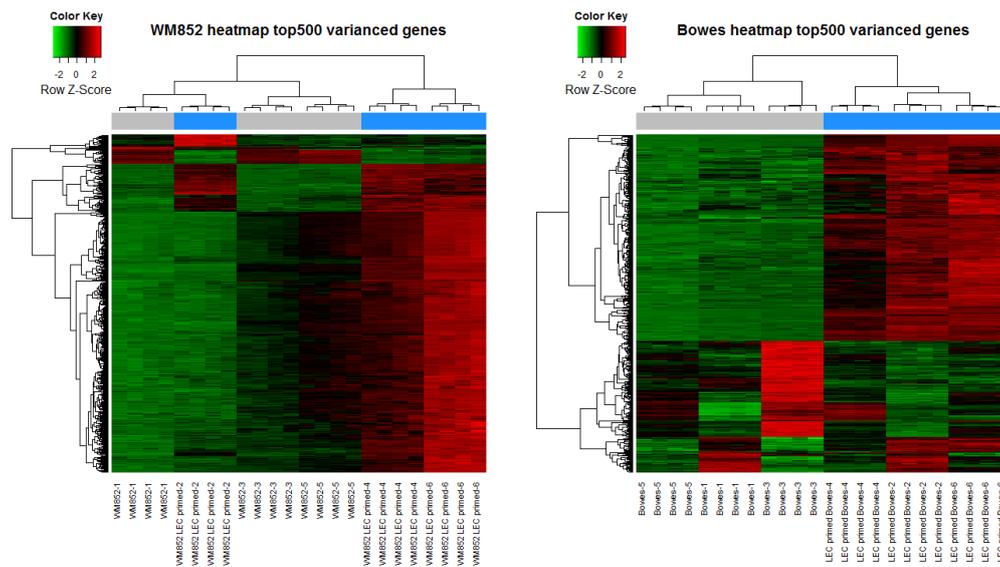


Figure 16. Variance based heatmaps showing the top 500 most differentially variance genes in both populations. Dendrograms and clustering are done automatically and blue column colour represents LEC treated samples. Green and red represents how much the specific sample deviates from gene's average across samples as green is to negative direction and red positive.

In WM852 samples, samples one and two show clear distinctive difference compared to the other samples. However, the distribution of genes and variance looked well-defined in both data sets.

Genes that influence the output of the above heatmap and other figures are in the data matrix that is shown in Table 4. Each column has important data stored in them:

- baseMean is the average of normalized factors
- log2FC is the fold change between groups
- FoldChange in base 10 format
- lfcSE is standard error of log2fc
- stat is wald test statistic
- pvalue tells the probability value for true null hypothesis(no significant expression change)
- padj is Benjamin-Hochberg adjusted p-value (False detection rate adjusted)

Table 4. Example of six most and least expressed genes sorted by fold change from both populations. The gene names are coded in this version of the publication to not expose the results before the acceptance of the manuscript under preparation.

	Ensembl ID	baseMean	log2FC	FoldChange	lfcSE	stat	pvalue	padj
W	Gene WM1	64.1	6.58	95.7	0.37	17.6	4.5E-69	1.2E-66
M	Gene WM2	137.5	6.20	73.5	0.25	24.9	1.8E-136	2.3E-133
8	Gene WM3	15.1	5.40	42.2	0.39	13.9	5.5E-44	5.3E-42
5	Gene WM4	50.4	-2.60	-6.1	0.16	-15.9	9.3E-57	1.6E-54
2	Gene WM5	906.9	-2.72	-6.6	0.09	-30.4	3.4E-203	1.2E-199
	Gene WM6	46.5	-2.80	-7.0	0.16	-17.2	2.2E-66	4.8E-64
	Ensembl ID	baseMean	log2FC	FoldChange	lfcSE	stat	pvalue	padj
B	Gene BO1	200.5	5.04	32.9	0.14	36.7	3.9E-294	1.8E-290
O	Gene BO2	538.9	4.51	22.7	0.15	30.7	5.1E-207	1.4E-203
W	Gene BO3	467.5	4.45	21.8	0.15	29.0	3.0E-185	5.3E-182
E	Gene BO4	118.7	-1.00	-2.0	0.18	-5.7	1.5E-08	4.0E-07
S	Gene BO5	18.0	-1.04	-2.1	0.16	-6.6	2.9E-11	1.1E-09
	Gene BO6	583.3	-1.08	-2.1	0.17	-6.3	3.6E-10	1.2E-08

In the RNA-sequencing analysis the amount of the data produced could be high, and the study could yield over thousand significantly expressed genes where singling out important genes could turn out to be problematic. There is really no good and simple solution to this but pathway and enrichment analysis could provide some hints of the biological importance and give good leads for further studies. Enrichment analysis is discussed in chapter 7.3.

The overall direction of the whole population expression change can be also visualised with DESeq2 function plotMA (Figure 17), which is also a useful data diagnostics tool.

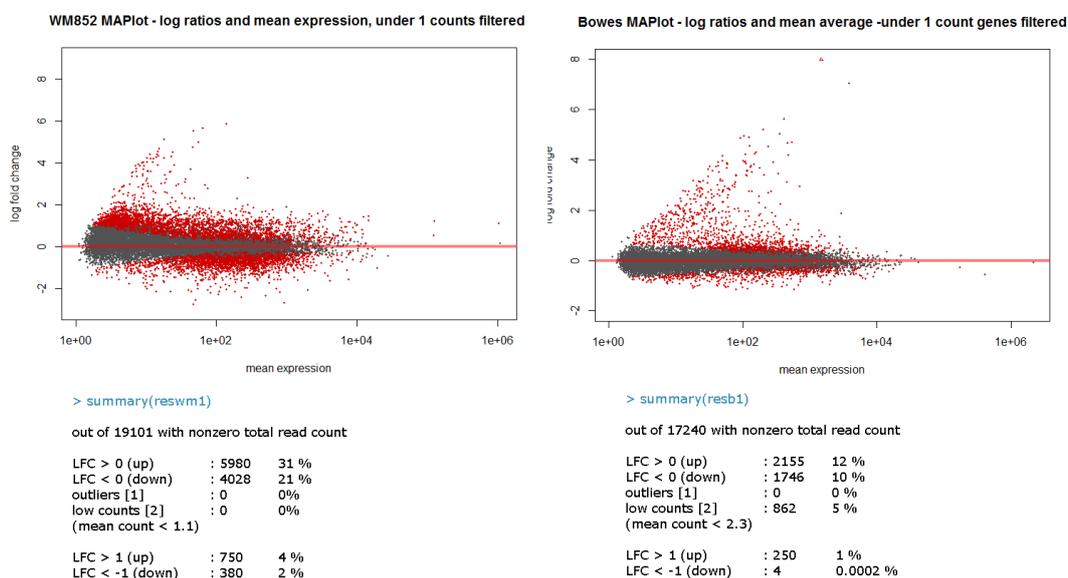


Figure 17. MA plots of the results, showing genes mean expression in relation to its log<sub>2</sub> fold change. On red colour, genes with significant expression changes with  $\alpha < 0.01$ . WM852 left, Bowes right.

From the 62000 gene IDs that were part of the reference genome 19101 (WM852) and 17240 (Bowes) passed the DESeq2 algorithm that also discards genes with low counts. WM852 had a total of 10008 significantly expressed genes, while Bowes had 3901. In the log<sub>2</sub>FC range of less than -1 or over 1, there were approximately 1130 significantly differentially expressed genes in the LEC primed WM852 and 254 in the LEC primed Bowes populations. WM852 had been cultured for 48 hours which was twice as long as compared to Bowes. Culturing time difference could indicate the difference in the amount of genes differentially expressed in the two cell populations. Both populations had remarkably more up regulated than down regulated genes, which might be an indication of the experimental set-up favouring up regulation of gene expression.

## 7.2 Methods of result verification

When performing a large scale RNA-sequencing data analysis the validity is a probable concern (Zheng, et al., 2011). To this end, three validity threats were identified as

possible influencers of the result outcome. First threat for the validity of the results is a mechanical experimental error for example in cell during the culturing phase or in measurement of correct nucleotides in RNA-sequencing. Second threat to validity is the error in analysis phase. Third validity threat lies in the interpretation of the results.

As for the first threat, all cell-handling steps are well documented and repeatable and no minor or major mistakes were discovered. Second threat is more problematic to control as it exposes our data quality to major and possibly minor variations. Major variations are easier to filter out and eliminate as they are more blatant and obvious. Minor variations might contribute to approximately  $\pm 0-10\%$  changes in the final values and results (Rapaport, et al., 2013). Minor errors might be well concealed because the effect might occur at the whole population level and nothing really stands out as a possible error source in the results. However, all the statistical analysis on the R platform can be easily repeated with the same end results. Furthermore all the analysis are done using standard adjustments. The publications describing the procedures offer complete workflows and manuals and are well cited and at the same time. The probability for the third validity threat depends on personal expertise, experience level and on the aspect that the subject is studied. To avoid the threats, all project steps were exposed to peer review and during the thesis work I also had the possibility to compare the differential expression results of WM852 to one previously performed by a graduate student (Icay, 2015) in System Biology group, Genome-Scale Biology Research Program Unit, Faculty of Medicine, University of Helsinki. The Systems Biology group completed their differential expression analysis with their RNA-sequencing data pipeline where Cufflinks is integrated as differential expression analyser (Icay, 2015). Cufflinks is a slightly different method but the results for the top genes should be the same nearly the same (Zhang, et al., 2014). From the same raw data we were able to obtain results that were nearly matching for the top genes. Figure 18 validates that our results did not suffer from major errors and the differential expression analysis was successful.

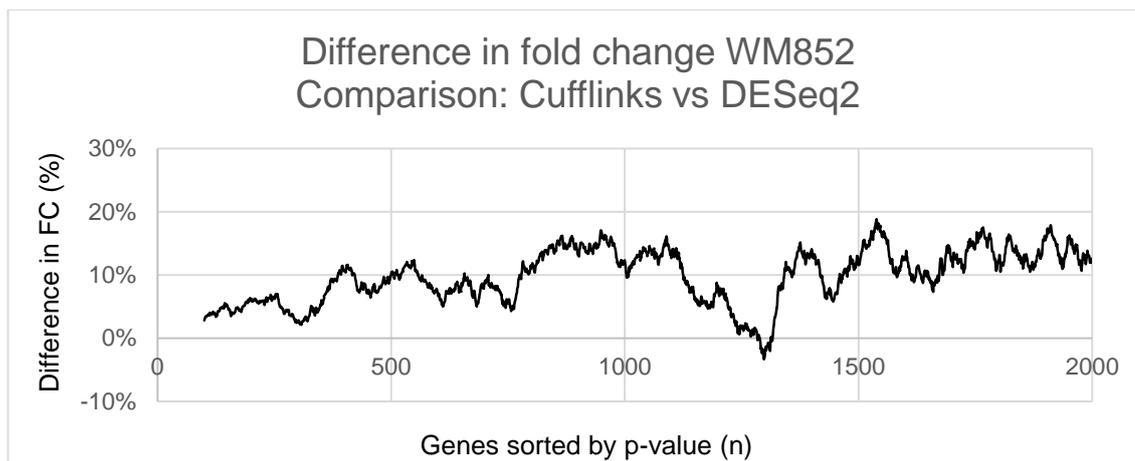


Figure 18. Moving average of fold change difference to validate the result. First 2000 genes sorted by p-value.

Most significant genes show approximately ~5% difference in fold change. The quality varies below the top genes and settles to around 10-15% of difference. However, DESeq2 is more sensitive as an analysis method than Cufflinks (Zhang, et al., 2014), and the two methods use different algorithms. Therefore, the difference between results can be partly explained with this. As the data obtained by DESeq2 turned out to yield similar results as Cufflinks, we chose to use the DESeq2 workflow to analyse RNA-sequence data from the Bowes cells as described in 7.1.

### 7.3 Enrichment and pathway analysis results

Focusing on single genes might be interesting, but for larger scale changes in expression, enrichment result based on KEGG and GO annotations can be more informative (Kaneisha & Goto, 2000). Here we used Generally Applicable Gene-set Enrichment (GAGE) method to perform a large variety of different enrichment group tests with different setups (Luo, et al., 2009). Because of the research in progress in this project, only selected outcome can be presented in this thesis. Figure 19 visualizes a heatmap produced by GAGE showing KEGG and GO enrichment groups. GAGE compares the groups with “paired” comparison where treated and untreated samples are of equal length and one-on-one paired according by the original experimental design (Luo, et al., 2009).

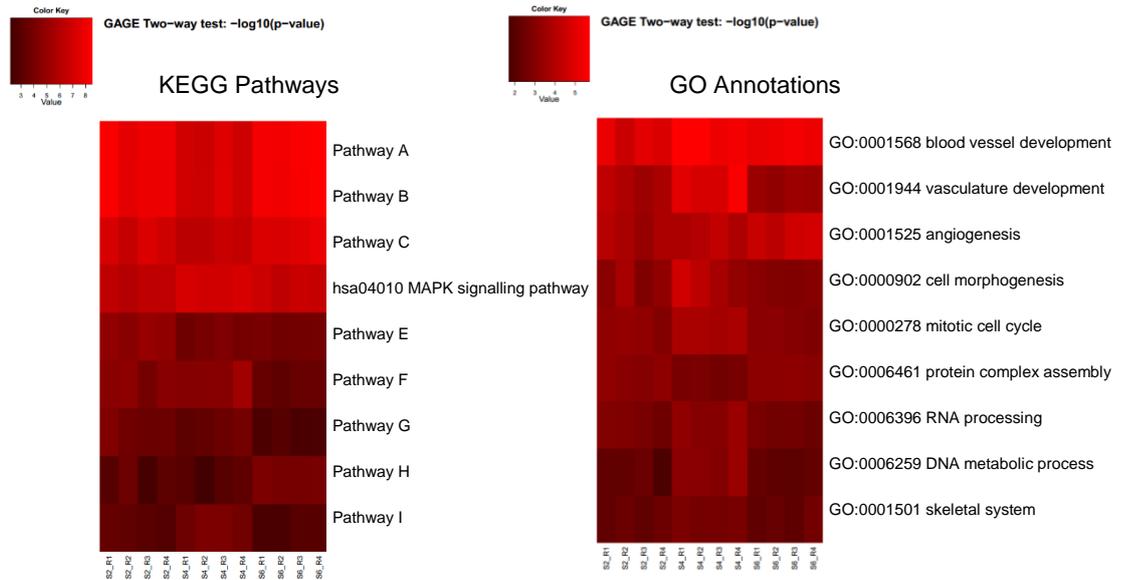


Figure 19. GAGE two directional test analysis, showing KEGG and GO annotation enrichment groups for Bowes cell line, where groups have statistical significance ( $\alpha < 1,0 \times 10^{-21}$ ). Both populations show sets of interesting groups after lymphatic endothelial cell treatment. Brighter red colour means lower p-value.

The primary results of the GAGE analysis are the enrichment scores that are outputted to tab limited text file and heatmaps (Figure 19) that are the visualization of the scores. Heatmaps reflect the degree to which a gene set is significantly expressed. GAGE calculates the scores by gene set differential expression test based on one-on-one comparison between samples from two experimental conditions untreated vs treated. The p-values shown in the Figure 19 are a result of a global p-value based meta-test on the negative log sum of p-values from all one-on-one comparisons. Only gene sets with highest significance  $\alpha < 1,0 \times 10^{-21}$  are presented in Figure 19.

Enrichment results derived from GO show variety of groups that have been shown to be active in malignant cancer development (Penn, 2008). The blood vessel development is a process where aim is the formation and maturing of capillary and vascular vessels and in the end carrying blood to tissues. The vasculature development is an interconnected tubular multi-tissue structure that contains fluid that is actively transported around the organism. Angiogenesis refers to blood vessel formation when new vessels emerge from the proliferation of pre-existing blood vessels. Cell morphogenesis refers to a developmental process in which the size or shape of a cell is generated and organized. Mitotic cell cycle, the most common eukaryotic cell cycle (Figure 2) which canonically

comprises four successive phases called G1, S, G2, and M and includes replication of the genome and the subsequent segregation of chromosomes into daughter cells (The Gene Ontology Consortium, 2015). Cancerous cells might enrich these groups normally if compared to normal equivalent stromal cells. Here remarkably the treatment with LECs cause the cancer cells to significantly express groups shown in Figure 19 compared to same cancer cells.

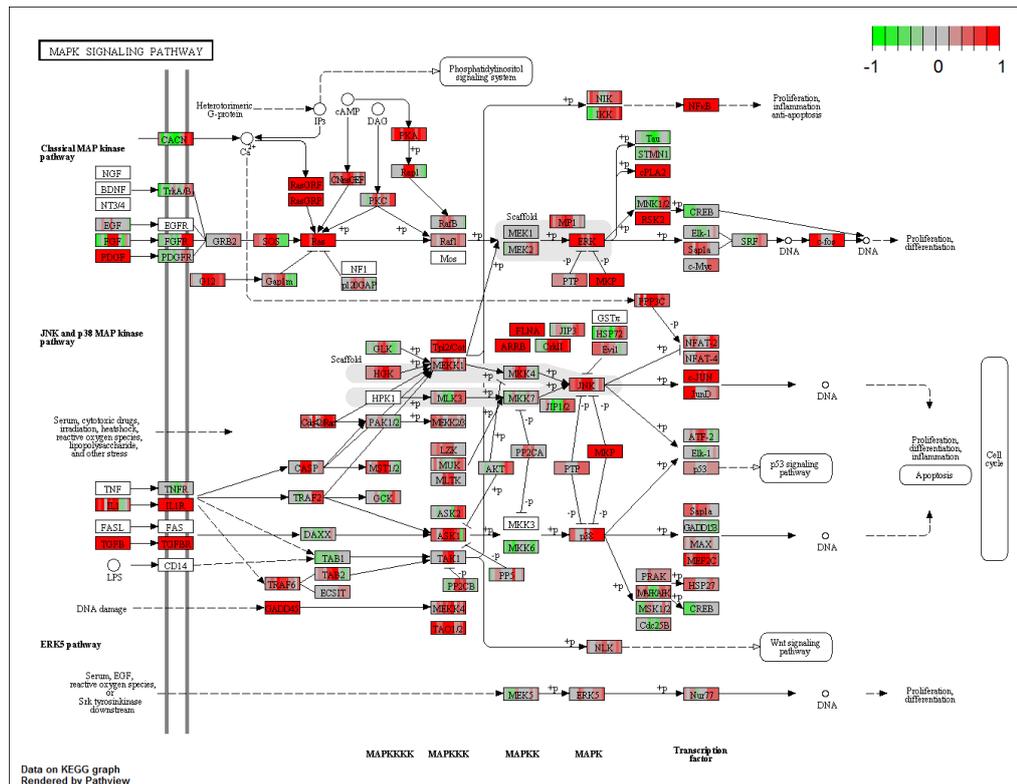


Figure 20. KEGG pathway called MAPK, which is an interesting signalling pathway in melanoma development. It is shown by directional test by GAGE, visualized by Pathview. Most detailed description of the graph can be viewed from [http://www.genome.jp/kegg/document/help\\_pathway.html](http://www.genome.jp/kegg/document/help_pathway.html)

One of the pathways which genes were found enriched in the analysis of the LEC primed melanoma samples was the Mitogen-activated protein kinase (MAPK) pathway. It is composed of a group of proteins that deliver extracellular signals from the cell surface to the DNA in the nucleus (Figure 20). MAPK is highly conserved module involved in proliferation, differentiation and migration of cells. It usually functions as on-off switch for genes, while mutations to this pathway cause uncontrollable growth where tumours and cancer originate from. It is an important target for therapeutical treatments that aim at providing better outcomes for patients. The MAPK signalling pathway shows groups of targeted genes significantly expressed and most of them are upregulated. The single

target could be red, green or grey at the same time as differences in samples and replications is taken into account and visualized in Figure 20. Deregulation of the MAPK pathway in melanomas takes place frequently due to activating mutations in the B-RAF and RAS genes or other epigenetic or genetic alterations, leading to increased signaling activity promoting cell proliferation, invasion, angiogenesis, metastasis, distant migration and survival reference, thus representing a relevant pathway in melanoma (Chin, 2003) (Cargnello & Roux, 2015).

## **8 Conclusions**

Sequencing data revealed extensive gene expression changes in both the LEC primed Bowes and WM852 melanoma cell lines when compared to Bowes and WM852 cell cultures alone. Bowes had fewer significantly differentially expressed genes, but the result could be due to different culturing time. However, both populations responded to the lymphatic endothelial cell treatment with in a similar manner. Enrichment results showed significant expression of pathways related to pathogenesis of cancer and tumour progression. These results indicate that the interaction and crosstalk with the LECs can contribute to the tumorigenic properties of melanoma cell lines WM852 and Bowes.

The most interesting genes identified during in this RNA-seq analysis are currently being validated by qRT-PCR and cell-based functional assays in the laboratory of Prof. Päivi Ojala. To protect the novelty of these findings for the manuscript in preparation, the identity of the genes and majority of additional findings cannot be presented in this thesis.

Finally, it is vital to understand the mechanisms of malignant melanoma development and its ability to metastasize to find ways to improve the treatment modalities of this disease. The research done here will hopefully help not only to build a deeper understanding of the contribution of the stromal microenvironment to melanoma metastasis but also to develop more effective and specific ways to predict melanoma progression and to inhibit metastasis.

## 9 References

- Alberts, B. et al., 2015. *Molecular Biology of The Cell*. s.l.:Garland Science.
- Anders, S. et al., 2010. Differential expression analysis for sequence count data. *Genome Biology*.
- Anders, S., Pyl, P. T. & Huber, W., 2015. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2), pp. 166-169.
- Balch, C. M. et al., 2009. Final Version of 2009 AJCC Melanoma Staging and Classification. *Journal of Clinical Oncology*, Volume 27, pp. 6199-6206.
- Bijian, K. et al., 2013. Targeting focal adhesion turnover in invasive breast cancer cells by the purine derivative reversine. *British Journal of Cancer*, Issue 109, pp. 2810-2818.
- CancerLink - Neoplastic Diseases Reviews, 2010. *MELANOMA BIOLOGY*. [Online] Available at: <http://cancerlink.ru/enmelbiology.html> [Accessed 17 03 2015].
- Cargnello, M. & Roux, P. P., 2015. Activation and Function of the MAPKs and Their Substrates, the MAPK-Activated Protein Kinases. *Microbiology and Molecular Biology Reviews*, 75(1), pp. 50-83.
- Chen, C. S. et al., 2003. Cell shape provides global control of focal adhesion assembly. *Biochemical and Biophysical Research Communications*.
- Chial, H., 2008. Genetic Regulation of Cancer. *Nature Education*, 1(1), p. 67.
- Chi, D. V., 2012. MYC on the Path to Cancer. *Cell*, Volume 149, pp. 22-35.
- Chin, L., 2003. The Genetics of Malignant Melanoma: Lessons from Mouse and Man. *Nature Reviews*, Volume 3, pp. 559-570.
- Chipster, 2015. *Chipster*. [Online] Available at: <http://chipster.csc.fi/> [Accessed 15 03 2015].

Chowdhury, S. & Sarkar, R. R., 2015. Comparison of human cell signaling pathway databases - evolution, drawbacks and challenges. *Database*, pp. 1-25.

Cunningham, F. et al., 2015. Ensembl 2015. *Nucleic acid research*, Volume 43.

Emmett, M. S. et al., 2011. CCR7 mediates directed growth of melanomas towards lymphatics. *Microcirculation*.

Engholm, G. F. J. et al., 2014. *NORDCAN: Cancer Incidence, Mortality, Prevalence and Survival in the Nordic Countries, Version 7.0 (17.12.2014)*. [Online] Available at: <http://www-dep.iarc.fr/NORDCAN/English/frame.asp> [Accessed 1 April 2015].

Erdman, F. et al., 2012. International trends in the incidence of malignant melanoma 1953-2008-are recent generations at higher or lower risk?. *International Journal of Cancer*.

Ewing, B. & Green, P., 1998. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Research*.

Ferlay, J. et al., 2014. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*.

Fonseca, N. A., Brazma, A. & Marioni, J., 2014. RNA-Seq Gene Profiling - A Systematic Empirical Comparison. *PLoS ONE*.

Gingeras, T. R., 2007. Origin of Phenotypes: Genes and transcripts. *Genome Research*.

Gurell, B. et al., 2008. Nuclear MYC protein overexpression is an early alteration in human prostate carcinogenesis.. *Modern Pathology*, pp. 1156-1167.

Icay, K., 2015. *M.SC Ph.D Student* [Interview] (11 04 2015).

Illumina, 2015. *Total RNA Sequencing*. [Online] Available at: [http://www.illumina.com/applications/sequencing/rna/total\\_rna-seq.html](http://www.illumina.com/applications/sequencing/rna/total_rna-seq.html) [Accessed 03 05 2015].

Kallio, A. M. et al., 2011. Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC Genomics*.

Kaneisha, M. & Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1), pp. 27-30.

Kim, S. Y. et al., 2015. Metaanalysis of BRAF mutations and clinicopathologic characteristics in primary melanoma. *Journal of American Academy of Dermatology*.

Langmead, B. & Salzberg, S. L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, Issue 9, pp. 357-359.

Lasken, R. S. & McLean, J. S., 2014. Recent advances in genomic DNA sequencing of microbial species from single cells. *Nature*.

Lesk, A. M., 2014. *Introduction of Bioinformatics*. 4th ed. Pennsylvania: Oxford University Press.

Liu, Y., Zhou, J. & White, K. P., 2014. RNA-seq differential expression studies: more sequence or more replication?. *Bioinformatics*, pp. 301-304.

Love, M., Anders, S. & Huber, W., 2015. *RNA-Seq workflow: gene-level exploratory analysis and differential expression*. [Online] Available at: <http://www.bioconductor.org/help/workflows/rnaseqGene/> [Accessed 21 04 2015].

Love, M. I., Huber, W. & Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*.

Luo, W. & Brouwer, C., 2013. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, 29(14), pp. 1830-1831.

Luo, W., Friedman, M., Shedden, K. & Hankeson, K., 2009. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, Volume 10, p. 161.

Meyerson, M., Gabriel, S. & Getz, G., 2010. Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics*, 11(10), pp. 685-696.

Muller, C., 2008. Histology of melanoma and nonmelanoma skin cancer.. *Adv Exp Med Biol*, 624(59), pp. 215-226.

Nagano, M. et al., 2012. Turnover of Focal Adhesions and Cancer Cell Migration. *International Journal of Cell Biology*.

Nature America, Inc., 2013. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature Protocols*, pp. 1765-1786.

PDQ Melanoma Treatment—for health professionals , 2015. *Melanoma Treatment—for health professionals (PDQ®)*. [Online] Available at: <http://www.cancer.gov/types/skin/hp/melanoma-treatment-pdq#section/all> [Accessed 26 04 2015].

Pekkonen, P., 2015. *Cell plasticity in cancer: Cues from virus host interaction*. Helsinki: University of Helsinki.

Penn, J. S., 2008. *Retinal and Choroidal Angiogenesis*. s.l.:Springer.

Plessis, L., Skunca, N. & Dessimoz, C., 2011. The what, where, how and why of gene ontology—a primer for bioinformaticians. *Briefing of Bioinformatics*, 12(6), pp. 723-735.

R Core Team, 2015. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria: <http://www.R-project.org/>.

Rapaport, F. et al., 2013. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*.

Solunetti, 2015. *Solunetti: virukset ja syöpä*. [Online] Available at: [http://www.solunetti.fi/fi/solubiologia/virukset\\_ja\\_syopa/](http://www.solunetti.fi/fi/solubiologia/virukset_ja_syopa/) [Accessed 08 04 2015].

Suomen syöpärekisteri, 2015. [Online] Available at: <http://www.cancer.fi/syoparekisteri/tilastot> [Accessed 25 04 2015].

Tatti, O. et al., 2011. Membrane-Type-3 Matrix Metalloproteinase (MT3-MMP) Functions as a Matrix Composition-Dependent Effector of Melanoma Cell Invasion. *PLoS ONE*.

The Gene Ontology Consortium, 2015. Gene Ontology Consortium: going forward. *Nucleic Acid Research*, Volume 43.

Wang, Z., Gerstein, M. & Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. *National Institutes of Health*, pp. 57-63.

Zhang, Z. H. et al., 2014. A Comparative Study of Techniques for Differential Expression Analysis on RNA-Seq Data. *PLoS ONE*.

Zheng, W., Chung, L. M. & Zhao, H., 2011. Bias detection and correction in RNA-Sequencing data. *BMC Bioinformatics*.

## Chipster run parameters

Step 1

-----

Dataset name: A006-co-culture-elute-3-exp-GAATTCGT-run20150316N\_S6\_L004\_R1\_001.fastq.gz  
Created with operation: Import / Import data

Step 2

-----

Dataset name: trimmed.fastq.gz  
Created with operation: Preprocessing / Trim reads for several criteria with PRINSEQ  
Parameter Input file format: FASTQ  
Parameter Base quality encoding: Sanger  
Parameter Trim 5-prime end by quality: 0  
Parameter Trim 3-prime end by quality: 5  
Parameter Quality score calculation method: minimum quality value  
Parameter Quality score comparison condition: less than  
Parameter Window size for quality calculation: 1  
Parameter Step size used to move the quality window: 1  
Parameter Trim left A/T tails: 10  
Parameter Trim right A/T tails: 10  
Parameter Trim left poly-N tails: 10  
Parameter Trim right poly-N tails: 10  
Parameter Minimum length: 20  
Parameter Write a log file: yes

Step 3

-----

Dataset name: tophat.bam  
Created with operation: Alignment / TopHat2 for single end reads  
Parameter Genome: Homo\_sapiens.GRCh38  
Parameter Use annotation GTF: yes  
Parameter When GTF file is used, ignore novel junctions: no  
Parameter Base quality encoding used: Sanger - Phred+33  
Parameter How many hits is a read allowed to have: 20  
Parameter Number of mismatches allowed in final alignment: 2  
Parameter Minimum anchor length: 8  
Parameter Maximum number of mismatches allowed in the anchor: 0  
Parameter Minimum intron length: 70  
Parameter Maximum intron length: 500000  
Parameter Library type: fr-unstranded

Step 4

-----

Dataset name: htseq-counts.tsv  
Created with operation: RNA-seq / Count aligned reads per genes with HTSeq  
Parameter Reference organism: Homo\_sapiens.GRCh38.78  
Parameter Chromosome names in the BAM file look like: 1  
Parameter Does the BAM file contain paired-end data: no  
Parameter Was the data produced with a strand-specific protocol: no  
Parameter Mode to handle reads overlapping more than one feature: union  
Parameter Minimum alignment quality: 10  
Parameter Feature type to count: exon  
Parameter Feature ID to use: gene\_id  
Parameter Add chromosomal coordinates to the count table: no

## **R run parameters and script**

R script is located at: <http://pastebin.com/f9Yrvibv>