Transcriptome analysis of cartilage in homeostasis and osteoarthritis using deep sequencing technology

Yaobo Xu

Doctor of Philosophy

Institute of Cellular Medicine

February 2015

Abstract

Osteoarthritis (OA) is the most prevalent type of joint diseases. It is associated with the progressive degradation of articular cartilage and disease progression can lead to total destruction. In order to study the molecular changes in OA cartilage, I compared the transcriptome of OA and healthy cartilage using two technologies, expression microarrays and the recently introduced RNA sequencing (RNAseq) technology. RNAseq is based on the next-generation sequencing, enabling the investigation of the transcriptome in single nucleotide resolution. In this PhD project, I optimized the cartilage RNA extraction protocol first and then used the optimized protocol to extract RNAs from both OA and healthy cartilages samples. Before the RNAseq experiment, the quality of the RNA samples was checked with real-time PCR and gene expression microarray experiments. Using microarray and RNAseq data, I found novel OA associated genes and canonical pathways. With the RNAseq data, the knowledge of the OA transcriptome was further extended, including differentially expressed transcripts, novel transcripts in cartilage, alternative splicing events and differential allelic expressions. The performance of the RNAseq was also compared with the microarray technology, revealed both advantages and the limitations of the technology.

Acknowledgements

Upon the completion of the thesis, I would like to thank all of my supervisors, Prof. David Young, Dr Mauro Santibanez-Koref, Prof. John Loughlin, Dr Daniel Swan and Prof. Drew Rowan, for their kind and patient supports, especially David and Mauro who gave me the most supports and encourages. I would like to thank my parents who provide me the fees for my study and their kindest love. My wife takes the credit as well. The beautiful woman supported me, took care of me and married me at last before I completed my student life. Thanks you all of the colleagues who shared their knowledge with me and helped me though the whole study.

Table of contents

Lists of tables and figures

Figures:

Tables:

Additional Table (not printed because of their sizes. Available online https://github.com/byb121/Thesis_2015/tree/master/Thesis_2015/Additional%20tables):

List of Abbreviates

BP	Biological process
CC	Cellular components
cDNA	Complementary DNA
ChIP	Chromatin immunoprecipitation
CpG	Cytosine-phosphate-guanine
DAVID	Database for Annotation Visualization and Integrated Discovery
ddNTPs	Dideoxynucleotides triphosphates
DE	Differentially expressed
dNTP	Deoxynucleotide triphosphates
ENCODE	Encyclopedia of DNA Elements
FDR	False discovery rate
FPKM	Fragments per kilo-bases of transcripts per million reads
GO	Gene ontology
GSEA	Gene set enrichment analysis
INDELs	Insertions and deletions
IPA	Ingenuity pathway analysis
lincRNA	Long non-coding rnas
MAQC	Microarray Quality Control
MeDIP	The combination of Methyl-DNA immunoprecipitation
MF	Molecular functions
MiGS	MBD- isolated Genome Sequencing
MPSS	Massively parallel signature sequencing
NGS	Next generation sequencing
NOF	Neck of femur
OA	Osteoarthritis
PCA	Principal component analysis
PCR	Polymerase chain reaction
RIN	RNA integrity value
RNAseq	RNA sequencing
RPKM	Reads Per Kilo bases of the gene per Million bases of read
RRBS	Reduced representation bisulphite sequencing
SAGE	Serial analysis of gene expression
SNPs	Single nucleotide polymorphisms
TSS	Transcription starting sites

Chapter 1 Introduction

1.1 Osteoarthritis

Osteoarthritis (OA) is the most prevalent type of joint disease, which is associated with the progressive degradation of articular cartilage and total destruction with the disease progression. The deficiency of the cartilage cushion between bones of the joint correlates with pain, stiffness and eventually loss of mobility (Fabio, 1998). It is commonly observed in the joint of the hand, knee, hips and spine (Felson, 2006). OA affects at least 5.2 million people in UK and around 60% people aged over 60 years of age (Campian, 2002). Because of the high incidence rate of the disease and the long term of medical needs of the OA patients, the disease creates a considerable social and economical burden.

OA is a complex disease. The disease is now recognized as a different process from aging but an age-related joint disorder, as aging decreases the chondrocytes' ability to maintain the healthy status, therefore the risk of OA development increases (Loeser, 2009). Pathology of the disease involves multiple factors, including mechanics, environment and genetics.(Felson *et al.*, 1997; Mahr *et al.*, 2006; Iliopoulos *et al.*, 2007) Mechanical factors, such as increased body weight burden on the knee and hip joints caused by obesity, are considered to have predominantly a role in the initiation of the disorder (Pelletier *et al.*, 2001). Genetic factors also play an important role in OA development. Since 1941, certain types of OA have been proved to be strongly related to genetics (Stecher and Hersh, 1944). More recent studies have shown a greater correlation of the disease status in identical twins than in non-identical twins, indicating a role for genetic factors in OA predisposition (Fernandez-Moreno *et al.*, 2008). In recent years, a number of genome-wide associated studies (GWAS) have been conducted to search for candidate genes associated with the OA susceptibility (Nakamura *et al.*, 2007; Mototani *et al.*, 2008; Dieguez-Gonzalez *et al.*, 2009; Valdes *et al.*, 2010). One of the largest studies, "arcOGEN" studied 7410 patients and revealed 5

loci that are significantly associated with the disease (Zeggini *et al.*, 2012). Another GWAS study of European and Asian population showed a single nucleotide polymorphism of *GDF5* is strongly associated with OA susceptibility (Chapman *et al.*, 2008). The outcomes of these studies not only contribute to the understanding of the OA mechanism but also can be used to identify high-risk individuals (Valdes and Spector, 2010). However, follow on studies of these have not answered neither the cause of the disease or the cure so far.

The complexity of OA is also due to the fact that the whole joint is involved in the disease process (Brandt *et al.*, 2006). As chondrocytes are likely to be involved in the initiation and progression, studies of molecular changes of chondrocytes during the development of OA are more likely to provide insights into the genetic mechanisms of this joint disease.

Different therapies have been used but so far none is able to reverse the OA progression. Current treatments for OA can vary depending on the severity of the disease symptoms but they all concentrate on pain relief and the symptom of inflammation. Non-pharmacological treatments, such as self-management and exercise provide only limited benefits comparing to pharmacological treatments. Pharmacological treatments, such as the use of acetaminophen and nonsteroidal anti-inflammatory drugs are effective on pain relief (Kennedy and Moran, 2010). However, these symptomatic treatments lack the ability to stop or reverse disease progression and eventually many patients will need joint replacement surgery, which currently remains the only cure.

In the surgery the affected joints are replaced with plastic or metal implants. It helps to reduce the pain and reinstall some function of the joint. However, the surgery has risks such as myocardial infarction and dislocation of the implants (Katz, 2006) and some patients cannot be persuaded to undertake the surgery because of their concerns regarding the associated risks (Hamel *et al.*, 2008). Gene-based therapy, a novel approach enabled by the advances in the knowledge of the disease and the improvements of the gene transfer methods, promises site-specific treatment and long-term solution of OA (Madry and Cucchiarini, 2013). This emphasizes the importance of further our understanding of the molecular mechanisms of the OA and identification of potential gene targets.

1.2 Articular cartilage

1.2.1 Cartilage and extracellular matrix

Cartilage is an avascular, aneural and alymphatic tissue that can be found in many places in the human body. Articular cartilage surrounds the ends of long bones to provide protection and cushion during use. Chondrocytes are the only cell type found in cartilage. However, they represent less than 2% of total tissue mass and less than 5% of the total volume (Adams *et al.*, 1992; McKenna *et al.*, 2000). Cells in cartilage are enclosed by a highly cross-linked extracellular matrix (ECM) consisting mainly of collagen and proteoglycans. More than 40 different molecules have been identified in ECM with different structures, functions and distributions. Some of the molecules whose functions have been revealed are related to the genetic disorder (Roughley, 2001). The structure of ECM provides protection for chondrocytes and the elasticity of the tissue, which is necessary for drawing water back to the matrix after compressed. This pumping action helps nutrition supply of chondrocytes.

As the destruction and loss of the articular cartilage is the most obvious feature of OA, the pathology of cartilage degradation has received intensive interest of researchers. Earlier studies in 1950s revealed composition differences between fibrillar cartilage and healthy cartilage of the same joint (Matthews, 1953). Lately studies have presented evidence showing changes of inflammatory mediator activities resulting in the cartilage destruction (Loeser, 2008). As the sole cell type in cartilage, chondrocytes are often the focus of research. Initial stages of OA features increased chondrocyte proliferation and synthesis of ECM proteins, growth factors, cytokines, and other inflammatory mediators (Loeser, 2008). With OA progression, cells enter into a catabolic state with increased expression of a number of proteins including matrix metalloproteinase (MMP) *(MMP1, MMP3, MMP9, MMP13 MMP14*), aggrecanases (*ADAMTS5, ADAMTS4, ADAMTS9*), regulatory proteins (*IL1*, TNFa, toll-like receptors, etc.), matrix proteins (collagens type II and X, aggrecan, etc.) and several transcription factors, such as *SOX9* (Goldring and Goldring, 2010).

It is now known that OA involves not only the cartilage but also the ligaments, periarticular muscle, nerve, subchondral bone, meniscus and synovial fluid (Brandt *et al.*, 2006; Loeser, 2008). However, supported by the fact that OA progress can be halted by preventing cartilage loss (Glasson *et al.*, 2005) and that chondrocytes are involved in the initiation and progression of OA, detailed study of the molecular and genetic changes occurring within cartilage during the disease development is more critical to interpret the pathology of OA.

1.2.2 RNA extraction from cartilage tissue

RNA extraction from cartilage is problematic because of the low density of cells and firmly cross-linked proteoglycan ECM network. Furthermore, as most cartilage material is coming from patients who are undergoing joint replacement, limited amounts of cartilage can be collected. The quality assessment of extracted RNA from cartilage also appears to be complex (Clements *et al.*, 2006). In addition, the current genome-wide expression profiling technologies, such as cDNA microarray, are demanding both in terms of the quantity and the quality of the input RNA required for accurate measurements. Therefore, it is critical to use a protocol that can extract high quality RNA from relatively limited amounts of cartilage. The guanidinium thiocyanate-phenol-chloroform extraction method (Chomczynski and Sacchi, 1987) is commonly used to extract RNA from cartilages. However this includes a few modifications of the original method, such as:

1. Tissue is snap frozen and milled under very low temperature (Fabio, 1998);
2. Trizol (Invitrogen Life Technologies), a mono-phasic solution of phenol and guanidine isothiocyanate, is introduced into the method;
3. Membrane binding-washing system (RNeasy mini kit, etc.) is used to purify RNA instead of precipitation, which increases RNA yields and purity.

The recently invented phenol/chloroform-free filter-based system (RNAqueous™) has also been successfully demonstrated for RNA extractions from human and bovine cartilage and showed better results in terms of RNA quality and yield (Ruettger *et al.*, 2010).

1.3 Next generation sequencing

1.3.1 The evolution of nucleotide sequencing

Sanger sequencing has been widely used to determine DNA sequence since it was first introduced in 1977 (Sanger *et al.*, 1977). The crux for Sanger's method is to utilize dideoxynucleotides, which will terminate a DNA chain after being added onto it. The process usually starts with the amplification of DNA fragments by cloning, or polymerase chain reaction (PCR). The amplified product is then mixed with fluorescent labelled dideoxynucleotides triphosphates (ddNTPs) in four different colours (for A, T, G and C), normal nucleotides and a polymerase which catalyzes the extension of a DNA. After a number of rounds of denaturation, primer annealing and primer extension, a set of DNA copies with labelled ends is generated. Each base of a DNA fragments can be determined by using capillary-based technology to separate end-labeled DNA copies by size. Modern Sanger sequencing can determine the sequence of DNA fragments with 1000 base pairs in length and costs approximately $0.50 per kilobase (Shendure and Ji, 2008).

The correct order of the sequence of Sanger sequencing is depending on the size separation step and the sequence can only be viewed after this. Newly introduced "sequencing by synthesis" approaches allow the sequence to be read in real-time (Ronaghi *et al.*, 1996). In these methods, the single-stranded DNA is immobilized on a solid surface first and then a sequencing adaptor is hybridized. Using the single-strand as template, the second stand is synthesised with the repeated cycle of incubation with different deoxynucleotide triphosphates (dNTP) and washing. The DNA polymerase catalyzed extension of a dNTP will release the pyrophosphate (PPi), which can trigger a light emission through ATP sulfurylase and luciferase. The strength of the light signals is correlated with the number of the same dNTP being added. By continuously monitoring of the light signals, the bases and their orders of the DNA sequence will be revealed. (Nyren *et al.*, 1993; Ronaghi *et al.*, 1996).

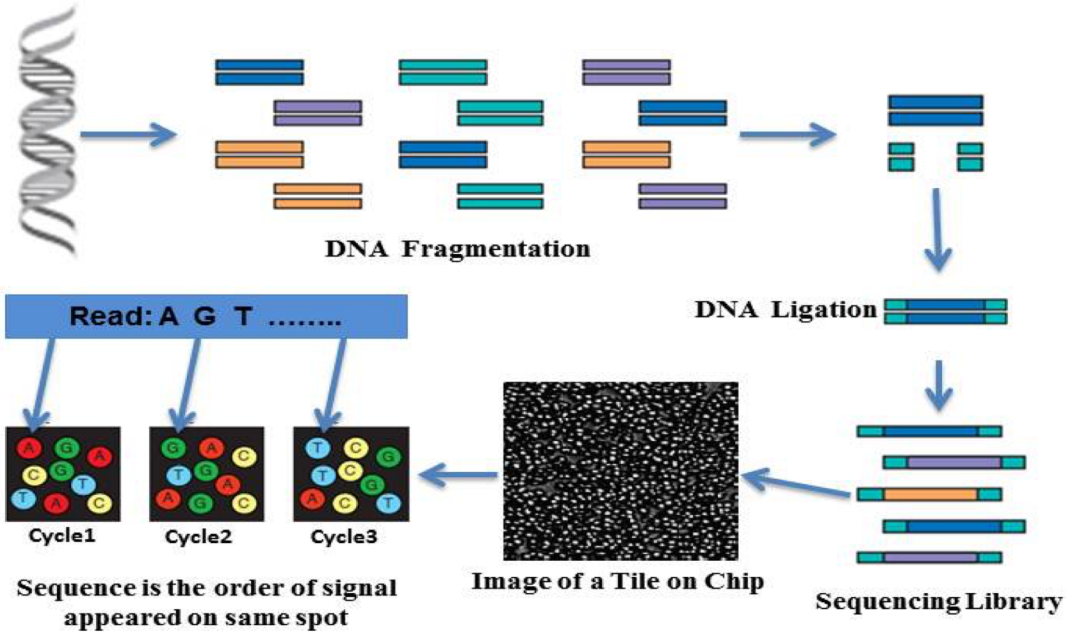


Figure 1.1 Basic work flow of NGS. The DNA is fragmented first then ligated with sequencing adaptors to construct the sequencing library. As the adaptors are complementary to the shorts sequences printed on the flow cell, these fragments can be fixed on the surface of the flow cell. After several rounds of PCR amplification, a cluster of identical short sequences is formed for each of the fragments. Multiple-rounds of sequencing-by-synthesis is then started on each cluster, light signals emitted are captured by a high-resolution camera and stored as images. Eventually a sequencing read can be derived from a series of signals of a same position of the images.

Next generation sequencing (NGS) is based on pyro-sequencing and cyclic array strategy (Shendure *et al.*, 2005). In NGS, different colour labelled ddNTPs are used instead of normal dNTP. The mixture of the 4 ddNTPs are used to extend the templates of the sequenced DNA instead of only one type dNTP in the extension cycle. A laser is then used to excite the light signals and a high-resolution camera will take photos to record the coloured signals. The incorporated ddNTPs are then chemically modified in the washing cycle to allow further extension. As the DNAs are immobilized on a surface, a sequence can be derived from a series of such coloured signals from the same position of the photos. With high-resolution cameras, extensions of thousands of DNA sequences can be recorded on one photo, which enabled parallel sequencing on a massive scale.

Although this technique is provided by several commercial platforms, such as Roche Applied Science 454 sequencer, Illumina Solexa Genome Analyze and the Applied Biosystems SOLiD platform, the basic workflow of each is similar (Figure1.1). Starting with random fragmentation of genomic DNA, the process then requires these uniform sized fragments to be ligated with two unique adapters. Then PCR amplification of each fragment is started at a fixed spot on the surface of a flowcell. After numbers of cycles of sequencing-by-synthesis of these clones, colour changes of a spot are recorded by a high-resolution camera on images. Sequence reads can then be determined by analysis of these images. The key of NGS is to amplify one DNA fragment on a fixed spot, so that the same position on an image always indicates a cluster of clones from a single DNA fragment. Interestingly, compared to magnetic beads based amplification used by 454 sequencer and SOLiD system, Illumina Genome analyzer uses a unique bridge PCR technique. NGS also allows reads to be generated from both ends of a DNA sequence, called paired-end reading and theoretically increases the mapping accuracy.

To date (May 2014), the human genome can be sequenced to 15-fold coverage within a week with cost of around £1,300, comparing to the 13 years world-wide efforts and $2.7 billion cost of the Human Genome Project launched in 1990 (Lander *et al.*, 2001). Read length has also achieved more than 100bp.

1.3.2 Applications of NGS

With the rapidly and continually dropping cost of sequencing in recent years, multiple applications of the NGS have been developed and large number of studies have been accumulated (240 NGS published studies by December 31, 2011 (Nyren *et al.*, 1993). It can certainly be used in whole genome re-sequencing, but because of the cost issue most of re-sequenced organisms are small genomes. Other applications include:

RNA sequencing (RNAseq): A population of RNA can be deep sequenced after reverse transcripted into cDNAs. Transcriptome derived by RNAseq not only can be used to identify and quantify known/unknown genes but also contains rich information of single nucleotide polymorphisms (SNPs), insertions and deletions (INDELs), structure of transcripts and single base resolution of genes positions. Such information can be used to determine alternative splicing events, RNA-editing events and allelic expressions, etc.

ChIP-seq: By using chromatin immunoprecipitation (ChIP) before high-throughput sequencing (Seq), this method allows binding sites of transcription factors and DNA-binding proteins to be identified genome widely (Raha *et al.*, 2010). By using ChIP-seq technique, Dustin Schones reported that position shifting of nucleosome between activated and rest human CD4 + T cells (Schones *et al.*, 2008).

Exome sequencing: Instead of sequencing the whole genome, exons can be pulled out from genomic DNA by using custom/commercial DNA tiling arrays to construct sequencing libraries. It significantly reduces the sequencing cost while still represent the most functionally relevant sequences. A number of studies have used the technique to identify pathogenic variants for a variety of mono-allelic diseases (Choi *et al.*, 2009; Ng *et al.*, 2009; Kim *et al.*, 2010; Ng *et al.*, 2010; Walsh *et al.*, 2010).

Genome-wide methylation study: The combination of Methyl-DNA immunoprecipitation (MeDIP) assays with NGS provides positions of methylated cytosine-phosphate-guanine (CpG) sites of a whole genome. (Jacinto *et al.*, 2008) But considering the specificity of Anti-5-methyl cytosine antibody used in MeDIP, David Serre's group recently demonstrated a new approach called MBD- isolated Genome Sequencing (MiGS), which utilize methyl CpG binding domain precipitation of genomic DNA to select fragments with methylated CpG sites (Serre *et al.*, 2010). However, the specificity of this method is also in doubt. (Thu *et al.*, 2010) Direct sequencing of bisulphite treated DNA can reveal methylated CpG sites as well, as the sodium bisulphite can convert unmethylated cytosine to uracil while methylated cytosine remain unchanged. (Chatterjee *et al.*, 2012) The downside of this protocol is the expensive cost of sequencing the whole genome. While later introduced Reduced Representation Bisulphite Sequencing (RRBS) (Meissner *et al.*, 2005), on the other hand, requires only sequencing 1% of the whole genome but still contain the sequences from the majority of the promoter regions and other relevant genomic regions. (Gu *et al.*, 2011)

1.3.3 Data analysis of NGS

NGS is high-throughput, thus produces large amount of data. A typical sequenced human exome with averagely 50-fold coverage expected, will contain more than 50 million reads of 70 base in length, the total length of which is greater than the whole

human genome, as the distribution of the reads is usually not evenly spread. Dealing with data of this scale in an acceptable time length (several days) is demanding both in terms of the computing resources as well as bioinformatics software. Sophisticated workflows for the analysis of the data are also necessary. When study purposes are more specific, the workflows can be more complicated. These are discussed in the following sections.

1.4 Transcriptome

1.4.1 Transcriptome with RNA sequencing (RNAseq)

The transcriptome is the total set of the transcripts and their quantities in a cell, including mRNAs, small RNAs and other non-coding RNAs. It reflects genes that are actively expressed in a cell at a development stage or under a physiological condition (Wang *et al.*, 2009). Transcriptome analysis is necessary for studying gene expressions and regulation of a genome, as it presents critical information on classification of transcripts, transcriptional structures and abundances of genes. The human genome is complex in terms of its structure variations (Korbel *et al.*, 2007) and polymorphisms (Sachidanandam *et al.*, 2001), which also leads to the complexity of the transcriptome, alternative splicing events for example (Pan *et al.*, 2008). These strengthen the need for accurate characterization of genes, such as their expression abundances in different tissues/organisms, transcription starting sites (TSS) and alternative splicing patterns, in order to understand better their functions and regulation in specific biological processes.

The transcriptome can be analysed using several techniques/methodologies. The hybridization-based methods (microarrays) were introduced to map mammalian genomes in 2002 (Kapranov *et al.*, 2002). These methods are characterized by the step where fluorescently labelled cDNA are allowed to hybridize with designed high-density oligo-arrays. Since then, microarrays are widely used for genome-wide expression profiling analysis. It is high throughput and relatively inexpensive but suffers a number of limitations, such as sophisticated array design to ensure specific detections and also to avoid hybridization between probing-oligonucleotides (Casneuf *et al.*, 2007), limited

detection range and difficulties in comparison of results from different experiments. (Wang *et al.*, 2009)

The tag sequencing based methods, such as serial analysis of gene expression (SAGE) (Harbers and Carninci, 2005) and massively parallel signature sequencing (MPSS) (Brenner *et al.*, 2000), were then developed and designed to address the problems above. In these methods, instead of using hybridization a specific part of mRNAs are sequenced first and then the products are mapped to the known genes in the databases to determine which genes were detected and their relative abundances. These methods have better specificity but are very time consuming and expensive for large-scale genomes, owing to the Sanger sequencing as part of the process (Velculescu *et al.*, 1995).

Based on the newly invented NGS technology, RNAseq provides an innovative insight of the transcriptome compared to the conventional techniques. Common RNAseq protocols comprise the following steps: a set of RNAs are fragmented first and then reversed-transcribed into cDNA by random priming; these uniform sized cDNAs are then sequenced on a next generation sequencing machine using either single end or paired-ends sequencing strategy. Millions of sequencing reads are then generated and analysed with a bioinformatics workflow. The number of reads mapping to a gene is expected to be proportional to the gene abundance, thus the expression of genes can be inferred. Slight modification of the input material can adjust the sequencing focus, for example adding a poly-A tail filtering step before the reverse transcription can reduce the amount of reads from tRNA, rRNAs and other types of RNAs, thus increase the yields of reads from mRNAs, while micro RNAseq protocols include a size selection on RNAs before sequencing (Morin *et al.*, 2008). More recently, stranded specific RNAseq sequencing has been introduced to ensure only one strand of cDNAs were amplified for sequencing (Levin *et al.*, 2010).

With RNAseq, not only the gene expression changes can be measured in a more sensitive and comprehensive manner comparing to expression microarray analysis, but also information of other aspects of RNA molecules be measured, including expression changes of individual transcript variants (Trapnell *et al.*, 2010), splicing patterns (Pan *et al.*, 2008; Sultan *et al.*, 2008), sequencing variants (Barbazuk *et al.*, 2007; Medvedev *et*

al., 2009; Montgomery *et al.*, 2010; Li *et al.*, 2011a), accurate TSS mapping (Sultan *et al.*, 2008), allelic specific expressions (Heap *et al.*, 2010) and RNA-editing events (Picardi *et al.*, 2010; Peng *et al.*, 2012). Furthermore, as the technology does not require previous knowledge of the sequence of a gene to detect its expression, it is also possible to discover novel genes and transcripts (Pepke *et al.*, 2009). Allelic expression analysis and RNA-editing events were also of interest in the OA field. Therefore using RNAseq on OA cartilage may provide novel insights into the molecular changes in OA. More importantly, RNAseq requires relatively small amounts of RNAs to start (as low as 250pg (Ozsolak *et al.*, 2010)). This is particularly useful when RNA material availability is limited, such as healthy articular cartilage, although to ensure the sequencing quality of my study, several micro grams of RNAs were used. Recent advances in the technology even allow transcriptome profiling from a single cell (Tang *et al.*, 2009).

1.4.2 Transcriptome analysis of OA

At the start of this PhD project, there were a number of gene expression studies of OA and OA vs. normal articular cartilage studies, while comprehensive and genome-wide studies were rare before 2010. In 2001, Thomas Aigner's group found that several metalloproteinases were expressed differentially between early stage and end stage of OA. (Aigner *et al.*, 2001) Later in 2006, they reported different gene expression profiles between normal and early/late stage OA chondrocytes (Aigner *et al.*, 2006a). However, the studies could be biased by using chondrocytes as the source of RNA. Interestingly, the group also concluded that the gene expression profiles of cartilage with macroscopically less damage or even a normal appearance from a late-stage OA joint are still significantly different from healthy cartilage expression profiles. Similarly more recent findings were reported that the chondrocytes from intact and fibrillated OA cartilage of a single joint have the same total mRNA expression profiles.(Tew *et al.*, 2014) Other gene expression profiling studies suffer from the use of smaller data sets or are based upon animal models (Meng *et al.*, 2005; Sato *et al.*, 2006; Appleton *et al.*, 2007; Dell'accio *et al.*, 2008; Geyer *et al.*, 2009; Swingler *et al.*, 2009b). In 2010, Karlsson et al., (Karlsson *et al.*, 2010a) published the first comprehensive gene expression comparison study of OA and healthy knee cartilage using genome-wide cDNA microarray. In the study, over 1,400 genes reported as significantly changed with

around 60 genes of them not previously associated with OA. The finding exhibits the power of new technology in OA studies. However, the study did not correct the p-values for multiple test correction, thus false positives are likely to be introduced. To conclude, although OA is a complex and the most common form of joint disease and articular cartilage is the most evidently affected tissue by the disease, a comprehensive and accurate study of the molecular mechanism is still in need.

1.4.3 Bioinformatics for RNAseq data

The vast amount of data generated from next generation sequencing machines is only useful if interpreted correctly using bioinformatic tools. The next generation sequencing machines amplify and sequence a DNA fragment on a fixed spot, so that high-resolution photos can be taken to record colour changes of a spot after each round of pyro-sequencing. Hence, raw data of sequencing result should be considered as series of images rather than readable text files. Though sequencing platforms always come with software supplied by manufacturers to transform these images into sequencing reads. A process called Base Calling, there is evidence suggesting that the bioinformatician should make careful decision before using base calling software (Kao *et al.*, 2009). However, in reality, due to the difficulties in transferring and storing image files, sequencing service providers usually conduct the base calling stage. Thus sequence reads are the starting point of most bioinformatics analysis.

The analysis of RNAseq reads can be categorised into the following: gene/exon expression analysis, transcripts expression analysis, identification of transcripts (including their sequences and structures), identification of alternative splicing events, and identification of RNA sequence variants. The variants can be further used for allelic expressions analysis, identification of RNA-editing events expression quantitative traits loci (eQTL) analysis etc. As the size of RNAseq data is usually large (≥ 3 gigabytes for each sample in our experiment), all of the analysis of RNAseq data require intensive bioinformatic efforts. In 2011, there were only a few commercial software packages which could analyse RNAseq reads, such as Genespring GX 12 (Agilent Technologies, Inc, California, USA) or CLC Genomics Workbench (CLC bio, Aarhus, Denmark). But due to the cost of these software packages and their relatively limited functionality, freely available open source software was more popular in the next-generation

sequencing data analysis community. The major disadvantage of such open source software is that each of the programs/packages performs only one aspect of the analysis of RNAseq data. Thus for a complete analysis several pieces of software need to be linked into a workflow. The programs/packages often lack interactive interface, thus require at least a basic programming knowledge from the user. At the time, there were a number of open source software tools available, such as FastQC (Andrews), GSNAP (Wu and Nacu, 2010), Tophat (Trapnell et al., 2009), Bowtie (Langmead et al., 2009), DESeq (Anders and Huber, 2010), Diffsplice (Hu et al., 2013) etc, although only some of them were properly documented and maintained, and not all were published. As there was no commonly recognized protocol for RNAseq analysis available and most of the software tools present different characters in terms of running times, hardware requirements, and focuses of software design, it was a challenge to choose which tools to use to assemble a pipeline for our project.

Quality control of the raw reads

The quality control of short reads of sequencing data is the first step. RNAseq data is produced by the high-throughput next generation sequencing machine, thus contains millions of short reads, and the quality control is complicated. FastQC is simple to use and allows a quick check of several statistics describing the quality of reads. Several very useful statistics include: Total Number of Reads contained in the sequencing results, Per Base Sequence Quality, Per Base Sequence Content, Per Base N Content, Sequence Length Distribution and Duplicated/Over-Represented Sequences. Each of these represents a unique insight of into the quality of reads: Total Number of Reads is the most important quality indicator, as fewer reads in RNAseq results lack of detection of genes that have relatively lower expression, thus reduces the integrity of the transcriptome; Per Base Sequence Quality can reveal the fraction of low quality bases at each read position, this represents the reliability of the reads; Per Base Sequence Content exhibits base content of each read position of all reads, un-even distribution of the content means the sequencing library is not random, either containing over-represented sequences or biased fragmentation when generating the library using hexamer random priming (Hansen *et al.*, 2010); Per Base N Content can reveal if any base position has “N”, which indicate the base can be any of the 4 bases; Sequence Length Distribution shows if all reads are of the same length; Duplicated/Over-

Represented Sequences detects over-represented sequences. FastQC also compares the sequences with known sequencing adaptors to identify adaptor contamination. With this information, one can decide the key parameters for quality control, such as the quality score threshold, how many low quality bases need to be trimmed from the ends of sequencing reads and whether adaptor need to be removed. Adaptor sequences can be removed with Cutadapt (Martin, 2011). Trim Galore (Krueger) uses FastQC and Cutadapt to automate the quality control of short reads including both low quality bases removal and adaptor sequences removal. For paired-end sequence data, it also maintains the reads in pairs after the QC process. Many aligners used to map reads require this maintenance of paired ends.

Mapping Software (Aligner)

By 2011, several software tools had been developed to align short read sequences to reference genomes. The earliest aligners include ELAND (Cox, 2007) and Maq (Li *et al.*, 2008). Their accuracy has been proven in several studies. (Schmidt *et al.*, 2008; Maher *et al.*, 2009; Perkins *et al.*, 2009; Xue *et al.*, 2009) They both use similar algorithms that are based on hash tables. Comparing to ELAND, Maq is distributed as open source but it does not support multiple threads, which makes it very time consuming when dealing with large dataset on common desktop machines. Compared to aligners that are based on hash tables, aligners using Burrows-Wheeler transformation, such as BWA (Li and Durbin, 2009) and Bowtie are faster and require much less memory, but tend to be less sensitive. All of the aligners mentioned above only support small gapped (less than 7bp) alignments, thus are not appropriate for mapping RNAseq data, as a sequencing read may be originated from two or more exons that are not in a close range on genomic DNA. In contrast, Novoalign (www.novocraft.com) and GSNAP, which use hash tables of the reference sequences, allow mapping reads to exon-exon junctions. Novoalign features better accuracy comparing to other aligners (Li and Homer, 2010), while GSNAP can accept known splicing sites, SNPs and RNA-editing events and tolerates these variants while mapping. Tophat, on the hand, utilizes Bowtie to support gapped alignment in a two-round mapping procedure. In the first round Tophat tries to map reads to the reference genome and identifies potential exons

by searching the genome for regions with reads aligned and determine the spliced junctions of them. In the second round, it aligns the reads that were not mapped in the first round to the junctions of the exons. The algorithm of Tophat favours the mapping of RNAseq data, since it does not rely on existing annotation of the reference genome. It also has the ability to detect novel splicing events.

Before the advent of SAMtools in 2009 (Li *et al.*, 2009), each mapping software had its own file format to record alignment results, which often required conversion to be recognized for the downstream analysis. SAMtools introduced the SAM format as a standard format for mapping results and this was soon adopted by users. BAM format is the binary format of SAM files but more compressed to save storage space of alignments files. SAMtools also contains a set of tools to manipulate alignment files and extract specific information from them.

Expression analysis

After mapping reads to the reference genome, many analysis tasks can be conducted with existing software. In theory, the number of reads mapped to a gene is proportional to the abundance of the gene. Software that identifies differentially expressed genes uses counts of reads aligned to each gene as a starting point. Such count data is easily to produce with tools such as the bioconductor package ShortRead (Morgan *et al.*, 2009), BEDtools (Quinlan and Hall, 2010) and htseq-count (Anders *et al.*, 2014). As ShortRead relies on the R (R Core Team, 2012) environment, which is slower compared to scripts written in other programming languages (such as Perl, C or Java), BEDtools and htseq-count are faster and easier to use. The later also produces results that can be easily imported into DESeq. EdgeR (Robinson *et al.*, 2010) and DESeq are the two most common tools to determine differentially expressed genes/exons. Both of them used a negative binomial model for gene counts of RNAseq data. They are also used to determine differentially expressed exons when counts data of exons are the input.

Transcripts assembly

Since a good quality RNAseq library could contain 5-25% of reads (read-length) that can be mapped to exon-exon junctions (Engstrom *et al.*, 2013), it is feasible to assemble expressed transcripts with RNAseq data to determine the structure and sequence of the

transcripts, as well as transcript abundance. Several published software tools try to achieve this with different algorithms. Velvet (Zerbino and Birney, 2008) is designed to de novo assemble short reads using de Bruijn graphs into the genome or transcriptome, depending on the origin of reads. The assembled transcriptome can then be used as reference to mapping reads, thus abundances of the transcripts can also be estimated. Trinity (Grabherr *et al.*, 2011) utilizes the same similar algorithm but is focused on assembling the transcriptome and includes the downstream expression estimations. Both software tools do not require existing knowledge of the transcriptome, however, constructing de Bruijn graphs is memory intensive. It may take more than 80GB RAM for large mammalian genomes like that of *homo sapiens* (Illumina, 2009). In contrast, Cufflinks (Trapnell *et al.*, 2010) takes advantage of existing annotation of the reference genome and use this as template to assemble transcripts. This saves running time and reduces memory requirement. CummeBand (Trapnell *et al.*, 2010), developed by the same authors as Cufflinks, allows users to visualize the assembly results as well. There is also a published protocol to use the software for the identification of differentially expressed genes and transcript assembly from RNAseq data (Trapnell *et al.*, 2012), as a response of the popularity of Cufflinks in the field. BitSeq (Glaus *et al.*, 2012) determines differentially expressed transcripts in a unique way, which does not require assembling the sequences of transcripts first. It uses existing knowledge of the transcriptome and uses a Bayesian approach to estimate gene expressions from RNAseq data.

Alternative splicing events

Alternative spliced events (ASEs) can also be identified by comparing multiple RNAseq libraries. Several open source tools are available for this purpose. Cuffdiff (Trapnell *et al.*, 2010) use the assembled transcripts libraries of Cufflinks as input. Transcripts of different libraries are compared first to determine duplicates, and then merged into a total reference library without including any duplicates. Cufflinks is then used to assemble transcripts of each RNAseq library again using the merged library as reference. The abundances of the transcripts in each library are also estimated during the process. Alternative spliced exons can then be identified. DEXSeq (Anders *et al.*,

2012), developed by the same group as DESeq, requires no assembly step, which means much less CPU time is required comparing to de novo assembly algorithms. It uses read count data as input, which is the same as DESeq, when identifying differentially expressed exons. But instead, DEXSeq identifies the exons that have differential usage between RNAseq libraries. The usage of exons is derived by comparing coverage of exons of the same gene. When an ASEs happens, the usage of one or more but not all exons of a gene will be changed, thus DEXSeq can be used to identify ASEs and is relatively fast but relies on accurate and complete annotation of the reference genome. In contrast, Diffsplice does not require transcriptome annotation. It does not assemble the full length transcripts neither, instead it detects splicing events by searching for gapped aligned reads and measure the abundances of such events for determination of ASEs. All of the above software rely on correct mapping and number of reads originated from exons junctions (or spanned over the junction for paired-end reads), thus an accurate aligner and extensive sequencing (eg: 500 times coverage of the transcriptome (ENCODE, 2009)) are required.

Due to the complex composition of RNAseq data from the diversity of genes in terms of their different isoforms, abundances, repeated sequences and sequencing variants they may carry, the accuracy of all of the transcriptome assembly software and software for ASE identification still need to be improved. (Schliesky *et al.*, 2012; Engstrom *et al.*, 2013; Vijay *et al.*, 2013). It is logical to believe that paired-end reads with longer length can provide better coverage on exon junctions, better mapping accuracy and eventually reduce the difficulties in transcriptome assembly. Furthermore, recently developed strand specific RNAseq can also provide information of transcription directions (Parkhomchuk *et al.*, 2009). Together, these all will certainly improve the accuracy of transcript assembly of RNAseq data.

Identification of sequence variants and their applications

With the single base resolution of transcriptome produced by RNAseq data, sequencing variants on the RNA level can be determined, thus allelic-specific expression analysis and RNA-editing event identification can be performed. To my knowledge, in 2011 there was no available software specializing in these two tasks. But for searching

variants on RNA level, several tools that were developed for identification of DNA variants could be adapted, such as Varscan (Koboldt *et al.*, 2012), SAMTools and GATK (McKenna *et al.*, 2010)). Several aligners also have such function implemented, such as Maq and Novoalign. When there is a heterozygous variation detected in a transcript, allelic-specific expression of the transcript can be obtained from the coverage of each base. In humans the most common form of RNA-editing change is adenosine to inosine, which is translated modified to guanosine (Peng *et al.*, 2012). A change that is consistent with the pattern of RNA-editing indicates possible occurrence of the editing event. When comparing RNAseq libraries of two conditions, whether the allelic-specific expression and RNA-editing event is associated with the transcriptome expression difference between the conditions can be tested. However there were no available software tools for this sophisticated analysis during the time of the completion of this PhD thesis, so I wrote in-house Perl scripts for the analysis.

Overall, RNAseq data analysis is still in its infancy and challenging, software tools and algorithms are immature. It is similar to the early years following the emergence of the genome microarray. The situation will change in the future when both the sequencing technology and the analysis methods improve. The accumulation of the knowledge of genomes and transcriptomes will also benefit the analysis.

1.5 Aims of the study

1.5.1 To define the transcriptome of OA and normal cartilage

In this project, we planned to study the differences of transcriptomes of OA and healthy cartilage using both microarrays and RNAseq technology. By comparing the transcriptomes, new evidence of known gene regulations in OA and novel regulation factors would be revealed. From the RNAseq data, more information of the transcriptome, including sequence of transcripts, transcription start sites, expression levels, differential expressed genes, splicing patterns, sequence variants on the RNA level and novel transcripts, would be obtained to understand the cartilage and the disease on molecular level, and ultimately provide possible targets to cure the disease.

1.5.2 To define the workflow to analysis the RNAseq data

As existing open source and commercial software for analysis of the next generation sequencing results is not mature, we would take advantage of different characteristics of popularly used and publication proven software and assign them in different parts of the whole bioinformatics workflow. We will try to connect all of used tools into one workflow. With more investigations on novel software, part/most of the workflow would be adjusted and more functions would be added accordingly. Eventually, we would develop a workflow that requires minimum adjustments to analysis RNAseq data automatically.

1.5.3 To compare the accuracy of RNAseq and microarray in terms of detecting differentially expressed genes

As this project was conducted in the early era of the RNAseq technology, it would be of interest to compare the performance in identification of differentially expressed genes using RNAseq and microarray platforms. This would reveal the advantages and limitations of the RNAseq in gene expression studies.

Chapter 2 Methods and materials

2.1 Reagents and commercially available kits

2.1.1 Reagents

Penicillin-streptomycin solution (10000 U/mL and 10 mg/mL respectively) and Nystatin suspension (10000 U/mL) were obtained from Sigma-Aldrich (Poole, UK). Phosphate buffered saline (PBS) was purchased from Lonza (Wokingham, UK).

2.1.2 Commercially Available Kits

RNeasy® Mini Kit and RNeasy® Midi Kit were purchased from Qiagen (Crawley, UK). E.Z.N.A.™ DNA/RNA Kit was purchased from Omega (Georgia, USA)

2.1.3 Molecular Biology Reagents

Real time qRT-PCR primers and probes were purchased from Sigma- Aldrich (Poole, UK). Moloney Murine Leukaemia Virus (M-MLV) reverse transcriptase and Platinum® SYBR® Green qPCR SuperMix-UDG were purchased from Invitrogen. TaqMan® Gene expression Arrays and TaqMan® Universal PCR Master Mix (2X) were purchased from Applied Biosystems (Foster City, CA, USA). SYBR® Advantage® qPCR Premix (1X) and ROX reference Dye (50x) were purchased from TaKaRa Biomedicals (Wokingham, UK). GeneRuler™ 1 kb DNA ladder was purchased from Fermentas Life Sciences (York, UK).

All other standard laboratory chemicals and reagents, unless otherwise indicated, were commercially available from Sigma-Aldrich, Fisher Scientific, Invitrogen or BDH Chemicals (Poole, UK).

2.2 Methods

2.2.1 Cartilage sample collection

Human articular cartilage samples were obtained from consented patients undergoing joint replacement surgery due to either end-stage hip OA or intracapsular neck of femur fracture (NOF) with Ethical Committee approval from the Newcastle and North Tyneside Health Authority. Joints were inspected macroscopically and scored using a scheme adapted from Noyes classification(Kijowski et al., 2006) to include the presence of osteophytes (Table 2.1) by a blinded experienced orthopaedic surgeon. This adaptation to the Noyes classification (which is commonly used for athroscopic knee cartilage scoring) was necessary because there are currently no accepted classifications for the macroscopic scoring of hip cartilage (Sampson, 2011). Samples scoring ≤1 were considered normal (control) while those scoring ≥5 were classified as osteoarthritic (Table 2.1). After joints were washed extensively with PBS, macroscopically normal full-depth cartilage was collected, snap frozen in liquid nitrogen and then stored at -80°C prior to RNA extraction.

Score	Criteria
0	No erosion, no osteophytes
1	Small erosion, no osteophytes
2	Small erosion, small osteophytes
3	Small erosion, large osteophytes
4	Large erosion, no osteophytes
5	Large erosion, small osteophytes
6	Large erosion, large osteophytes

Table 2.1 OA Scoring Criteria. Cartilage sample were scored from 0-6 according to criteria based on the Noyes classification.

Bovine nasal cartilage was sourced from a local abattoir. Macroscopically normal cartilage was processed removed into PBS containing antibiotics, cleaned thoroughly and cut up into small pieces, then snap frozen in super-cooled n-hexane for RNA/DNA extraction.

2.2.2 RNA extraction from bovine and human cartilage

Sample Preparation: The cartilage was ground with a freezer mill (Retsch, Mixer Mill MM 200, Leeds, UK). Metal vials containing cartilage were cooled with liquid nitrogen

whenever possible during the process. The cartilage was grounded with 5 cycles of 1 min grinding at an impact frequency of 25 Hz and 2 min cooling in liquid nitrogen.

RNeasy® Midi Extraction: RNA was extracted from ground powder of cartilage using RNeasy® Midi kit (Qiagen, Limburg, Netherlands). Cartilage powder was mixed with buffer RLT from the kit. After vortexing with β-mercaptoethanol, the homogenate was centrifuged at 9500 x g for 1 hour at 4°C. Supernatant was transferred on to RNeasy® Midi column. The column was then washed and RNA eluted according to manufacturers' instructions. The flow-through from the first column wash with Buffer RW1 was retained and used for extraction of contaminating DNA. The method extracting the DNA is described in 2.2.3 of this chapter. RNA samples were quantified using a NanoDrop® spectrophotometer (NanoDrop Technologies, Wilmington, Delaware, USA) and stored at -80°C.

E.Z.N.A.™ DNA/RNA Kit Extraction: Different form other nucleic acid extraction kits, E.Z.N.A.™ DNA/RNA Kit provides an approach to obtain pure DNA directly without additional precipitation step. 150-300mg of cartilage were lysed with 700µL of the lysis buffer provided in the kit, 350 µL of the buffer was used for cartilage samples less than 150mg. For bovine cartilage, no more than 200mg were used in order to obtain complete homogenate. Supernatant was loaded on to a DNA column of the kit after the centrifugation of the homogenate. The flow through of DNA column was loaded onto the RNA column provided in the kit. Both columns were then washed and RNA/DNA eluted according to manufacturer's instructions. RNA samples were quantified using the NanoDrop® spectrophotometer and stored at -80°C.

TRIzol® RNeasy® Extraction: RNA was extracted from powdered cartilage using TRIzol® Reagent (Invitrogen) and purified with RNeasy® mini kit. The TRIzol® was immediately added to the samples in a ratio of 5 mL TRIzol® to 300 mg cartilage. This solution was mixed thoroughly using a vortex and incubated at room temperature for 15 min to ensure tissue was fully disrupted. Insoluble material was removed by centrifugation of the homogenate at 20,000 x g for 10 min at 4°C. The supernatant containing RNA was mixed with chloroform in a ratio of 450 µL chloroform per 750 µL TRIzol®. This solution was then vortexed briefly and incubated at room temperature for 10 min prior to centrifugation at 12,000 x g for 15 min at 4°C. The colourless upper

aqueous phase was recovered and mixed with a half volume of 100% ethanol. Samples were applied to the supplied spin columns and centrifuged at 9,500 x g for 15 sec at room temperature. The columns were washed and RNA eluted according to manufacturer's instructions. RNA samples were quantified using the NanoDrop® and stored at -80°C.

2.2.3 DNA extraction from bovine and human cartilage

DNA was obtained from cartilage samples by using EDNA E.Z.N.A.™ DNA/RNA Kit. 150-300mg of cartilage was lysed with 700µL lysis buffer from the kit, 350 µL of the buffer was used for cartilage sample less than 150mg. For bovine cartilage, no more than 200mg was used to obtain complete homogenate. Supernatant was loaded on to DNA column of the kit after the centrifugation of the homogenate. The column was then washed and DNA is eluted according to manufacturer's instructions. DNA samples were quantified using the NanoDrop® spectrophotometer.

2.2.4 Quality assessment of nucleic acids samples

Agrose gel: 0.8% (w/v) Agarose gels were prepared by dissolving the required amount of agarose in x1 TAE buffer through boiling. Ethidium bromide (3,8-Diamino-5-ethyl-6-phenylphenanthridinium bromide) solution was added to cooled agarose at a final concentration of 0.2 µg/mL. Gels were poured, allowed to set and the required amount of RNA/DNA loaded in loading buffer. Bands were separated at 70V for approximately 40 min and visualized on a ChemiGenius II BioImager (Syngene, Cambridge, UK). The software also measures brightness of each band on the gel digitally. As the brightness reflects the concentration of each band, RNA/DNA concentration was calculated by comparing bands of samples to bands of RNA/DNA ladder used on the same gel.

Agilent Bioanalyzer 2100: The quality of RNA samples was further checked on an Agilent Bioanalyzer 2100 platfort. The Agilent 2100 Bioanalyzer is a microfluidics-based platform that separates RNA molecules depending on their sizes and detects quantification of each molecule. Result can be analyzed with 2100 Expert software (Syngene, Cambridge, UK). A very important output value of the software is RNA

integrity number (RIN), which reflects degradation degree of RNA samples (Schroeder *et al.*, 2006).

2.2.5 Quantitative real time PCR (qRT-PCR)

Reverse Transcription Using RNeasy® Mini Kit: Complementary DNA (cDNA) was synthesized from 0.25 µg of total RNA in a volume of 9 µL. RNA was combined with 2 µg of random hexamers (p(dN)6) (GE Healthcare, Little Chalfont, UK) and incubated at 70°C for 10 min. Samples were then transferred to ice and a reaction mixture containing 10 mM DTT, 0.25 mM dNTP, 100 U MMLV, 4 µL 5X First-Strand Buffer (Invitrogen) and 2 µL dH2O added to each sample giving a final reaction volume of 20 µL. The reactions were incubated at 42°C for 1 hour and subsequently diluted 1:50 in dH2O for quantification of target gene expression or 1:250 in dH2O for quantification of house-keeping gene expression. Diluted samples were stored at 4°C prior to analysis and 4 µL aliquots were used in each PCR reaction. Undiluted samples were stored at -20°C.

TaqMan® Probe-Based Real-Time qRT-PCR: Real-Time qRT-PCR reactions were prepared by combining 4 µL cDNA with 4.7 µL TGE 2x buffer and 300 nM of each primer and 150 nM probes in a final volume of 10 µL. Cycling conditions were: 95°C for 10 min and 40 cycles of [95°C for 15 sec, 60°C for 1 min]. For TaqMan® Gene Expression Assays (Applied Biosystems), 5µL cDNA with 4.5 µL TGE 1x buffer and 0.5 µL assay solution in a final volume of 10 µL was used. Genes analyzed include: *COL1A1, COL2A1, C2, CTSK, GPC1, IL6R, MMP11, MMP13, NLRX1, PCSK1, PCSK6, SPRINT1, TMPRSS4* and *XRCC5*.

SYBR® Green Real-Time qRT-PCR: PCR reactions were prepared by combining 4.7µL cDNA with 4.8 µL SYBR® Advantage® qPCR Premix (1X), 0.2µL ROX reference Dye (50x) and 400 nM of each primer in a final volume of 10µL. Cycling conditions were: 95°C for 10 min and 40 cycles of [95°C for 15 sec, 60°C for 1 min], followed by a standard dissociation curve. Genes checked includes: *TLR7* and *SOD2*.

For both TaqMan® probe-based and SYBR® Green qRT-PCR methods the relative quantification of gene expression was performed using the ABI PRISM 7900HT Sequence Detection System (Applied Biosystems). Target gene expression was

normalized to 18S/GAPDH expression levels using the calculation $2^{-\Delta Ct}$. Statistical analysis of differential expression between OA and NOF cartilage samples using real-time qRT-PCR data was performed using the Mann-Whitney U test. P-values ≤ 0.05 were considered statistically significant.

2.2.6 cDNA microarray

Extracted RNA samples were sent to Cambridge Genomic Services (www.cgs.path.cam.ac.uk) for microarray expression profiling. Illumina whole genome expression array HumanHT-12 V3 (Illumina Inc., Illumina United Kingdom, Saffron Walden, UK) was used to profile gene expression of RNA samples according to the manufacturer's protocol.

2.2.7 RNAseq

RNA samples were extracted and then sent to ARK-genomics (www.ark-genomics.org) for mRNA deep sequencing. The quality of RNAs was checked using the Agilent (Agilent Technologies, Inc., California, U.S.) Bioanalyser 2100 and only RNAs with a RNA integrity value (RIN) of greater than 7 was used for the sequencingx. During the library preparation DNA analysis on the Agilent Bioanalyser 2100 was used to check for the size range of the inserts. Kapa library Quant kits (Kapa Biosystems, Wilmington, U.S.) was used to quantify the libraries and the amounts that were loaded onto the flow cells. Illumina (Illumina Inc., California, U.S.) Truseq RNA sample preparation Kit V2 was used and standard library preparation protocol suggested by the manufacture was followed. During the preparation, mRNAs were selected to construct the sequencing libraries. Illumina Genome Analyzer IIx was used to sequence the libraries with one library per lane. Illumina CASAVA (1.7.0) was used for the base calling and quality score calculations. The raw sequencing data in FASTQ format was received via FTP transfer.

2.2.8 Functional and pathway analysis of differentially expressed genes

Gene Set Enrichment Analysis (GSEA) was used to investigate enriched functions of differentially expressed genes (Subramanian et al., 2005a). All expressed genes were ranked according to their fold changes with differentially expressed genes placed into the either end of the ranked list, depending on the change direction. The list was then

used as input of GSEAPreranked tool to enrichments of sets of genes that were classified according to the gene ontologies for molecular function, cellular component and biological process(Ashburner *et al.*, 2000). Results with P-value ≤ 0.01 (Fisher's Exact test) and false discovery rate (FDR) ≤ 0.25 were considered as significant. Ingenuity Pathway Analysis (IPA) (Ingenuity Systems, www.ingenuity.com) was used to identify canonical pathways associated with the differentially expressed genes. All of the differentially expressed genes were included in the analysis. P-values ≤ 0.05 were used to filter results.

2.2.9 Protein interaction network analysis

The interactions of the protein products of up- and down-regulated genes in OA samples were analysed with the use of the search tool STRING (Szklarczyk *et al.*, 2011). Because of the its limitation on number of input genes, differentially expressed genes with only fold change ≥ 2 were taken. I used STRING with three data sources ('Co-occurrence', 'Co-expression' and 'Experiments') to detect and predict interactions between proteins. Confidence threshold was set to 0.4, which is the default of the tool. Genes were then ranked by their number of connections and significances (corrected P value) of differential expression, the top 5% of genes in the list were classified as a hub.

Chapter 3 Nuclear acids extraction

3.1 Introduction

In this project, we planned to analyse information of the both transcriptomes and the methylome, thus both high quality RNA and DNA would need to be extracted from a same cartilage sample. However, extraction of nucleic acids from cartilage tissues is not as easy as from isolated cell lines, because of its low density of cells, highly organized extracellular matrix and also the very limited availability of the samples. Others have successfully isolated cells from cartilage, by enzymatic digestion, prior to nucleic acid purification (Jakob *et al.*, 2003). However, we decided to not do this because of concerns of altering the transcriptome. In order to achieve the RNA quality requirement for both microarray and RNAseq experiment, we tested several extraction procedures, including our existing lab protocol and also other published methods (Ruettger *et al.*, 2010) at the time. Both bovine nasal and human articular cartilage samples were used to test the methods, as human cartilage is very limited. We used the RNA Integrity Number (RIN) of the RNAs extracted with different protocols to evaluate the RNA quality. The RIN is an algorithm developed to evaluate the RNA integrity using electrophoretic RNA measurements from an Agilent 2100 bioanalyzer.(Schroeder *et al.*, 2006) It ranges from 0-10 with higher values indicating better integrity. The recommended RIN for genome-wide expression profiling experiments is usually ≥ 7.

Qiagen (Qiagen, Crowly, UK) RNeasy® Midi Kit was used for both RNA and DNA extraction in the lab. The protocol takes more than 5 hours for one sample, although several samples could be processed in parallel. The technique requires more than 700 mg of cartilage as input. In contrast, Omega (Omega Bio-tek, Norcross, US) E.Z.N.A.™ DNA/RNA Kit and Qiagen RNeasy® Mini Kit require less of cartilage (100 mg or more) as input and less processing time, which can potential reduce the chance of RNA degradation. Using TRIzol® reagent with RNeasy® Mini Kit could provide better

RNA yields and quality, but might not work for human cartilage tissue samples (Ruettger *et al.*, 2010). Life (Life Technologies, Carlsbad, US) TRIzol® reagent can also be used to purify genomic DNA, however it will compromise the genomic DNA and reduces its digestion efficiency by restriction enzymes (Xu *et al.*, 2008), and our priority was to extract sufficient high quality total RNA to construct transcriptome analysis, thus genomic DNA was not purified with the method. As NOF samples had never been collected by ourselves before, to ensure that OA and NOF cartilage samples could be distinguished, the quality of extracted RNAs were further verified using real-time PCR for expressions of previously known to be differentially expressed between OA and NOF cartilages. Sixteen genes were selected, including *COL1A1, COL2A1, SOD2,* (Aigner *et al.*, 2006a) *MMP13* (Bau *et al.*, 2002), *MMP11* (Aigner *et al.*, 2001), *XRCC5, TLR7* (Zhang *et al.*, 2008), *PCSK1, PCSK6,* (Malfait *et al.*, 2008) *C2, SPINT1 (Milner et al., 2010)*, *IL6R* (Kotake *et al.*, 1996), *NLRX1* (Radwan *et al.*, 2013), *CTSK* (Morko *et al.*, 2004), *TMPRSS4* and *GPC1* (Zhang *et al.*, 2003).

3.2 Results

3.2.1 Comparison of yields and quality of RNA extraction procedures

RNA was extracted from bovine nasal and human cartilage tissues by using three different procedures: RNeasy® Midi Kit, E.Z.N.A.™ DNA/RNA Kit and TRIzol® with RNeasy® Mini Kit. The concentration of the RNA products was measured using Nanodrop. The RNA integrity numbers were obtained using Bioanalyzer. TRIzol® with RNeasy® Mini Kit gave best yield among all of the approaches (Table 3.1), which is around 5.7 μg RNA per gram of bovine nasal cartilage on average and 21.4 μg RNA from a gram of human articular cartilage.

	RNeasy Midi Kit (ug RNA/g cartilage)	E.Z.N.A. Kit (ug RNA/g cartilage)	TRIzol RNeasy Kit (ug RNA/g cartilage)
Bovine nasal cartilage	3.34±0.96 (n=10)	0.66±0.14 (n=13)	5.72±1.92 (n=8)
Human articular (knee/hip) cartilage	2.84 (n=1)	5.72±1.92 (n=8)	21.38±2.55 (n=2)
Human fat pad	NA	48.48 (n=2)	NA

Table 3.1: Comparison of Total RNA yields of different tissue samples using different extraction procedures. "n" indicates the total number of RNA samples extracted with each procedure. The standard error of mean is indicated as "±". Trizol/Mini protocol produced the best RNA yields from cartilage. The RNA from human fat pad was used to ensure the protocol of using E.Z.N.A kit was working properly. However, the kit could not extract sufficient amount of RNA from the cartilage samples.

On average only 0.66 ug of RNA was eluted from the RNA column of E.Z.N.A. ™ DNA/RNA Kit. To ensure that the small yield of the RNA was not due to user error with the E.Z.N.A. ™ protocol being followed, human fat samples were used for the extraction. As anticipated, the extraction of the RNA was successful and the yield of RNA was more than from the cartilage samples.

RNeasy® Midi Kit produced similar amounts of RNA as the combination of TRIzol® and RNeasy® Mini Kit when using the bovine cartilage samples, but less when using human cartilage tissue. Comparing the RIN numbers of RNAs extracted from bovine cartilage using the two procedures, using TRIzol® with RNeasy® Mini Kit was better (Figure 3.1 and Table 3.2). The procedure also performed equally well when extracting RNAs from the human cartilage sample (Table 3.3).

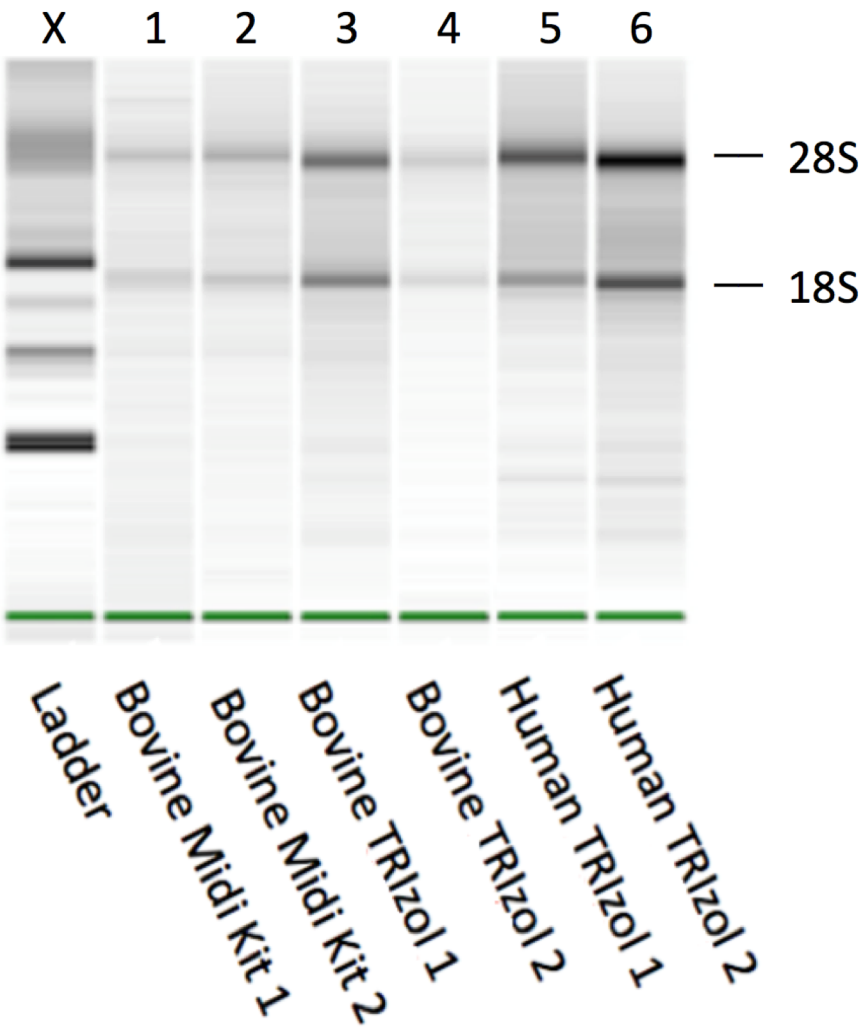


Figure 3.1: Comparison of the quality of RNAs extracted from the bovine and human cartilage using different procedures: The image is the pseudo-gel results converted from the Bioanalyer data of the RNA samples extracted using different procedures. The two bands of the rRNAs are clearer for samples extracted with TRIzol RNeasy Mini Kit than RNeasy midi kit.

	RNeasy Midi Kit	TRIzol RNeasy Kit
Bovine cartilage	4.85 (n=2)	7.3 (n=2)
Human cartilage	NA	7 (n=2)

Table 3.2: Comparison of RIN of RNA samples extracted from bovine cartilage by using different extraction procedures. "n" indicates the total number of RNA samples extracted with each protocol. RIN was obtained using Bioanalyzer. Trizol/Mini protocol produced RNAs in better quality.

	NOF1_1	NOF1_2	NOF1_3	NOF1_4	OA1_1	OA1_2	OA2_1	OA2_2	OA3_1	OA3_2	OA3_3	Mean
RIN	8.6	8.2	8.4	8.9	8.1	7.4	9.1	8.9	8.6	8.7	9	8.54 ±0.15
Yields (ug/g)	34.00	90.13	27.87	39.47	56.67	45.47	53.20	33.87	23.33	25.47	4.13	39.42 ±6.73

Table 3.3 RIN and yields of RNA extracted from human femoral heads cartilage using TRIzol RNeasy Mini Kit. The minimum RNA is 7.4. The average yield is 39.42 ug RNA from a gram of cartilage tissue.

The performance of the protocol of TRIzol® with RNeasy® Mini Kit on human articular cartilage was then verified with additional 3 cartilage samples collected from human femoral heads, including 1 NOF sample and 3 OA cartilage samples. The quality of the RNA samples was again checked using a Bioanalyzer (Figure 3.2). The average yield was 39.42ug RNA per gramme of cartilage tissue. The minimum RIN was 7.4 and the mean was 8.54. As the performance of the procedure was consistent and it was the only procedure produced RNAs with quality that met our requirements for microarray and RNAseq experiments, the procedure was chosen to extract total RNAs for following experiments.

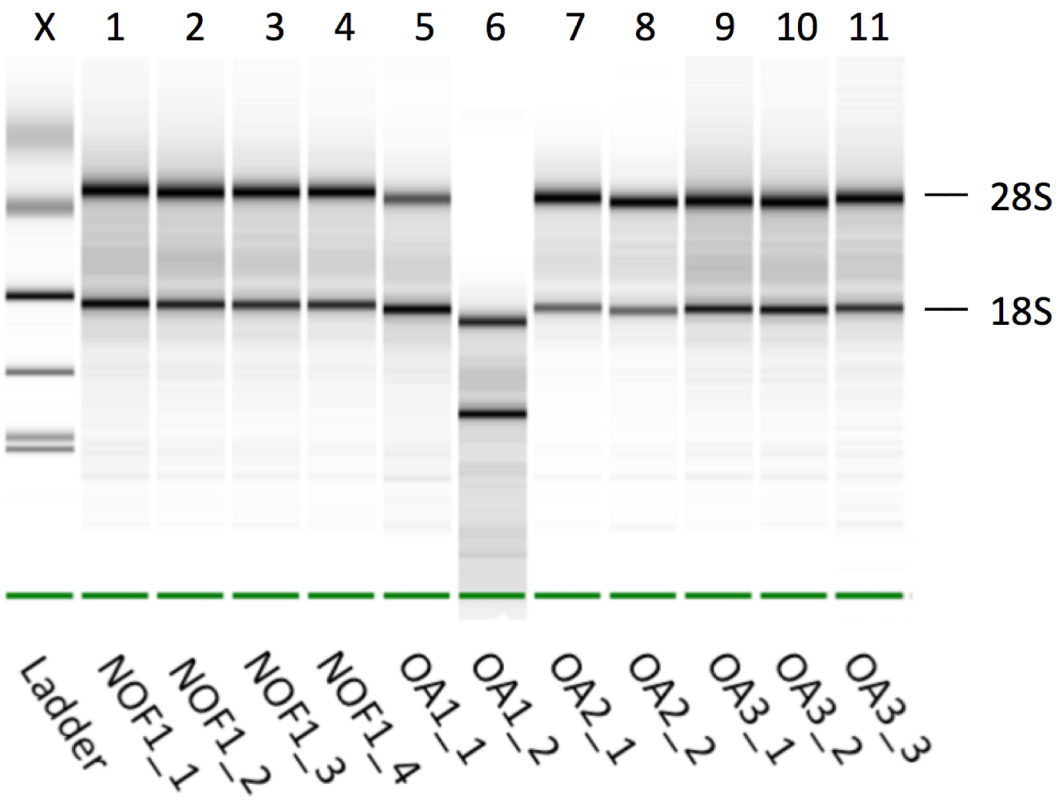
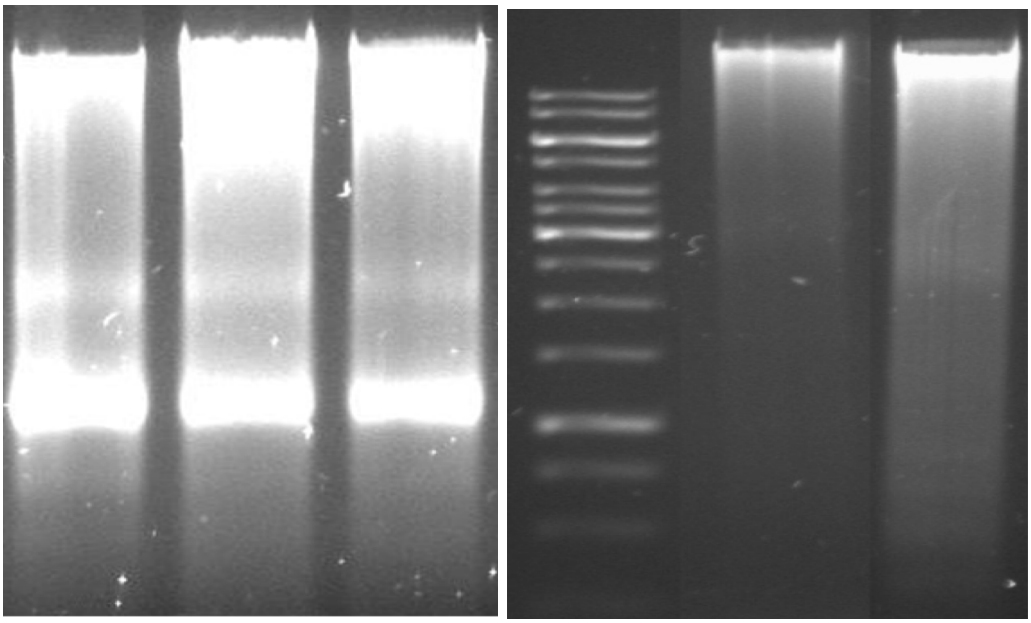


Figure 3.2: Quality of RNA extracted from both human OA and NOF cartilage samples using TRIzol with RNeasy Mini Kit. The image is the pseudo-gel converted from the Bioanalyer data of the RNA samples extracted using TRIzol with RNeasy Mini Kit. It shows the clear bands of the rRNAs and the consistency of the RNA quality across samples. Lane 6 has a late migration probably due to dirt on the electrode cartridge used for the experiment.

3.2.2 Comparison of DNA extraction yields and quality

The priority of the study was to obtain the transcriptome of cartilage. RNA extracted with RNeasy® Midi Kit did not meet the minimum quality requirement thus DNA was not further purified using the kit. Unlike RNeasy® Midi Kit, DNA extraction with E.Z.N.A. ™ DNA/RNA Kit does not require an extra purification step, thus was obtained along with RNA products. DNA was not extracted with the RNeasy® Mini Kit due to the concerns of TRIzol mentioned above. The DNA yield from bovine cartilage by using E.Z.N.A™ DNA/RNA Kit is 11.52 (n=13, ±1.22) μg per gram of cartilage and 3.86 (n=8, ±0.20) μg for human cartilage. Some of the bovine DNA samples showed obvious RNA contamination upon agarose gel electrophoresis as rRNA bands were observed on the gel, but this was not observed for human cartilage DNA samples (see Figure 3.3).



A B

Figure 3.3: Agarose gel of DNA samples extracted from bovine (A) and human (B) using E.Z.N.A. kit. Extra bands can be seen on the gel of the bovine DNA samples. No other bands were observed on the gel of the human cartilage DNA samples.

3.2.3 Expression profiles of the extracted RNAs using real-time PCR

In order to further confirm the RNA quality extracted using the TRIzol® with RNeasy® Mini Kit, in total 46 RNA samples were extracted from 17 OA cartilage samples and 19 NOF cartilage samples using the protocol. The expression of 16 genes that were know to be differentially expressed between OA and NOF cartilage were then determined using qRT-PCR, including *C2, CTSK, GPC1, IL6R, MMP11, MMP13, NLRX1, PCSK1, PCSK6, SOD2, SPINT1, TLR7, TMPRSS4, XRCC5, COL1A1* and *COL2A1* (Table 3.4 and Figure 3.4). Equally amounts of RNA were reversely transcribed. Two housekeeping genes, *18S* and *GAPDH*, were also measured and their expressions (Ct) were determined to be 17.80 ± 0.31 and 23.65 ± 0.08 respectively. When comparing the expression of *GAPDH* and *18S* between the OA and NOF samples, *GAPDH* showed the least variability in expression and was thus chosen for the normalization. *GAPDH* was used to normalize Ct values of the other genes. Four RNA samples had more than 5 undermined cycle values for all genes ($Ct > 40$), thus considered as low quality samples and removed from further analysis and experiments. Except *MMP13* all of the other genes were detected and found significantly differentially expressed (P-value ≤ 0.05) in

OA samples comparing to the NOF (Table 3.4 and Figure 3.4). This confirmed both the quality of the RNA samples and also the expression profiles of the genes in OA cartilage.

Gene	Log2Fold Change (OA/NOF)	P-value	Expected Change
C2	2.66	2.78E-06	up
CTSK	2.49	5.21E-09	up
GPC1	2.39	5.98E-08	up
IL6R	-1.13	1.16E-04	down
MMP11	3.58	4.66E-05	up
MMP13	0.87	5.39E-01	up
NLRX1	1.73	9.35E-07	up
PCSK1	-2.3	3.67E-05	down
PCSK6	1.31	1.23E-02	up
SOD2	-3.89	5.00E-12	down
SPINT1	1.33	2.01E-04	up
TLR7	-0.61	4.03E-02	down
TMPRSS4	-3.42	3.42E-05	down
XRCC5	0.75	3.34E-02	up
COL1A1	3.03	1.20E-02	up
COL2A1	4.07	2.31E-08	up

Table 3.4: Gene expression differences between OA and NOF. The table shows the fold change in log 2 scale and the p-values (Mann-Whitney test) of the changes. The known change of the gene and the related publication are also included in the table. With the exception of MMP13, the remaining genes were significantly differentially expressed in OA samples, which were as expected and confirmed quality of RNAs extracted.

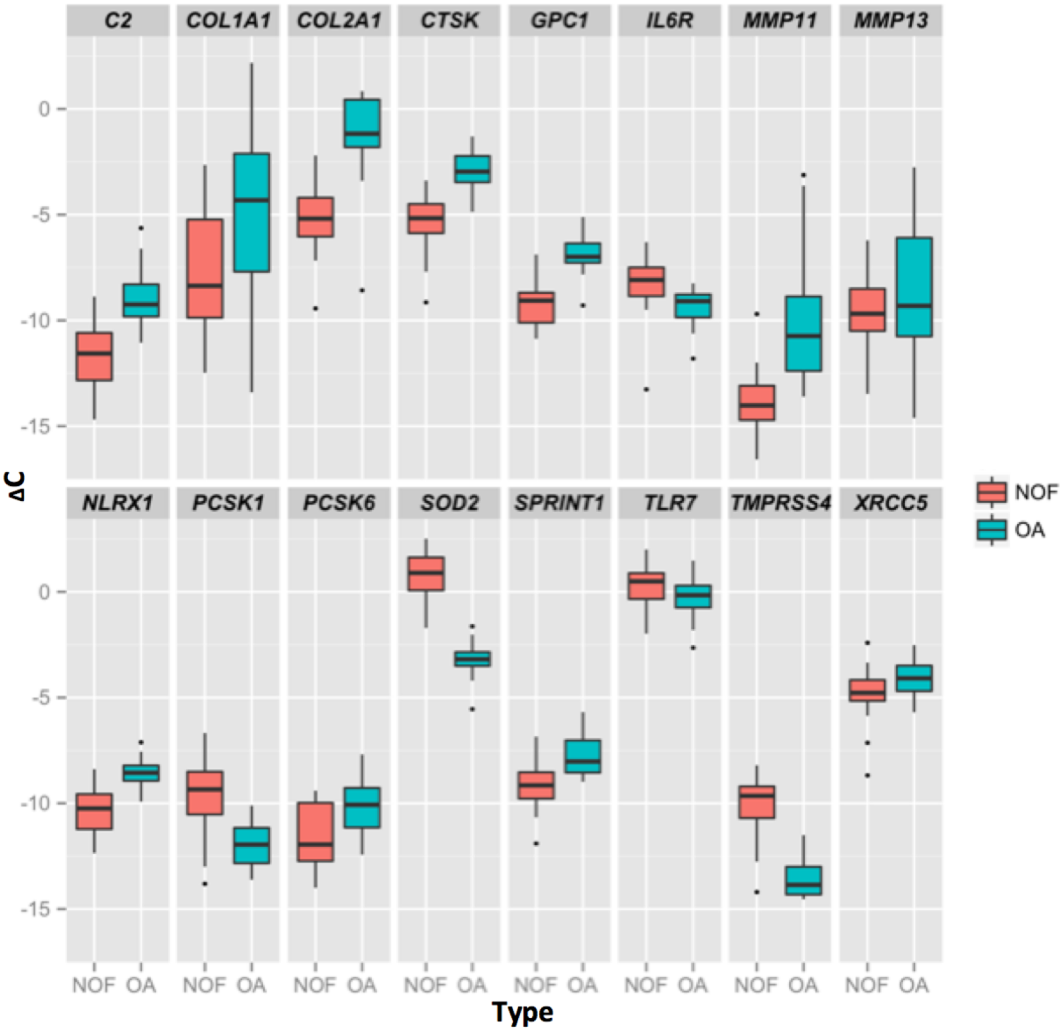


Figure 3.4: ΔCt of the genes determined using real-time PCR. The boxplot shows the ΔC values of the genes expressed in NOF samples (coloured in red) and OA samples (coloured in blue).

3.3 Discussion

RNA extraction from cartilage tissue samples is difficult as shown in this study. RIN of such RNA samples is rarely above 9.0, a value commonly observed for RNA extraction from cell lines. This could be because of the room temperature procedures, which increase the risk of RNA degradation. Importantly, the delay between surgical removal of the joint and collection of cartilage could have an impact on the integrity of RNA. Furthermore, how the consented tissues were treated within operating theatres was not

possible to control. These confounding factors have not been taken into account or tested when deciding that the TRIzol® with RNeasy® Mini Kit was the most effective. Contents in ECM also play a role in the decrease of RNA yield, mainly because of a blockage of columns and probably because of the actual composition of the cartilage macromolecules. A recent report showed that newly developed membrane technologies might provide a solution. (Ruettger *et al.*, 2010)

When using the TRIzol® with RNeasy® Mini Kit protocol some of the RNA products were not colourless but lightly pink. Though it was found latterly that by decreasing the ratio of TRIzol® to chloroform when lysing the cartilage the pink carryover could be avoided, in order to retain consistency between sample preparations the protocol was kept unchanged. As TRIzol® is the only pink reagent in the process, it could be the source of the suspicious colour. As TRIzol® was demonstrated to reduce digestion efficiency of genomic DNA by restriction enzymes (Xu et al., 2008), thus DNA was not purified after using TRIzol to extract RNA from cartilage samples.

The NOF RNA and cartilage samples used in the lab before were provided from collaborators at University of East Anglia and this was the first time that we collect NOF cartilage ourselves, thus we performed qRT-PCR experiment to ensure that the OA and NOF samples could be distinguished. Except *MMP13,* all of the rest genes showed differential expressions in OA samples as anticipated, indicating the reliability of the NOF cartilage.

Overall, in comparison of the other available procedure to extract RNA from cartilage samples, TRIzol® with RNeasy® Mini Kit produced high quality RNA with sufficient yields for our study to proceed. Although using TRIzol® raised concerns of genomic DNA quality, the procedure was chosen for this study to ensure the reliability of the transcriptome analysis results.

Chapter 4 Genome-wide cDNA Microarray ensured RNA quality and revealed commonality as well as discord between hip and knee OA

4.1 Introduction

cDNA microarray technology has been used for expression profiling for several years since its first introduction (Schena et al., 1995). The key advantage of the technology is it is high-throughput compared to previous gene expression detection techniques, such as standard RT-PCR. The expression of genes of the whole genome can be measured at once with the technology, which has enabled genome-wide expression comparisons. The availability of commercial cDNA profiling chips and standardization of the data analysis method also ensured the reproducibility of profiling results, although data from different laboratories and different platforms may present slightly differing results ('The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements,' 2006). Benefitting from its high throughput nature, a single microarray chip can have more than one probe targeting different regions of the same gene or different isoforms of that gene, thus the expression of the gene detected can be more reliable and reproducible.

The high-density of cDNAs printed on microarray chips provides high throughput ability but also created obstacles to its interpretation, such as cross-hybridization and down-stream data analysis. Because of the short length of the DNA oligonucleotides as probes (~25mer), they have limited specificity thus a transcript molecule could cross-hybridize on to a probe of another mRNA (Li and Wong, 2001). Both on the physical probe design level and the data analysis level (Chu et al., 2002; Wu et al., 2005), the problem has been largely addressed. The major concerns of the down-stream data analysis include 1) the large amount of expression data of the whole genome together with the technical background noise composed by the differences in the efficiency of labelling reactions and 2) production differences between microarrays (Aris et al.,

2004). The main aim of the data analysis is to remove or minimize the background noise and preserve the true biological difference between samples. Over years, several data normalization techniques have been developed, reported and made available in several software packages in order to achieve this (Quackenbush, 2002; Leung and Cavalieri, 2003). Comprehensive understanding of the technology and its characters was also established in large scale microarray-reliability focused studies, such as the MicroArray Quality Control (MAQC) project (Canales et al., 2006), their analysis pipeline became standardized and distributed both freely and commercially. Some of these are freely distributed with Bioconductor, such as lumi (Du et al., 2008) and affy (Gautier et al., 2004). Commercial packages, such as Genespring (Agilent Inc), also provide thorough and intuitive interface for the analysis, plus limited pathway analysis function. In our microarray data analysis, Genespring was used because of its ease of use.

Microarray experiment often results a list of genes, to get insight of the underlying biology requires functional and pathway analysis. For this purpose, a number of different software and databases of gene functions have been developed, covering simple over representation approach to more advanced pathway topology based approaches (Khatri et al., 2012). All of these rely on the existing knowledge of genes and their protein products. Several commonly used tools include the Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005a), the Database for Annotation Visualization and Integrated Discovery (DAVID) (Huang et al., 2008; Huang et al., 2009) and the Ingenuity Pathway Analysis (IPA) (Ingenuity Systems, www.ingenuity.com). GSEA and DAVID are freely available to the public while IPA is commercialized featuring an intensive database of functions and interactions of genes and proteins manually curated from published papers.

The first cartilage comparative study to use commercially available microarrays compared gene expression changes within 5 OA and 5 normal human knee patient cartilage biopsy samples (Karlsson *et al.*, 2010b). However, in the study the authors did not correct the p-values for multiple tests, so the differentially expression genes identified in the study may include a number of false positives. The sample size of the study was limited and included both male and female individuals.

Here, we used Illumina Human HT-12 V3 whole genome expression arrays to profile the transcriptome of cartilage collected from femoral heads OA and neck of femur fracture patients, importantly all of which were female. We also compared our findings to the similar study of knee OA.

4.2 Aims

The aim of the study was to investigate the gene expression of chondrocytes in hip OA compared to the relatively normal hip cartilage to provide a comprehensive understanding of disease pathology via identification of the pathways involved. To our knowledge, this was the first study to analyse gene expression changes in human hip OA at the whole-genome level. Furthermore, our project was to analyze the transcriptomes using RNAseq, thus the microarray experiment would not only provide many insights of the disease mechanisms but also will validate the subsequent RNAseq analysis.

4.3 Methods

4.3.1 Identify differentially expressed genes

Illumina whole genome expression array HumanHT-12 V3 (Illumina Inc., Illumina United Kingdom, Saffron Walden, UK) was used to profile gene expression of RNA samples according to the manufacturer's protocol. Raw expression data were analysed using Agilent GeneSpring GX 11 (Agilent Technologies, Santa Clara, California). Quality control of the raw data was performed and outlier samples were removed following the method described in (Oldham et al., 2008). R package Combat was used to adjust the data to remove any possible batch effect incurred during RNA preparation after removal of outliner samples. (Johnson et al., 2007; R Core Team, 2012). The data was then normalized with a quantile algorithm (Bolstad et al., 2003) and the baseline was transformed to the median of all samples within GeneSpring GX. Those probes with a flag value of 'Present' or 'Marginal' in ≥ 80% of either OA or NOF samples were selected for differential expression analysis.

4.3.2 Statistical analysis

From the microarray analysis the significance of differentially expressed genes was evaluated with a Mann-Whitney's test, which is non-parametric, and then corrected for multiple testing using the Benjamini-Hochberg (Benjamini and Hochberg, 1995) method. Differentially expressed genes with a fold change >= 1.5 and P-value 0.01 were included in further analyses. GSEA uses a permutation test procedure to evaluate the significance of gene ontology (GO) term enrichments. Because these P-values are not multiple testing corrected, a FDR (calculated as described (Subramanian et al., 2005b)) of 0.25 was used as threshold to control the number of false positives. The Fisher's exact test, implemented in IPA and other pathway analysis applications (Werner, 2008), was used to assess the significance of the association between a pathway and the differentially expressed genes. The Pearson product-moment correlation coefficient was used to test the correlation of fold changes of overlap genes between the Karlsson (Karlsson *et al.*, 2010b) and our study.

4.4 Results

4.4.1 Cartilage sample collection

In total 29 cartilage samples were collected from femoral heads of female OA donors (13 samples; median age = 71 yrs, shown in Table 4.1), and female NOF donors (16 samples; median age = 78 yrs). All donors were UK citizens of North European descent. The OA cartilage samples had obvious OA signs including degraded and fibrous cartilage with osteophytes and exposed subchondral bone. Compared to OA joints, the NOF femoral heads had fully intact cartilage with little fibrillation and no exposed bone (Fig. 4.1 A and B respectively).

NOF Sample name	Age	OA Sample name	Age
T030-3	73	DTOS1734-3	83
T021-3	84	DTOS1626-3	83
T032-3	82	DTOS1596-3	76
T014-3	84	DTOS1842-3	82
T012-3	81	DTOS1590-3	66
T018-1	94	DTOS1683-3	78
T003-3	85	DTOS1786-3	60
T007-3	71	DTOS1883-2	72
T013-3	72	DTOS1817-3	55
T023-3	68	DTOS1906-3	72
T011-3	69	DTOS1595-3	51
T024-3	80	DTOS1772-3	76
T023New-3	68	DTOS1567-3	64
T029-3-2	52		
T035-3	92		
T037-3	83		
T031-2	89		
NOF mean age	78.1±2.6	**OA mean age**	70.6±3.0

Table 4.1 Age of OA and NOF patients.

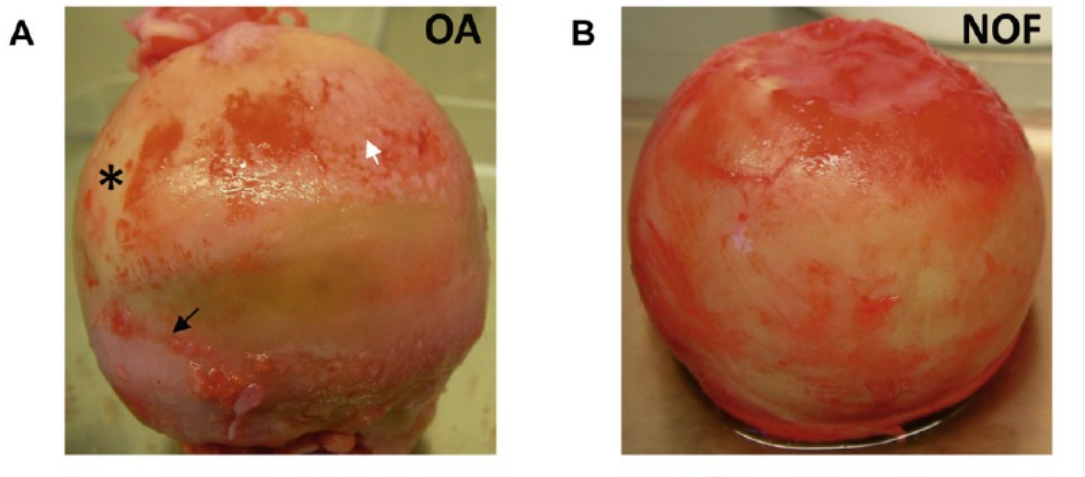


Figure 4.1 OA and neck of femur fracture (NOF) femoral heads and clustering. **A**, A typical OA femoral head with exposed bone (black arrow) and fibro-cartilage (white arrow). Macroscopically normal cartilage can be observed (*) which was collected and processed as described. **B**, A typical NOF femoral head with a covering of smooth healthy cartilage.

4.4.2 Quality controlled microarray data

As obvious outliers can be observed in the hierarchical clustering result of the microarray data (Figure 4.2 A), quality control on the samples before normalization of the microarray data was performed following the method described in (Oldham et al., 2008). Nine outliers were eliminated from the further analysis. In addition, two samples with macroscopic scores (2 and 3) in the middle range, which reflect unclear tissue types, were removed. Overall, microarray data of 19 samples were left having a mean modified Noyes score of 5.2 and NOF 0.6 (shown in Figure 4.3). Following microarray analysis, hierarchical clustering of cartilage samples based on expression of all genes passing the quality filter showed perfect segregation of the OA and NOF samples (Figure 4.2 B).

4.4.3 Differentially expressed genes and functional analysis

In total 1151 differentially expressed genes were identified (fold change ≤ -1.5 or ≥ +1.5, P-value ≤0.01) between the two sample groups (shown in Additional Table A4.1 available online: https://github.com/byb121/Thesis_2015/tree/master/Thesis_2015/Additional%20tables). These included 562 up-regulated and 589 down-regulated genes. Amongst these, 381 genes showed a ≥ 2 fold change of expression level. A number of these genes have previously been shown to be differentially expressed between hip OA and NOF cartilage including *ADAMTS1, ADAMTS5, ADAMTS9, MMP1, MMP3, MMP23* and *SOD2* (Kevorkian et al., 2004b; Davidson et al., 2006; Swingler et al., 2009a; Scott et al., 2010b) consistent with the data herein. The most robust down-regulation was observed for chemokine ligand 20 (*CCL20*) which showed more than 22 fold less expression in the OA group. Of the up-regulated genes, over 50% of the top 25 genes are expressed in the ECM including a number of collagen genes; *COL2A1, COL3A1, COL5A2, COL9A1* and *COL11A1*, all of which have previously been reported to be up-regulated in OA cartilage (Aigner *et al.*, 2006b).

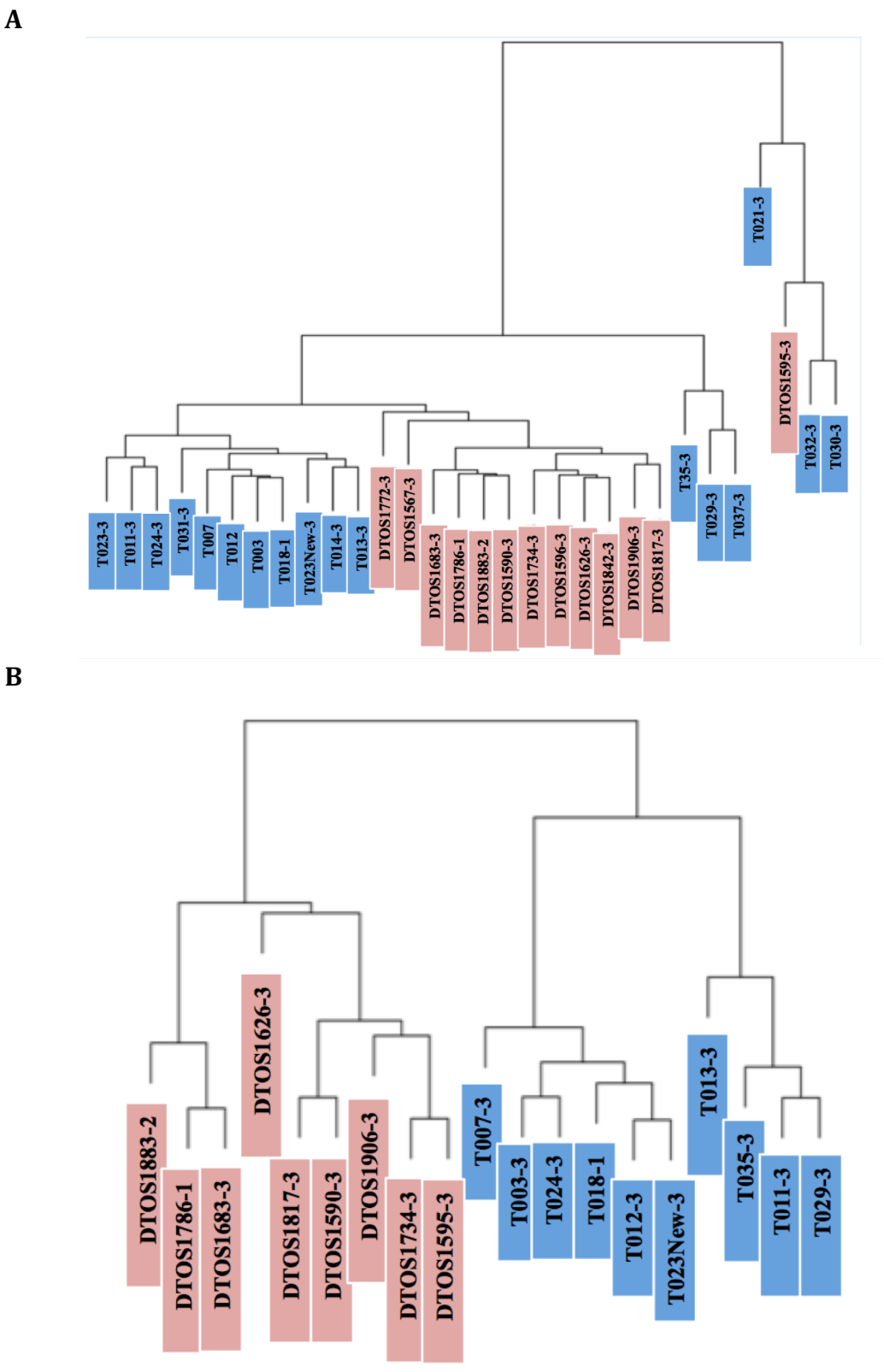


Figure 4.2 Hierarchical clustering of all samples based on the expression profiles before and after samples filtering. OA samples are in red background. NOF samples are in blue background. **A.** Before filtering, outliers can be seen on the left side of the figure. OA and NOF samples are not perfectly segregated. **B.** After filtering OA and NOF sample are perfectly segregated into two groups, which is consistent with both diagnosis and the blinded cartilage OA scores.

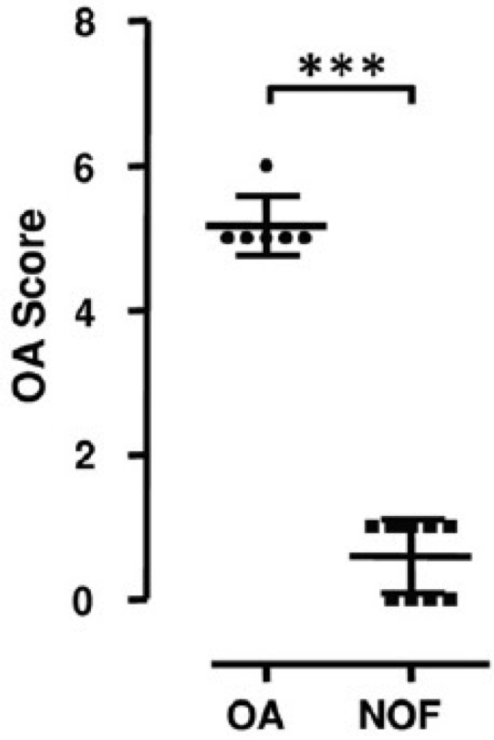


Figure 4.3 Blinded OA cartilage phenotype scores based on the Noyes classification of filtered samples. Closed bar, average NOF; open bar, average OA. Error bars represent standard deviation, *** represents $P<0.001$.

The predominant functions of up and down regulated genes within the three Gene Ontology (GO) categories (molecular function, cellular component and biological process) were assessed (Table 4.2; gene names of each enriched term are listed in Additional Table A4.2). In terms of 'molecular function', the results indicate increased hydrolases activity and decreased protein kinase activity in OA cartilage. With regards to 'cellular function', ECM was enriched in up-regulated gene lists, while nuclear regions are enriched in down-regulated genes. Although both up- and down-regulated genes lists contain approximately equal numbers, those with reduced expression were classified into a greater diversity of biological processes, such as stress responses, cell death and cellular processes.

Molecular Function - Up Regulated Genes:

GO term	DEG_Size	SIZE	FDR
HYDROLASE ACTIVITY HYDROLYZING O GLYCOSYL COMPOUNDS	6	29	0
HYDROLASE ACTIVITY ACTING ON GLYCOSYL BONDS	7	39	5.56E-04
CALCIUM ION BINDING	8	71	0.020
TRANSFERASE ACTIVITY TRANSFERRING HEXOSYL GROUPS	4	68	0.037
ION BINDING	12	208	0.053
TRANSFERASE ACTIVITY TRANSFERRING GLYCOSYL GROUPS	5	90	0.059
CATION BINDING	12	164	0.066
GLUTATHIONE TRANSFERASE ACTIVITY	1	15	0.083
ATPASE ACTIVITY COUPLED TO TRANSMEMBRANE MOVEMENT OF IONS PHOSPHORYLATIVE MECHANISM	3	15	0.119
DAMAGED DNA BINDING	0	18	0.121
OXIDOREDUCTASE ACTIVITY	11	223	0.126
PYROPHOSPHATASE ACTIVITY	11	198	0.166

Molecular Function - Down Regulated Genes:

GO term	DEG_Size	SIZE	FDR
PROTEIN KINASE ACTIVITY	19	243	0.160
TRANSFERASE ACTIVITY TRANSFERRING PHOSPHORUS CONTAINING GROUPS	22	358	0.175
PROTEIN SERINE THREONINE KINASE ACTIVITY	15	175	0.191
RNA BINDING	10	216	0.211
MRNA BINDING	5	20	0.224

Cellular Component - Up Regulated Genes:

GO term	DEG_Size	SIZE	FDR
EXTRACELLULAR MATRIX	14	78	0
PROTEINACEOUS EXTRACELLULAR MATRIX	14	77	0
ENDOPLASMIC RETICULUM	10	253	0
GOLGI APPARATUS	9	192	0
EXTRACELLULAR MATRIX PART	11	47	0
COLLAGEN	8	19	2.77E-04
EXTRACELLULAR REGION PART	19	226	4.00E-04
ENDOPLASMIC RETICULUM PART	3	83	4.57E-04
ORGANELLE MEMBRANE	4	266	5.29E-04
EXTRACELLULAR REGION	23	291	6.21E-04
NUCLEAR ENVELOPE ENDOPLASMIC RETICULUM NETWORK	2	79	0.002
MICROSOME	4	26	0.002
ENDOPLASMIC RETICULUM MEMBRANE	2	73	0.003
GOLGI APPARATUS PART	3	88	0.005
ENDOMEMBRANE SYSTEM	3	199	0.010
ER GOLGI INTERMEDIATE COMPARTMENT	1	20	0.011
VESICULAR FRACTION	4	28	0.012
MITOCHONDRIAL MEMBRANE PART	0	43	0.013
CYTOPLASMIC VESICLE	7	99	0.014
VESICLE	7	104	0.025
INTRINSIC TO ORGANELLE MEMBRANE	0	48	0.026
INTEGRAL TO ENDOPLASMIC RETICULUM	0	22	0.027

MEMBRANE			
CYTOPLASMIC MEMBRANE BOUND VESICLE	6	95	0.030
MEMBRANE BOUND VESICLE	6	97	0.033
INTRINSIC TO ENDOPLASMIC RETICULUM MEMBRANE	0	22	0.036

Cellular Component - Down Regulated Genes:

GO term	DEG_Size	SIZE	FDR
PORE COMPLEX	1	34	0.055
TIGHT JUNCTION	2	24	0.068
APICOLATERAL PLASMA MEMBRANE	3	26	0.068
NUCLEAR MEMBRANE PART	1	39	0.077
NUCLEAR PORE	1	30	0.153

Biological Process - Up Regulated Genes:

GO term	DEG_Size	SIZE	FDR
CELLULAR CARBOHYDRATE METABOLIC PROCESS	11	99	0.002
SKELETAL DEVELOPMENT	14	80	0.003
ORGAN DEVELOPMENT	31	418	0.004
ORGANELLE ORGANIZATION AND BIOGENESIS	16	387	0.009
NERVOUS SYSTEM DEVELOPMENT	15	295	0.014
PROTEIN FOLDING	1	52	0.016
VESICLE MEDIATED TRANSPORT	4	171	0.018
CARBOHYDRATE BIOSYNTHETIC PROCESS	5	37	0.021
GLYCOPROTEIN METABOLIC PROCESS	0	71	0.100
GENERATION OF PRECURSOR METABOLITES AND ENERGY	5	101	0.067
MEMBRANE ORGANIZATION AND BIOGENESIS	2	112	0.007
CARBOHYDRATE METABOLIC PROCESS	13	137	0.002
CHROMATIN MODIFICATION	0	47	0.103
PHOSPHOINOSITIDE METABOLIC PROCESS	0	26	0.096
PHOSPHOINOSITIDE BIOSYNTHETIC PROCESS	0	23	0.069
ESTABLISHMENT AND OR MAINTENANCE OF CHROMATIN ARCHITECTURE	0	65	0.073
GLYCEROPHOSPHOLIPID BIOSYNTHETIC PROCESS	2	27	0.059
DNA REPAIR	1	109	0.135
EXTRACELLULAR STRUCTURE ORGANIZATION AND BIOGENESIS	1	23	0.107
LIPOPROTEIN BIOSYNTHETIC PROCESS	0	23	0.123

Biological Process - Down Regulated Genes:

GO term	DEG_Size	Size	FDR
CELL PROLIFERATION GO 0008283	29	392	0
RESPONSE TO EXTERNAL STIMULUS	20	207	4.88E-04
RESPONSE TO WOUNDING	13	128	0.008
NUCLEOCYTOPLASMIC TRANSPORT	4	73	0.008
NUCLEAR TRANSPORT	4	73	0.008
REGULATION OF BIOLOGICAL QUALITY	20	291	0.008
TRANSLATION	6	153	0.009
NEGATIVE REGULATION OF PROGRAMMED CELL DEATH	13	129	0.009
NEGATIVE REGULATION OF DEVELOPMENTAL PROCESS	13	161	0.009
INFLAMMATORY RESPONSE	8	89	0.010
NEGATIVE REGULATION OF APOPTOSIS	13	128	0.011

REGULATION OF PROTEIN KINASE ACTIVITY	9	133	0.011
REGULATION OF TRANSFERASE ACTIVITY	9	137	0.012
REGULATION OF CELL PROLIFERATION	18	229	0.012
RESPONSE TO STRESS	37	409	0.012
PROGRAMMED CELL DEATH	29	356	0.012
APOPTOSIS GO	29	355	0.012
REGULATION OF CYCLIN DEPENDENT PROTEIN KINASE ACTIVITY	3	39	0.013
REGULATION OF DEVELOPMENTAL PROCESS	21	353	0.014
REGULATION OF KINASE ACTIVITY	9	135	0.014
REGULATION OF APOPTOSIS	21	279	0.014
REGULATION OF PROGRAMMED CELL DEATH	21	280	0.014
NEGATIVE REGULATION OF CELL PROLIFERATION	10	120	0.015
RNA EXPORT FROM NUCLEUS	0	16	0.018
REGULATION OF NUCLEOCYTOPLASMIC TRANSPORT	2	17	0.028
ANTI APOPTOSIS	7	103	0.028
CALCIUM INDEPENDENT CELL CELL ADHESION	2	16	0.031
DEFENSE RESPONSE	12	170	0.036
NEGATIVE REGULATION OF MAP KINASE ACTIVITY	2	16	0.045
REGULATION OF CATALYTIC ACTIVITY	10	212	0.045
REGULATION OF CELL CYCLE	11	158	0.050
NUCLEAR EXPORT	2	29	0.056
TRANSCRIPTION FROM RNA POLYMERASE II PROMOTER	29	380	0.060
IMMUNE RESPONSE	9	148	0.063
TRNA METABOLIC PROCESS	2	18	0.063
CELL DEVELOPMENT	33	464	0.086
REPRODUCTIVE PROCESS	10	104	0.094
NEGATIVE REGULATION OF TRANSPORT	1	17	0.099

Table 4.2 Functions Enrichment Analysis Result. Enriched functions of up and down regulated genes are listed in the table and separated into 3 GO term categories. DEG_Size, number of differentially expressed genes that contribute to the enrichment of the term. Size, number of expressed genes associated with the term. FDR, False discovery rate.

4.4.4 Molecular pathways and protein interaction networks

The entire list of differentially expressed genes was found to be significantly ($P \leq 0.05$) associated with 60 canonical pathways (Table 4.3), a number of which have previously been associated with OA (Giatromanolaki *et al.*, 2001a; Giatromanolaki *et al.*, 2003a; Velasco *et al.*, 2010b; Huang *et al.*, 2011b; Li *et al.*, 2011b). Interestingly, the pathway analysis identified a possible role for *IL17* signalling in OA, generally based upon the altered expression of *AKT3, MAP2K2, MAPK1* and *NFKB2*. In all, eight cancer signalling pathways showed a significant over-representation within the dataset. Within these cancer pathways a total of 45 genes were differentially expressed, with altered

AKT3, CDKN1A, FZD2, FDZ4, FDZ7, FDZ9, MAP2K2, MAPK1, PDGFC and *SMO* expression providing a common link between the majority of these pathways.

Ingenuity Canonical Pathways	P-value	Ratio
Aryl Hydrocarbon Receptor Signalling	1.62E-05	0.15
IL17A Signalling in Fibroblasts	2.29E-05	0.26
Colorectal Cancer Metastasis Signalling	2.75E-05	0.13
Glioblastoma Multiforme Signalling	4.47E-05	0.14
ILK Signalling	8.91E-05	0.13
Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis	9.12E-05	0.12
Molecular Mechanisms of Cancer	1.00E-04	0.1
Glycosaminoglycan Degradation	1.15E-04	0.3
Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis	2.09E-04	0.11
Pancreatic Adenocarcinoma Signalling	4.57E-04	0.14
Factors Promoting Cardiogenesis in Vertebrates	5.37E-04	0.15
Basal Cell Carcinoma Signalling	7.24E-04	0.17
Role of IL17F in Allergic Inflammatory Airway Diseases	7.59E-04	0.2
IL17 Signalling	9.55E-04	0.16
Wnt/β-catenin Signalling	1.05E-03	0.12
TREM1 Signalling	1.07E-03	0.18
PI3K/AKT Signalling	1.66E-03	0.12
Human Embryonic Stem Cell Pluripotency	1.95E-03	0.12
Ovarian Cancer Signalling	1.95E-03	0.12
p53 Signalling	3.09E-03	0.14
Arginine and Proline Metabolism	3.09E-03	0.15
Oncostatin M Signalling	3.16E-03	0.21
Interferon Signalling	3.16E-03	0.21
Hepatic Fibrosis / Hepatic Stellate Cell Activation	3.63E-03	0.12
Role of IL17A in Arthritis	3.98E-03	0.15
Role of NANOG in Mammalian Embryonic Stem Cell Pluripotency	4.79E-03	0.12
O-Glycan Biosynthesis	5.13E-03	0.21
Bladder Cancer Signalling	5.89E-03	0.13
HIF1α Signalling	7.41E-03	0.12
Prostate Cancer Signalling	7.76E-03	0.12
Protein Kinase A Signalling	8.32E-03	0.09
Circadian Rhythm Signalling	8.51E-03	0.18
p38 MAPK Signalling	9.33E-03	0.12
PTEN Signalling	1.10E-02	0.11
IL-17A Signalling in Airway Cells	1.12E-02	0.13
Caveolar-mediated Endocytosis Signalling	1.12E-02	0.12
Corticotropin Releasing Hormone Signalling	1.17E-02	0.11
Inhibition of Angiogenesis by TSP1	1.35E-02	0.18
IL-8 Signalling	1.35E-02	0.1

Role of Wnt/GSK-3α Signalling in the Pathogenesis of Influenza	1.58E-02	0.12
Endoplasmic Reticulum Stress Pathway	1.91E-02	0.22
Ascorbate and Aldarate Metabolism	1.91E-02	0.22
Production of Nitric Oxide and Reactive Oxygen Species in Macrophages	1.95E-02	0.1
IL-12 Signalling and Production in Macrophages	2.69E-02	0.1
Role of Tissue Factor in Cancer	2.69E-02	0.11
Dendritic Cell Maturation	2.69E-02	0.09
Amyloid Processing	2.88E-02	0.13
Intrinsic Prothrombin Activation Pathway	2.95E-02	0.16
LXR/RXR Activation	3.39E-02	0.11
Role of IL-17A in Psoriasis	3.72E-02	0.23
Urea Cycle and Metabolism of Amino Groups	3.80E-02	0.16
NRF2-mediated Oxidative Stress Response	3.89E-02	0.09
VDR/RXR Activation	3.98E-02	0.11
Reelin Signalling in Neurons	4.27E-02	0.11
Hypoxia Signalling in the Cardiovascular System	4.27E-02	0.12
Glycosphingolipid Biosynthesis - Neolactoseries	4.37E-02	0.17
Cell Cycle: G1/S Checkpoint Regulation	4.57E-02	0.12
Atherosclerosis Signalling	4.68E-02	0.1
Ephrin Receptor Signalling	4.68E-02	0.08
Glioma Invasiveness Signalling	4.90E-02	0.12

Table 4.3 Associated pathways of the differentially expressed genes. Sixty pathways in total were identified as associated with the differentially expressed genes. Several of these (bold-italicised) have been previously reported as associated with OA. Six pathways are related with *IL17* signalling. The ratio column is the proportion of differentially expressed genes divided by the total number of genes associated with a pathway.

Network analysis was performed on differentially expressed genes with fold change ≥ 2 using STRING. In total, 21 networks were found enriched in the genes, of which 12 consisted of only two nodes (Fig. 4.4). Five up- and 14 down-regulated genes were assigned as hubs (Table 4.4). Gene *SPARC* and *COL2A1* had the most interactions and were concentrated to the largest network of 27 genes. Seven up-regulated collagens in OA were also in this network.

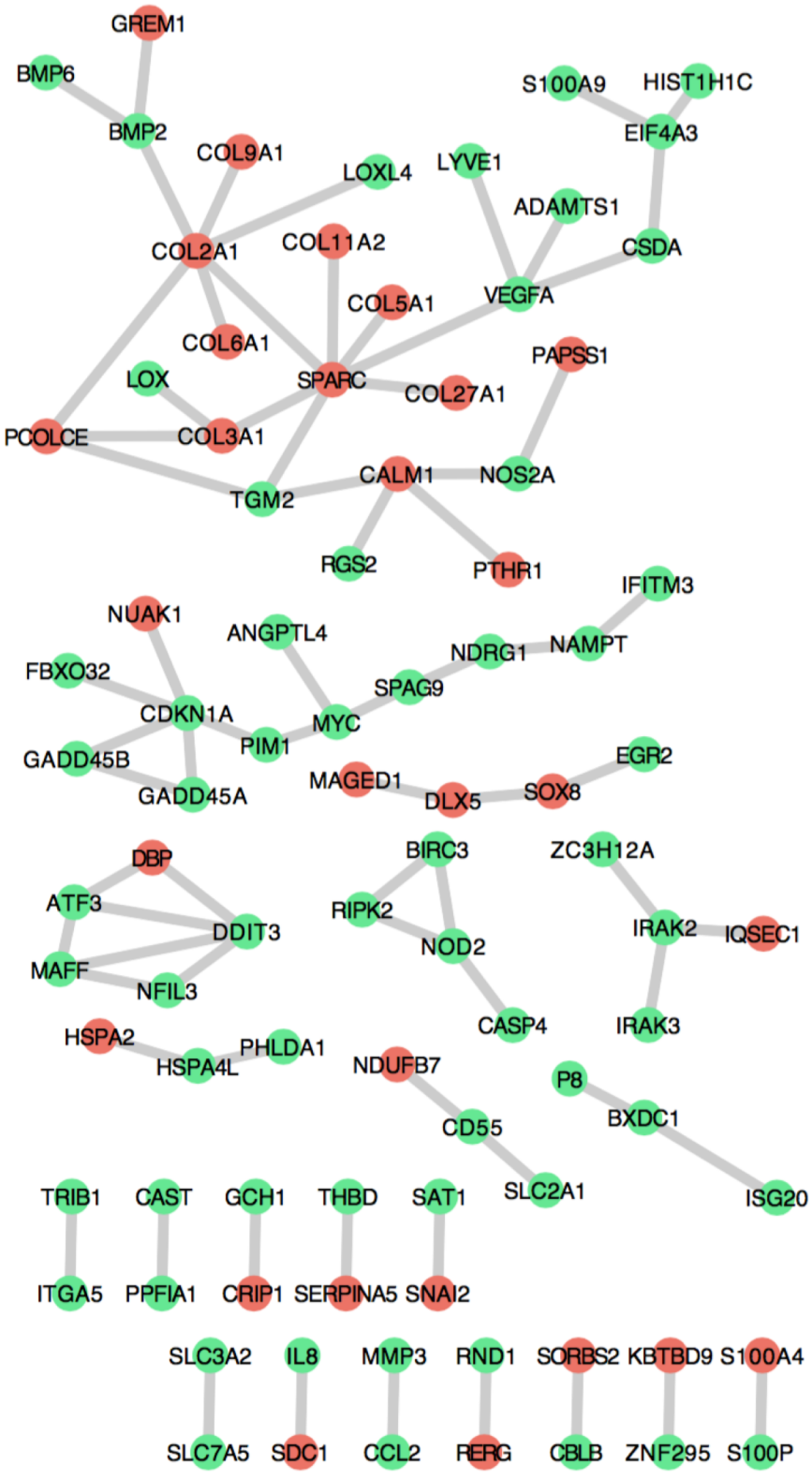


Figure 4.4 Protein interaction networks of genes. Network analysis was performed on differentially expressed genes with fold change ≥ 2 using STRING. Down-regulated genes are in green and up-regulated genes are in red. Genes with no interactions were not shown in the figure.

Symbol	Corrected p-value	Fold Change	Connections
SPARC	3.43E-05	5.65	7
COL2A1	1.50E-05	4.50	6
CDKN1A	2.56E-10	-6.44	5
CALM1	1.16E-08	2.45	4
DDIT3	1.38E-07	-4.37	4
VEGFA	2.02E-05	-2.71	4
ATF3	4.89E-09	-5.07	3
TGM2	6.25E-08	-6.93	3
NOD2	1.55E-07	-4.49	3
BMP2	3.07E-07	-7.21	3
MYC	2.19E-06	-3.15	3
EIF4A3	6.08E-06	-2.04	3
PCOLCE	1.28E-05	3.40	3
MAFF	4.60E-05	-3.17	3
IRAK2	1.78E-04	-2.59	3
COL3A1	3.70E-03	4.04	3
NFIL3	1.08E-09	-3.50	2
HSPA4L	1.03E-08	-2.16	2
SPAG9	1.20E-08	-2.54	2

Table 4.4 Differentially expressed genes that have more than 5 interactions with other differentially expressed genes.

4.4.5 Comparison of knee vs. hip OA gene expression

This is the first study to comprehensively investigate gene expression changes of hip OA cartilage. However, as mentioned earlier, a similar study using knee OA versus normal cartilage has been performed (Karlsson *et al.*, 2010b). In total, 1423 genes showed differential expression in this knee study, of which only 265 genes were also differentially expressed here (Figure 4.5 A and Additional Table A4.3). In fact, of these overlapping genes, only 183 (13% of the 1423) showed regulation in the same direction although there was a significant positive correlation for such a trend (Figure 4.5 B and C). Genes that increased in both tissues with disease included a large number of collagens, including *COL2A1, COL3A1, COL5A1, COL5A2, COL6A1, COL8A2, COL11A1, COL12A1* and *COL13A1*. Remarkably, given the small gene overlap, pathway analysis showed that 35 out of 60 canonical pathways associated with our gene list ($P \leq 0.05$) are also associated with knee OA (Figure 4.5 D and Additional Table A4.4). Of these, 'Role of Macrophages, Fibroblasts and Endothelial Cells in

Rheumatoid Arthritis' and 'Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis' contained the largest number of overlapping differentially expressed genes. The 'TREM1 signalling' pathway contained the highest ratio of differentially expressed genes (around 9% of all genes within that pathway).

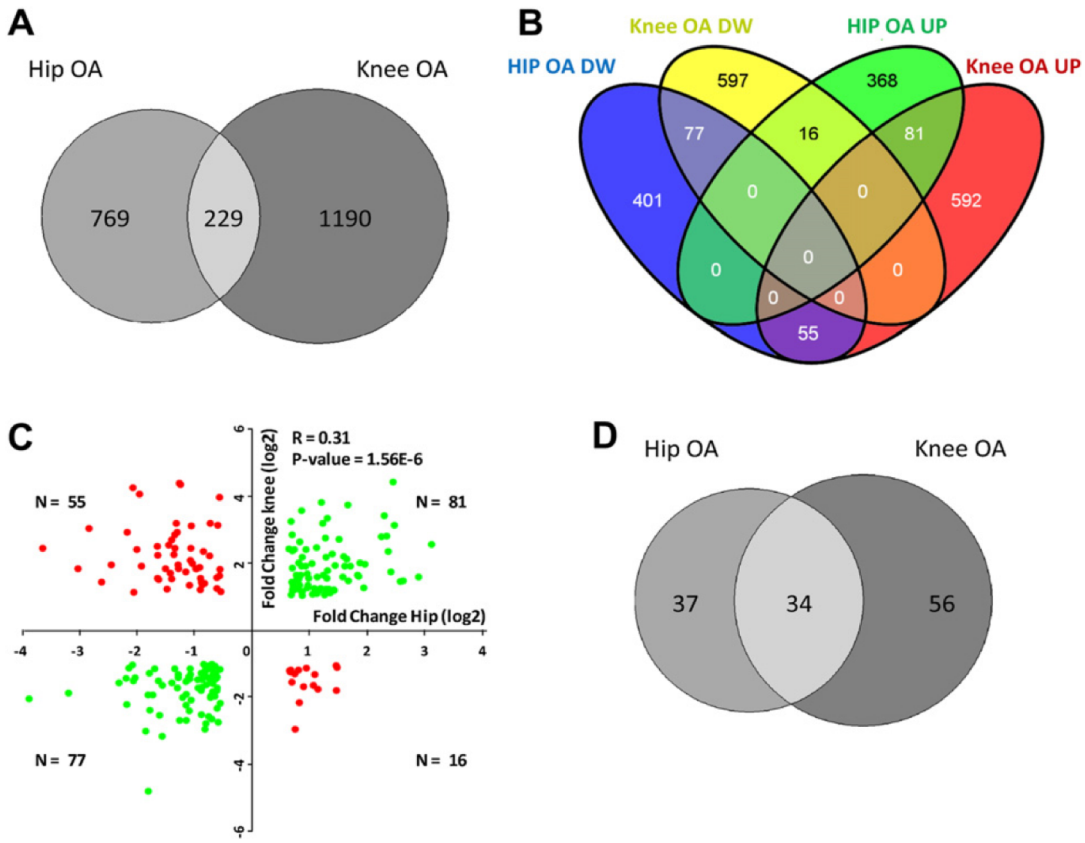


Figure 4.5 A comparison between hip and knee OA gene expression changes. A, size-dependent Venn diagram of the differentially expressed gene lists from the two studies. Only 265 genes were identified as differentially expressed in common between both studies. **B**, 4-way Venn diagram of hip and knee differentially expressed genes. DW, down-regulated; UP, up-regulated. **C**, scatter plot of fold changes of the overlapped genes. X-axis represents the fold change result of the genes in this study (hip) and Y-axis the knee OA study (Karlsson *et al.*, 2010b). Green points indicate genes that are significantly differentially expressed in the same direction in both studies. Red points indicate genes that are regulated in the opposite direction. N, number of genes in each quadrant. Despite the large discrepancy in differentially expressed genes, the fold changes of these overlapping genes are significantly co-related. **D**, size-dependent Venn diagram of the pathways identified as associated with knee and hip OA gene lists.

4.5 Discussion

To my knowledge, this is the first report of comprehensively relate gene expression changes in humans between normal (NOF) and OA hip cartilage at the whole genome level. Several studies have compared gene changes using similar samples via a real-time RT-PCR technique but have generally focussed on specific genes, originally metalloproteinase family members (Kevorkian et al., 2004b; Davidson et al., 2006), and more recently the entire degradome (Swingler et al., 2009a). Although the datasets were different, reflecting the differences in techniques, all the proteases expressed differentially in both screens followed identical expression patterns in terms of direction (up- or down-regulation), which validated our samples and the microarray analysis approach that I have taken here and provided the confidence to proceed to the RNAseq experiment.

Besides the RNA extraction method, a well-designed tissue collection and handling strategy is required in order to obtain accurate and reliable expression profiles, because many factors, such as the tissue storage time in different temperatures or reagents (Mutter *et al.*, 2004; Espina *et al.*, 2009; Hatzis *et al.*, 2011), can affect the RNA integrity thus introduce variance into the data. In the microarray experiment, considering the sensitivity of the technology and the difficulties in RNA extractions from the human cartilage tissue, the RNA samples from all of our available tissues were used, as the quantity of RNAs was the only concern at the time. Inevitably, a careful quality control on the raw data was performed and 1/3 of the samples were marked as outliers. However, there is no clear biological factor explaining the difference of expression profiles between these samples and the others. RNA integrity numbers, age of patients and Noyes scores of the outliers are all similar with those retained. This implies that the difference lies in the tissue collection and handling method. Each joint tissue might be handled differently before collection from operating theatres. Furthermore, there were different time intervals between surgery completion and collection as well as between collection and harvesting cartilage from the tissue. These may introduce extra non-genetic differences between RNA samples. However, due to the limited availability of tissue samples, especially NOFs, these were unavoidable and required thorough quality control on the expression data.

In total, 1151 genes were found differentially expressed with a fold change ≥1.5 and p-value ≤ 0.01, in line with the only other whole genome human cartilage study (Karlsson *et al.*, 2010b). Using functional enrichment analysis 'receptor and transmembrane receptor signalling activity' appeared up regulated in OA. 'Oxidoreductase activity' was significantly down-regulated which is in keeping with a previous report that found a down-regulation of oxidative damage defence genes, including *SOD2* and *SOD3*, in OA cartilage (Aigner *et al.*, 2006b; Scott *et al.*, 2010a). A decrease in oxidative defence/reductase activity would lead to an increase in reactive oxygen species and oxidative stress, a process that has been demonstrated in OA and shown to negatively affect chondrocyte function (Yudoh et al., 2005).

The Ingenuity canonical pathway analysis identified several pathways based upon *IL17* or cancer signalling. Common to these were a number of genes including *MAP2K2*, *MAPK1* and *AKT3*. In fact, these factors were also present in several other highlighted pathways, including phosphatidylinositol 3-kinase (*PI3K*) signalling in B lymphocytes. The CC-chemokine, *CCL20* was the most down regulated gene identified from our array and also appears to be an important target for *IL17* signalling (Onishi and Gaffen, 2010). *CCL20* binds to the receptor *CCR6* and interestingly signals via the Akt and/or ERK MAP Kinase pathway. *CCL20* acts as a chemoattractant for *CCR6*-expressing cells such as dendritic cells. There has been little research on *CCL20* in OA pathogenesis but it has been suggested that it contributes to ECM-bone remodelling (Lisignoli et al., 2009). However, in rheumatoid arthritis (RA), *CCL20* produced by synovial cells is thought to play a pivotal role in the recruitment of arthritogenic Th17 ($CD4^+$ T cells that secrete *IL17A*) cells to the inflamed joint (Hirota et al., 2007).

SPARC was the most connected gene in the STRING analysis. The gene encodes osteonectin and found up-regulated in OA in several studies (Nakamura *et al.*, 1996; Nanba *et al.*, 1997). It plays a vital role in collagen mineralization (Termine *et al.*, 1981; Maurer *et al.*, 1995), which explains its interactions with the collagen in the network. *ATF3* (activating transcription factor 3) is a transcription factor induced by various stress signals and proposed to be a hub of the cellular adaptive-response network (Yatsugi et al., 2000), it was also among the gene with the most STRING connections from our study. *MYC* expression in OA cartilage correlates with apoptosis of the chondrocytes (Hai et al., 2010). Both *GADD45α* and *GADD45β* were also found in the

enriched networks with both having been previously identified as down regulated in OA cartilage as part of a microarray analysis. *GADD45β* has been proposed to play a role in chondrocyte homeostasis via the regulation of collagen gene expression and the promotion of cell survival (Ijiri et al., 2008).

Herein gene expression differences between hip and knee OA were also compared, providing lists of genes and molecular pathways common to both disorders. Common between the two tissues was an increase in a large number of ECM-associated genes, especially collagens, purportedly as an ineffectual repair response (Dell'accio and Vincent, 2010). Similar increases in matrix genes with OA have been reported by others (Aigner *et al.*, 2006b). Importantly, this work also highlights gene differences between OA in both tissues. A clear difference in the hip and knee datasets is present when examining the expression of metalloproteinases. The expression of the collagenase gene *MMP1* and the aggrecanase gene *ADAMTS5*, along with *ADAMTS1*, were all increased in knee OA cartilage but decreased in hip OA cartilage, an observation that has been reproducibly observed (Kevorkian et al., 2004b; Davidson et al., 2006; Swingler et al., 2009a). This could suggest the disease mechanisms are fundamentally different, reflect the stage of disease at which hip or knee replacement surgery occurs, or be due to the normal comparator group, healthy tissue (Karlsson *et al.*, 2010b) or NOF fracture herein. In a preliminary study, gene expression within NOF fracture or post-mortem was largely similar, validating the use of NOF as a control tissue (Kevorkian et al., 2004b). The OA femoral tissue used within this study was end-stage disease with significant cartilage damage, as exemplified by the high score using the modified Noyes system (Kijowski et al., 2006). However, I only examined gene expression in macroscopically normal cartilage taken from the OA patients, yet significant gene expression differences were observed between this cartilage and the NOF control tissue. This finding supports the concept that even OA cartilage tissue of healthy macroscopic appearance is not necessarily free of disease (Aigner *et al.*, 2006b). The OA tissue used was taken from patients undergoing joint replacement surgery therefore the gene expression herein will likely be distinct from those occurring at initiation or during disease progression. The relatively small number of overlapping differentially expressed genes between our hip and the knee OA studies may reflect a difference in the tissues or the analysis undertaken (Miklos and Maleszka, 2004). However, even taking this into consideration, a remarkable number of pathways appeared to be conserved features of the disease in

both tissues. These included ‘Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis’ and ‘Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis’ perhaps unsurprisingly given their link to arthritis, and Wnt/b-catenin, IL-8 and IL-17 Signaling. However, ‘Oncostatin M Signaling’ contained the highest proportion of differentially expressed genes within a given pathway (approximately 21%). Oncostatin M in combination with other pro-inflammatory cytokines has been extensively linked to the induction of metalloproteinases by chondrocytes (Rowan and Young, 2007). ‘triggering receptor expressed on myeloid cells (TREM1) signalling’ also contained a high proportion of differentially expressed genes from both tissue gene lists. TREM proteins are a family of cell surface receptors that participate in diverse cell processes including inflammation, where they act in concert with other receptors to amplify an inflammatory response, and bone homeostasis where they play a role in osteoclastogenesis (Klesney-Tait et al., 2006). The role of TREM signaling in chondrocytes and cartilage remains to be determined, though TREM-1 is up-regulated in RA synovium (Kuai et al., 2009) and is proposed as a new therapeutic target in the disease (Kim et al., 2012). A number of pathways associate only with hip OA cartilage including pathways connected with proteoglycans ‘Glycosaminoglycan Degradation’ and ‘O-Glycan Biosynthesis’ and more novel pathways such as ‘Circadian Rhythm Signaling’. Knee OA specific pathways included ‘IGF-1 Signaling’ and ‘Ephrin Signaling’, both of which have established links with the disease or chondrocytes (Denko and Malemud, 2005; Kwan Tat et al., 2009).

In conclusion, to my knowledge this is the first study to compare gene expression changes in osteoarthritic femoral hip cartilage to that of patients with no signs of OA at the whole-genome level. These data have identified a number of novel pathways, such as IL- 17 signaling along with a number of genes that appear integral to signaling, including *CCL20* and *MAPK1,* and subsequent responses, such as *ATF3*, all of which may have a role in OA pathogenesis. Importantly, by comparing gene expression changes between hip and knee OA both commonality, such as the ‘TREM1 signaling pathways’, and discord such as the ‘O-Glycan Biosynthesis’ in hip and ‘IGF-1 signaling’ in knee OA are observed. Although proteolytic loss of cartilage typifies OA (Rowan et al., 2008) in both hip and knee joints, our observations add to the notion that the molecular mechanisms underpinning such destruction may have some unique and site specific differences. Exploitation of these may offer the potential for more tailored,

joint-specific therapies that circumvent the negative aspects associated with direct metalloproteinase inhibition (Rowan et al., 2008).

Chapter 5 Transcriptome analysis of RNAseq data

5.1 Introduction

Via the microarray study, not only the quality of RNAs extracted from cartilage but also the differences at the gene level between OA samples and NOF were identified. A number of canonical pathways were found to be associated with these differentially expressed genes. Apart from those well studied and largely expanded cancer pathways, several known OA associated pathways were found among the list (Giatromanolaki *et al.*, 2001a; Giatromanolaki *et al.*, 2003a; Velasco *et al.*, 2010b; Huang *et al.*, 2011b; Li *et al.*, 2011b) as well as number of other pathways that were not reported to be associated with hip OA before, such as TREM1 signaling and IL17 signaling pathways. By comparing these pathways to those reported in a knee OA study, conducted with the similar technologies, it was revealed that the molecular mechanisms underpinning the cartilage destruction might have some specific joint site differences. In addition, 1151 genes were found significant differentially expressed (DE) in OA. Among these genes, increased expression of collagens (*COL2A1*，*COL3A1*，*COL5A2, COL9A1, COL11A1*) and *GDF10*, and decreased expression of aggrecanases (*ADAMTS5* and *ADAMTS9*) were observed, all of which were previously characterized for OA in numbers of publications (Cagnard et al.; Kevorkian et al., 2004b; Chou et al., 2013). Growth factors *GDF5*, a gene found associated with OA susceptibility through GWAS studies (Evangelou et al., 2009) and which plays an important role in cartilage maintenance (Francis-West et al., 1999), was also found significantly up regulated in OA in our experiment. A large decrease in expression of *SOD2* (fold change = -14.7) was found, indicating the oxidative stress in OA and confirming the finding of (Scott et al., 2010a).

Our results also highlighted the limitation of the microarray technology used to determine gene expression, which suffers from the limited dynamic detection range, high background noise and relatively low resolution. During quality control of the microarray data, data from almost half of the probes on the Illumina (Illumina Inc.

California, U.S.) chip were removed, because of either doubtable detection or very low detected expression in all of the RNA samples. Within DE genes called, several other well-described gene expression changes were not found. For example, up regulation of *ADAMTS2* and *MMP16* described in other studies were not found in ours (Kevorkian *et al.*, 2004a; Swingler *et al.*, 2009b). Up-regulation of *BMP2* was described as a characteristic of OA chondrocytes (Fukui *et al.*, 2003; Nakase *et al.*, 2003) but significant down-regulation was found in our study. Plus, for the canonical pathways that were found associated with the disease, only a small portion of genes involved in them was found significantly differentially changed. One of the reasons of these undetected DE genes could be the difference of tissues/animal models used in the study, but could also be due to the limitations of microarray technology. Furthermore, microarray technology is also incapable of detecting novel genes, as it depends on the prior knowledge of gene sequences. Moreover, the nature of OA, as a complex disease, involves multiple factors, such as differential allelic expression (DAE) of certain genes (Raine *et al.*, 2013; Syddall *et al.*, 2013); alternative splicing was also shown to be associated with the disease (Takada *et al.*, 2011). Such information cannot be obtained with microarray technology. Hence, to explore the transcriptome with a more sensitive technology and to understand the molecular changes of OA comprehensively, the recently emerged technology RNA sequencing was used to investigate the disease transcriptome of the hip cartilage further.

The scripts used for the data analysis, including bash commands, Perl and R scripts, are available on Github:
https://github.com/byb121/Thesis_2015/tree/master/Thesis_2015/scripts .

5.2 Aims of this study

Although we are beginning to understand the disease processes that occur in cartilage as OA progresses such research is still in its infancy. It is very likely that changes in the regulation genes of cartilage cells, chondrocytes, play a strong role in the disease process. Studies (Evans *et al.*, 2004; Menendez *et al.*, 2011; Ahmed *et al.*, 2012) have made progress in identifying these gene changes with the eventual hope that a drug can be made to block the action of a given gene/pathway and treat the disease, but technological limitations, combined with the difficulties in isolation of RNA/Nucleic

acids from cartilage, have prevented to study all gene changes simultaneously. However, recent advances of RNAseq now make it possible to identify novel gene changes in OA cartilage.

5.2.1 Identification of gene expression changes in addition to those observed with microarrays

RNAseq has several advantages in determining of transcripts abundance comparing to the microarray. Its detecting range is not limited and suffers less background noise, thus the normalization of RNAseq data is also less complicated. RNAseq can also be used to identify novel transcripts expressed in cartilage and their abundances. The knowledge of these can be complementary to the known molecular mechanism of OA as well as being potential targets of the disease therapy.

5.2.2 Determination of transcript expression changes in OA

With RNAseq data, abundance of transcripts on a genome-wide scale can be determined. A number of genes, often different genes, have been found differentially expressed in OA in many studies, but in fact the abundances of transcripts determine the phenotypes. Determining the DE transcripts of cartilage is a vital step to move our understanding of OA mechanism to the transcript level.

5.2.3 Identification of alternative splicing on a genome-wide scale

Alternative splicing variants of transcripts of certain genes were found to be associated with OA progression in cartilage (Berardi *et al.*, 2001; Parker *et al.*, 2002; DuRaine *et al.*, 2011), however, a genome-wide investigation of alternative splicing events between NOF and OA cartilage is absent to date. With the RNAseq data, the splicing junctions can be identified with/without assembly transcripts beforehand, as some of the RNAseq reads will cover these junctions, especially when using paired-end sequencing protocols.

5.2.4 Identification differential allelic expression on a genome-wide scale

Genome-wide association analysis of OA has revealed that a number of SNPs are associated with susceptibility of OA (Zeggini *et al.*, 2012). Several of them were further studied with allelic expression analysis to investigate their effects on gene expression in

cartilage (Ratnayake *et al.*, 2012; Raine *et al.*, 2013; Syddall *et al.*, 2013). It is also reported that the phenomenon of imbalanced allelic expression of certain genes is associated with OA progression in different types of joint tissues (Southam *et al.*, 2007; Egli *et al.*, 2009), therefore a genome-wide study of allelic expression of genes and their association with OA will further our understanding of the relationship between the OA susceptibility and SNPs, and hopefully some of them can be used as biomarkers for the disease. With the transcriptome in single-base resolution derived from RNAseq data, bases on both alleles of a single loci can be observed along with relative abundances of each. This enables allelic expression on a genome-wide scale.

5.2.5 Identification of RNA-editing events and evaluation of their association with OA

Genome-wide RNA-editing events can be identified with RNAseq data (Ramaswami *et al.*, 2013). With respect of the fact that thousands of such events can happen in human (Ramaswami *et al.*, 2013) and some of them may alter the protein function, they ought to be studied for their impact on OA.

5.2.6 To define a workflow for the analysis the RNAseq data

Existing open source and commercialized software for analysis of second-generation sequencing results are not mature. Most of the available software can only do one section of the whole pipeline for RNAseq data analysis. The exact approaches and parameter for such analysis are not commonly recognized neither. Thus it is essential to explore these factors in the project.

5.3 Methods

5.3.1 Overview of the workflow for RNAseq data analysis

At the start of the project there was no commonly recognized pipeline/software for the analysis of RNAseq data available. As described, commonly used and commercially freely available software has differing advantages, so the selection of the software assigned to different part of the whole bioinformatics workflow is critical, but can vary depending on the purpose of the analysis. For our RNAseq study, the aim was to reveal

the difference of transcriptome characteristics between OA and NOF hip cartilage, including differential expressions, alternative splicing events, transcript sequences and differential allelic expressions, thus the software chosen for use in each step of our pipeline was selected to favour the chance of detecting these differences.

Our pipeline is illustrated in Figure 5.1, which involves a number of software tools. To improve the accuracy of mapping, FastQC was used to first check quality of the reads, as it can provide several quality metrics of the reads, from which essential parameters for processing reads in the next step can be deduced. Trim-galore (Krueger) was then used to trim off low quality bases from both ends of reads and remove sequencing adaptors when they were found at 3' end. After quality control, Tophat was used to map reads to the human reference genome hg19. To identify differentially expressed genes, the number of reads that aligned onto each transcript for each sample was counted with htseq-count (Anders *et al.*, 2014), then DESeq (Love *et al.*, 2014) used to identify DE genes. BitSeq (Glaus *et al.*, 2012) was used to identify DE transcripts. In order to allow BitSeq to estimate the coverage of transcripts and likelihood of a read correctly mapped to its originated transcript, RNAseq reads were aligned to the human transcriptome (cDNAs of known human transcripts, ENSEMBL 74) with Novoalign (www.novocraft.com), which had better mapping accuracy comparing to other aligners at the time (Ruffalo *et al.*, 2011) and also allows multiple placements (up to 100 different places) of a read. To assemble transcripts, Cufflinks (Trapnell *et al.*, 2013) was used after reads were aligned to the human genome (hg19) with GSNAP (Wu and Nacu, 2010), as the algorithm of the software supports gapped alignments and takes advantages of known splicing sites and SNPs. Cuffmerge (Trapnell *et al.*, 2013) was then used to merge transcripts assembled using Cufflinks from different samples. The coverage of the transcripts was estimated at the same time, so the sequences of transcripts expressed in OA and NOF cartilage can be obtained along with the relative abundances. Although Cufflinks uses known transcripts as template, it has the ability to identify novel transcripts. For allelic expression analysis and identification of RNA-editing events, the base variants at the mRNA level first need to be identified. This was done with Samtools. A Perl script was written to extract the coverage of the variants from mapped reads, thus the abundance difference between the transcripts from the two alleles could be derived. In order to determine the association between the identified

RNA-editing events and OA, in house R (R Core Team, 2012) scripts were used.

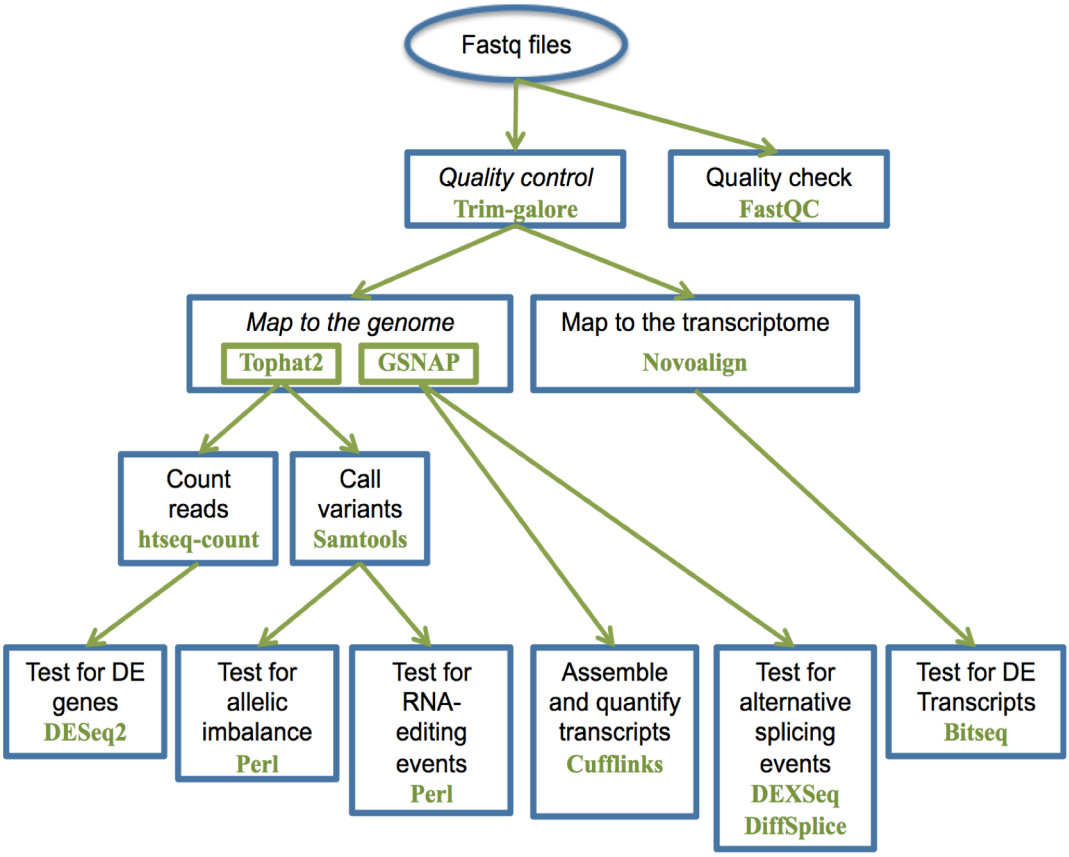


Figure 5.1 the workflow of the RNAseq analysis and software. Quality of the fastq files was checked with FastQC, low quality bases (Q<=20) and sequencing adaptor contaminations with Trim-glore. Quality filtered reads were then mapped to the human reference genome hg19 with Tophat2 and GSNAP. In order to call DE genes, reads mapped to each gene were counted with htseq-count from Tophat2 alignments then imported to R. DESeq2 were used to normalized the counts data and test for DE genes. The sequence variants were called with Samtools from Tophat2 alignment. Common heterozygous variants (found in 80% of samples in either condition) were used to calculate allelic imbalance of genes with the scripts written in Perl. All of the common variants were used to identify RNA-editing events and test if the frequencies of any RNA-editing event in the whole genome are significantly different in OA samples comparing to NOF samples. A script written in Perl was used for the test. In order to assemble and quantify transcripts, Cufflinks was used with GSNAP alignments. The alignments were also used as input of DEXSeq and DiffSplice to test for alternative splicing events in OA samples. To test for differentially expressed transcripts using Bitseq, Novoalign was used to map reads to the transcriptome with multiple placements allowed for reads so that Bitseq can weight the read alignment to transcripts and test for differentially expressed transcripts.

5.3.2 Quality assessment of raw reads

Quality of the raw sequencing reads was inspected with FastQC. Low quality bases (Q ≤ 20) on the 3' end of reads were trimmed off. Reads of length shorter than 20 after trimming were discarded. Possible contaminations of sequencing adaptors were removed with Trim Galore. Reads that passed the quality criteria were used as input for the subsequent analyses. To further assess the sequencing quality, these reads were mapped to the transcripts recorded in the ENSEMBL database (Version 75) with Novoalign (Version 2.07.13) and the human genome hg19 with GSNAP (Version 2012-06-20). The mapped reads were counted with Samtools (Version 0.1.18) and BEDtools (Version 2.15.0) (Quinlan and Hall, 2010) for number of mapped reads, number of reads mapped to the transciptome, number of reads mapped to the human genome. Number of reads mapped to genes were counted with htseq-count from the whole genome alignments of samples and then normalized with bioconductor CQN (Hansen *et al.*, 2012) package. Genes that have Reads Per Kilo bases of the gene per Million bases of read (RPKM) >= 0.3 in 80% of samples in either condition were considered as expressed and the normalized counts of these were then used to calculated sample distances to cluster samples and plot heatmaps with the function implemented in the DESeq package. The RPKM of 0.3 was suggested as a threshold when RNAseq was introduced in 2008 (Mortazavi *et al.*, 2008).

5.3.3 Identification of differentially expressed genes

The reads were then mapped with Tophat2 to the human reference genome (hg19). Reads mapped to each gene were then counted with ht-seq count. Genes were defined as annotated in GENCODE (GENCODE human genes V19). To filter unreliable gene counts from lowly expressed genes in samples, the expressed copy number of each gene for each samples were calculated as: Expressed copy number of a gene = (read counts of the gene in a sample * read length)/(exonic length of the gene). The exonic length of a gene is defined as the non-overlap total length of exons annotated for the gene. Genes that have at least 1 copy in 80% of samples in either condition were considered as expressed genes in hip cartilages. Bioconductor package DESeq2 (Version 1.4.2) were then used to test for differentially expressed genes from the counts data of all genes. P-values of detected expression changes were corrected with Benjamini & Hochberg algorithm. Genes that were identified as expressed genes and have at least two fold changes between OA and NOF samples with P-values ≤ 0.05 were considered as

differentially expressed genes for further analysis.

5.3.4 Identification of differentially expressed transcripts

Reads after quality control as described in (Chapter 5, 5.3.2) were then mapped to the ENSEMBL transcript sequences (Version 74) with Novoalign (Version 2.07.13). Reads were allowed to map onto multiple transcripts. Transcript abundances were measured and differentially expressed transcripts were called with BitSeq (Version 0.7.0). Transcripts that had at least 2 fold change between OA and NOF samples with positive probability of log ratio either ≥ 0.95 or ≤ 0.05 were considered as significantly changed.

5.3.5 Assembly of transcripts and identification of novel transcripts

Following the protocol recommended by the authors of the Cufflinks (Trapnell *et al.*, 2012), quality controlled reads were mapped with Tophat2 (Version 2.0.10) and then assembled with Cufflinks (Version 2.0.2). Transcripts that have fragments per kilo-bases of transcripts per million reads (FPKM) in lower 20% of 80% samples of either condition were not considered as expressed in cartilage. Expressed transcripts were compared with transcripts recorded in GENCODE (Version 14) to identify known and novel transcripts.

5.3.6 Identification of alternative splicing events

Human gene annotations were downloaded from GENCODE (Version 14) and processed with the annotation preparation script provided in the DEXSeq (Version 1.0.2) package for non-overlapped exons. Reads that mapped onto exons were counted and differentially used exons were tested using DEXSeq . Exons that have at least 2 fold change with P value ≤ 0.05 were considered as significant alternative splicing changes between OA and NOF cartilage samples.

5.3.7 Allelic expression analysis

After quality controlled reads were mapped with Tophat (Version 2.0.10), sequencing variants were identified with Samtools (Version 0.1.18) then filtered by the following criteria:

a) The variants call quality score is ≥ 13. The quality score that Samtools produced

for each called variant is Phred scaled possibilities of the call being error (Quality score = -10*lg*P*), thus 13 equals p-value of 0.05;

b) The total coverage of the variant in one of the samples is ≥ 10, which means at least 10 reads covered the position.

The variants that can be observed in 60% of samples in either group of samples were analyzed further. The imbalanced rates were calculated as the ratio of the non-reference allele depth to the total depth of the loci. The rates were tested with an in-house Perl script plus further statistical analysis in R. The student T-test was used to test the association between imbalanced rates and OA.

5.3.8 Identification of RNA-editing events

Identification of RNA-editing events is depending on the variants identified. In order to identify RNA-editing events, sequencing variants were first called as described in the above Methods section. Heterozygous variants that fit in the expected RNA-editing changes (A-to-G change) were selected for further filtering. Each variant was then filtered with the following criteria:

a) The variants call quality is ≥ 13;

b) The total coverage of the variant in one of the samples is ≥ 10.

The reads supporting any of the two alleles of the variant in all of the samples were then counted and used to denote the frequency of editing events observed. The association between potential editing events and OA were tested using an in house R-script (Supplementary file S5.1) to compare the difference of the frequency of each editing event in OA and NOF samples. A linear mixed-effects model was used to test the association between the frequency of the editing events and OA. Fisher's exact test was used when one of the two alleles has no supporting reads in either group of the samples. P-values <= 10e-8 were considered as significant changes. P-value threshold was decided as: 0.05/(total number of the variants). The number 0.05 is a commonly recognized false discovery rate threshold.

5.4 Results

5.4.1 Cartilage and RNA Samples

In total 16 cartilage samples were collected from femoral heads of female OA donors (10 samples; median age = 73 yrs), and female NOF donors (6 samples; median age = 81 yrs) (Fig. 5.1A). The ages difference between the two groups are not significant (Mann-Whitney p value = 0.14). Macroscopic scoring, using the same scoring method as described in Chapter 4, confirmed the two sample groups were significantly different ($P = 0.0024$) with OA samples having a mean score of 4.9 and NOF 0.8 (Figure 5.2 B). In fact, there were total 29 OA and 16 NOF samples were collected, but only total RNAs extracted from these samples had the required quantity and quality (~5ug with RIN ≥ 7) for RNAseq at the time.

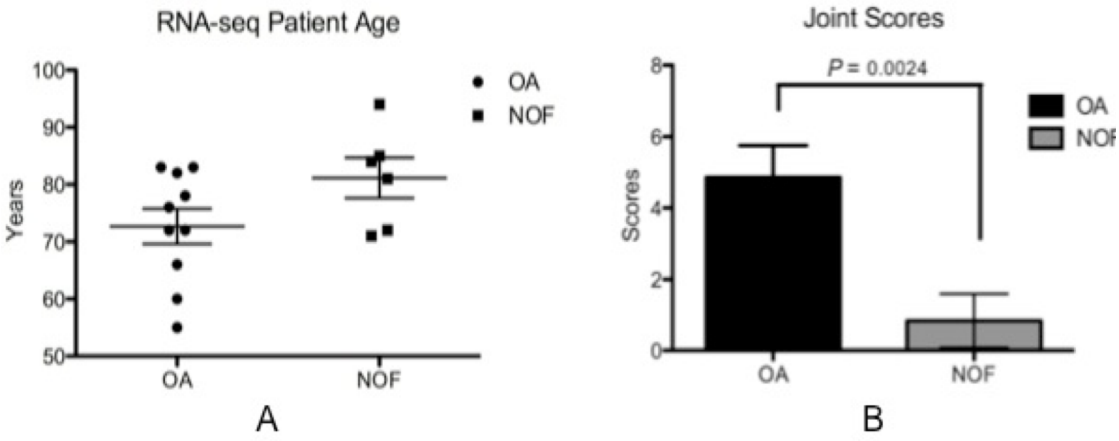


Figure 5.2 The ages of the patients and the modified Noyes scores of the joints. The figure shows **(A)** the ages of the patients and **(B)** the modified Noyes scores of the OA and NOF joints. Mean age of the NOF patients (81 years old) is greater than OA patients (73 years old). The difference between the two age groups is not significant using Mann-Whitney test (P = 0.14). The differences of the scores of the joints are significant between the two groups.

5.4.2 Quality of short reads and mapping

Quality of raw reads was checked with FastQC. Several aspects of the reads quality were plotted into figures, including: per base sequence quality, read length distribution and per base sequence content. Each of these represents a unique aspect of the read quality (Figure5.3). The quality scores drop along with positions of the reads. Low quality bases (Q ≤ 20) on the 3' end of reads were observed in all of the samples (Figure

5.3 A1). The length of all reads was 78bp, as we specified for the RNAseq experiment (Figure 5.3 B1). DNA base contents (A, C, G, T) were found inconstantly distributed in the first 13-14 bp of reads, which in fact is commonly observed across NGS data sets and originates from the random hexamer priming (Hansen *et al.*, 2010). The enrichments of random 5bp DNA polymers were also checked. This was calculated as the ratio of the observed frequency and the expected frequency of a polymer been seen on a position of reads using FastQC. Figure 5.3 D1 shows the patterns of most enriched polymers of one sample. Several short polymers were found enriched only within the first 14 bp of the reads, indicating potentially enriched sequencing patterns on those positions. The source of this polymer enrichment could be the bias from the hexamer priming or possible sequencing adaptor contamination at those positions. There were also polymers enriched between the 14^{th} base pair to the end of the reads, which indicates possible enrichment of duplicated reads. After quality control on the short reads, these quality statistics were checked again. These low quality tails were not found after removing low quality reads (Figure 5.3 A2). As low quality bases and sequencing adaptors contamination were removed, length of reads were not uniform, but most of the reads were still in their original length (Figure 5.3 B2). No short polymer enrichment was found after the 14^{th} base pair of the reads after removing adaptors contamination in the quality control step (Figure 5.3 D2).

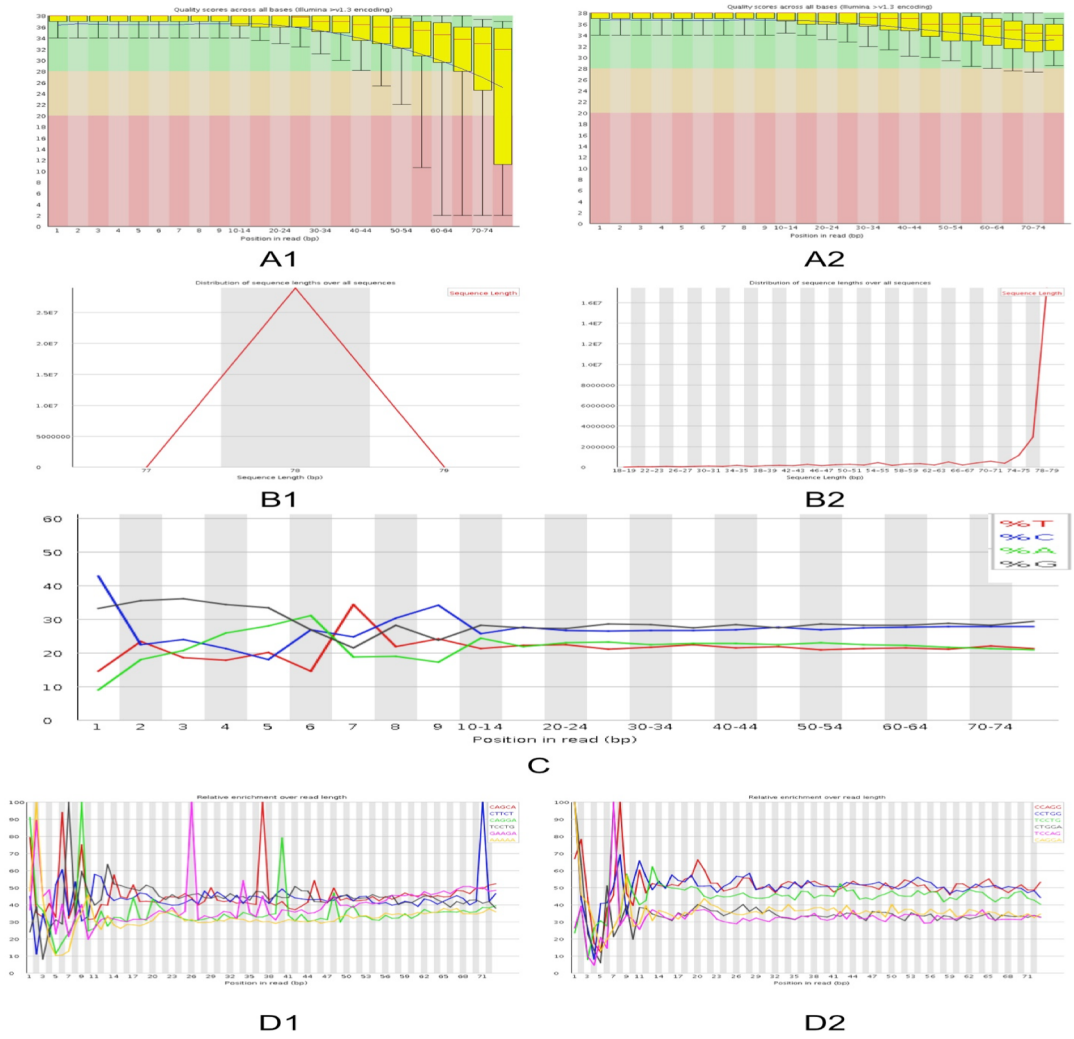


Figure 5.3: Quality control of the raw reads.
A: Box plots of base quality scores of read positions across all reads in a sample. The x-axis is the position of the reads. The y-axis shows the quality scores of the bases. **A1:** Before trimming off low quality 3' ends, the quality scores drops to below 20 near the end of the reads. **A2:** After trimming no bases have scores under below 20. **B:** Read length distribution. The x-axis is the length of reads, while y-axis shows the number of reads in a length. **B1:** Before trimming of low quality bases, all reads were in 78bp length. **B2:** After the trimming, the majority of the reads are still of their original length but some of the reads were shortened.
C: Percentage of each DNA bases in all reads on each position of the reads. The x-axis shows the positions of reads, the y-axis shows the base content in percentage on each position. A non-normal distribution is observed on the first 13-14bp.
D: Enrichment of short polymers on each position of reads. The x-axis shows the positions of the reads; y-axis shows the relative enrichment levels of random short polymers. **D1:** Short polymer enrichment can be observed both within the first 14 bp and beyond the position. **D2:** After trimming of low quality bases and removal of sequencing adaptor contaminations, polymer enrichments after the first 14 bp were removed, while for the first 14 bp, they remained, indicating the source of the enrichments are different.

The insert size of each sample was calculated using PICARD tool from the sequencing data. The insert sizes provided by the sequencing service provider were taken as expected, but how the insert sizes were determined by the company was not clear. The two sets of data were compared in the Table 5.1. The empirically determined insert sizes were shorter compared to the expected ones, however, samples with longer expected insert sizes appeared to have longer empirically insert size as well. This trend illustrated that the difference of the insert sizes could be from the different determining methods.

Sample	Type	Expected insert size	Expected standard deviation	Empirically determined insert sizes	Empirically determined standard deviation
N2080	**OA**	178	30	144.6	27.9
N1947	**OA**	183	30	143.0	34.2
N2049	**OA**	184	30	143.4	35.4
N2209	**OA**	189	30	153.5	31.7
N2004	**OA**	183	30	158.2	29.1
N1873	**OA**	180	30	151.7	31.7
N2062	**OA**	181	30	152.5	35.2
N2112	**OA**	195	30	156.7	54.1
N1901	**OA**	199	30	166.6	43.5
N1866	**OA**	198	30	166.7	40.3
N2002	**NOF**	190	30	150.7	32
N2024	**NOF**	189	30	155.3	33.7
N2060	**NOF**	196	30	165.9	30.8
N2064	**NOF**	185	30	157.5	37.1
N2059	**NOF**	198	30	163.0	46.3
N2120	**NOF**	180	30	142.1	35.4

Table 5.1: Expected insert size and empirically determined sizes. Expected insert sizes and the standard deviations were provided by the sequencing service provider.

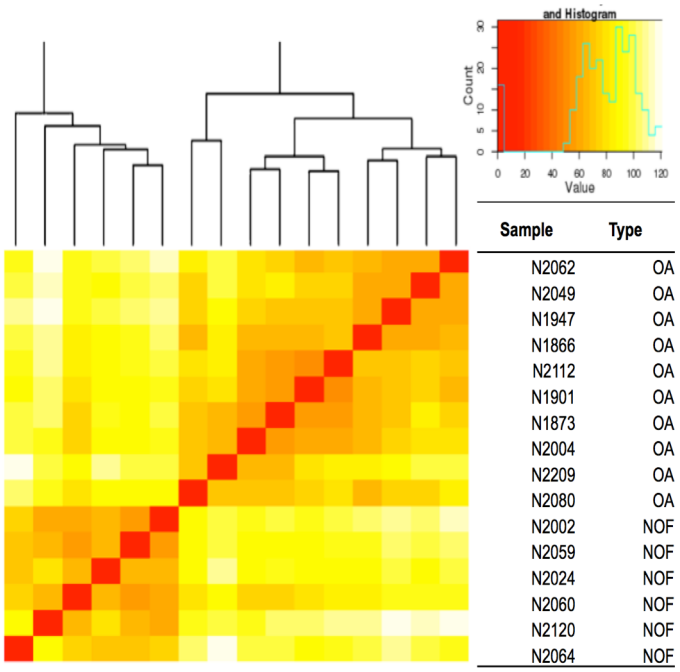
After removing 3' tails of low quality reads and sequencing adaptor contamination, on average around 27 million read pairs were retained for each sample, equaling on average 95% of the raw reads; Over 50% of the reads retained their original length (78 bases) and 75% were at ≥ 70bp long. When mapping reads to the sequences of ENSEMBL (Version 74) human transcripts, on average 81.6% of reads passed quality control could be mapped to the transcriptome, while 96.5% of reads can be mapped to

the human genome (Table 5.2). These mapped reads are equivalent to an average 47 fold of coverage to the human transcriptome, when considering the length of transcriptome is at approximately 80 million base pairs (calculated from the transcripts recorded in ENSEMBL Version 74). Table 5.2 also shows that around 15% of reads that mapped to the genome aligned to introns and intragenic regions.

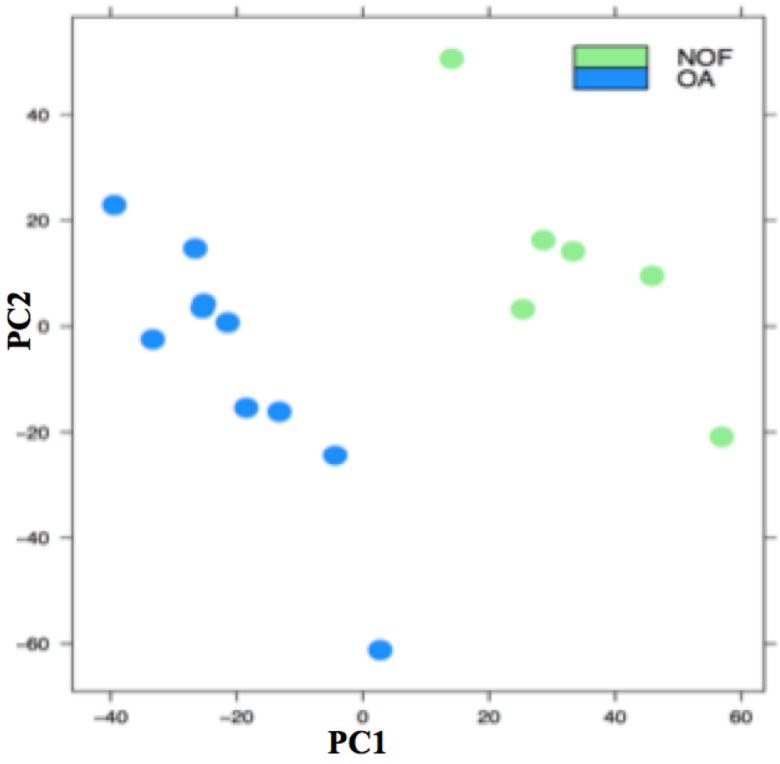
Sample	Type	Number of Reads before QC (million Pair)	Number of Reads After QC (million Pair)	Reads Passed QC (%)	Reads at least 70bp (%)	Total Base after QC (billion bases)	Mapped to Transcripto -me (%)	Mapped to the Genome (%)	Transcri- ptome Depth (Fold)	Equivalent Genome Depth (Fold)
N2080	OA	28.9	27.6	95.4	78.0	4.0	84.8	96.7	49.5	1.3
N1947	OA	29.5	26.4	89.5	68.7	3.7	83.2	96.8	45.7	1.2
N2049	OA	22.2	20.9	94.1	68.3	2.9	81.2	96.5	36.0	1.0
N2209	OA	29.0	27.8	95.9	73.7	3.9	86.2	96.3	49.1	1.3
N2004	OA	28.5	27.2	95.6	77.1	3.9	81.1	96.2	48.8	1.3
N1873	OA	26.7	25.1	94.2	79.4	3.6	80.7	96.4	45.2	1.2
N2062	OA	30.3	29.0	95.7	70.9	4.1	78.0	96.5	50.6	1.4
N2112	OA	28.7	27.5	95.5	70.9	3.8	76.7	96.1	47.8	1.3
N1901	OA	29.0	27.9	96.1	75.6	4.0	78.6	96.3	49.6	1.3
N1866	OA	26.8	25.2	94.0	76.0	3.6	81.3	96.9	44.9	1.2
N2002	NOF	30.6	29.1	94.9	73.0	4.1	83.6	96.2	51.2	1.4
N2024	NOF	26.7	24.8	92.8	72.2	3.5	83.0	96.7	43.5	1.2
N2060	NOF	28.7	27.6	96.2	75.1	3.9	80.7	96.4	49.0	1.3
N2064	NOF	30.8	29.4	95.6	72.4	4.1	76.2	95.7	51.6	1.4
N2059	NOF	27.7	26.2	94.8	73.4	3.7	82.3	96.4	46.2	1.2
N2120	NOF	29.3	28.1	96.0	73.8	4.0	87.2	96.1	49.6	1.3
Mean		28.3	26.9	94.8	73.7	3.8	81.6	96.4	47.4	1.3

Table 5.2: Mapping statistics of reads. The table presents several mapping statistics of the reads for each sample and their mean values of all samples. Around 95% of these reads passed our quality control filtering. On average, ~96% of reads can be mapped to the human genome while ~81% can be mapped to the transcriptome. These reads are equivalent to more than 47 fold coverage of the whole transcriptome.

Mapped reads to the whole genome were counted for each gene and then normalized to samples library sizes and GC content of genes in order to reveal whether expression profiles of genes can define the differences between OA and NOF samples. 24838 genes that had RPKM $\geq$ 0.3 in 80% of samples in either OA or NOF samples were considered as genes expressed in human hip cartilage. As expected, the hierarchical clustering of cartilage samples based on expression of expressed genes and the Principal component analysis (PCA) plots showed perfect separation of gene expression profiles of the OA and NOF samples (Fig. 5.3 A and B).



A



B

Figure 5.4 Distances between the samples. A. The heat map and the hierarchical clustering plot based on the expression profiles of the samples. **B.** The PCA plot of the samples based on the expression profiles of expressed genes. Both figures confirmed that the expression profiles of the samples could separate the OA and NOF samples into two groups. This confirmed of reliability of our expression profiles from the RNAseq data.

5.4.3 Differentially expressed genes in OA cartilage

With the raw counts data, the threshold of 1 expressed copy of a gene in 80% of either group of samples was considered as expressed. In total, 14,507 genes were found expressed in the cartilage samples. This number is smaller than the number of genes when using RPKM of 0.3 for filtering. In total, 1028 genes were identified as significantly differentially expressed with an adjusted P-value ≤ 0.05 and at least 2-fold change (Additional Table A5.1). 402 genes were found up regulated in the OA samples and 626 genes were down regulated. These include 812 protein coding genes, 72 long non-coding RNAs (lincRNA) and 144 other types of RNAs (miRNA, rRNA, snRNA, pseudo-gene etc) (Figure 5.5). It was noticed that all of the differentially expressed small RNAs are located on genomic repeated region.

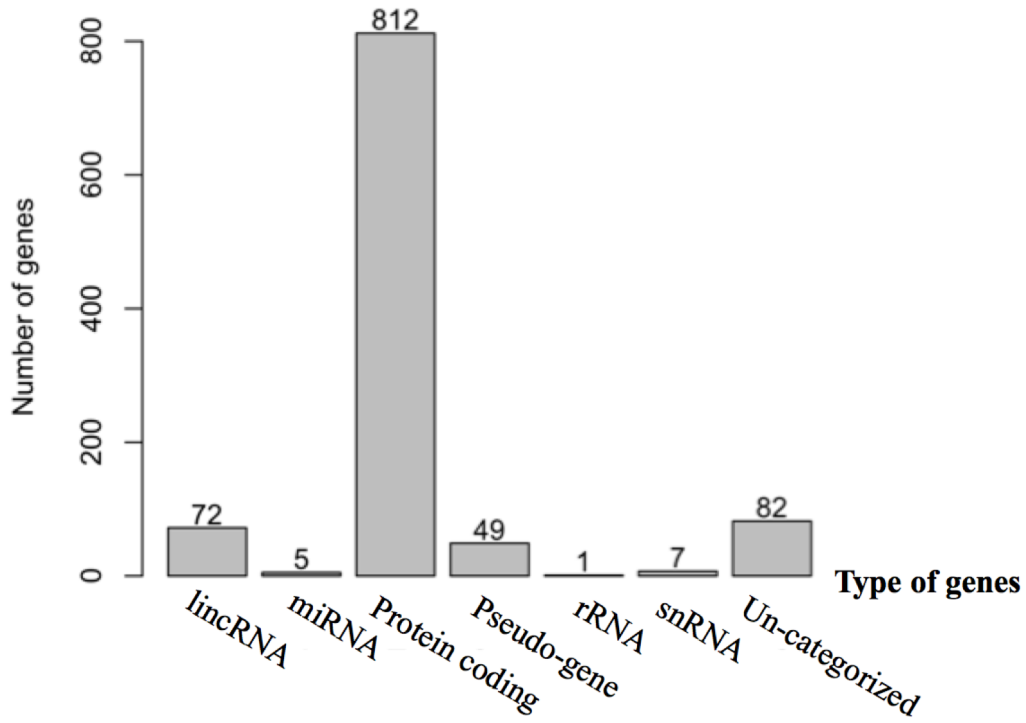


Figure 5.5 The composition of differentially expressed genes. The figure shows the number of differentially expressed genes of each type recorded in the GENCODE database.

Among the top 30 up-regulated genes and 30 down-regulated genes, down-regulated genes have generally larger fold changes. In fact, within the top 50 genes with highest fold changes, 45 of them are down regulated in OA cartilage. The expression of the

most down regulated gene *CCL20* in OA cartilage is only 1/60 in NOF cartilage samples, which is consistent with the observation of the microarray results. The expressions of matrix metalloproteinase 1 and 3 (*MMP1, MMP3*) and the heme oxygenase *HMOX1* in OA is less than 1/19 of their expressions in NOF samples. The serum amyloid A1 (*SAA1*) are among the most down regulated genes with a 25-fold change, but the variation of how many reads mapped to the gene in normal cartilage tissue is relatively large, ranging from 22 to 573. Bone morphogenetic protein-2 (*BMP2*) has almost 14 fold less expression in the OA samples with a reliable p-value (adjusted P-value = 7.43e-66), which again confirmed this observation from the microarray data, although it is contradict with other studies suggested (Fukui *et al.*, 2003).

In the up-regulated genes, MAM domain containing 2 (*MAMDC2*) has the largest fold change (>11 fold) in OA cartilage, while the function of the gene is unknown and it was not associated with OA before. A gene encoding voltage-gated potassium channel subunit, *KCNS1*, was found having almost 10 times more expression in OA cartilage samples, interestingly this gene has been reported to be associated with neuropathic pain (Costigan *et al.*, 2010). Significant up regulated expression of the gene encoding asporin (*ASPN*), type II collagen alpha I (*COL2A1*) and growth differentiation factor 10 (*GDF10*) were observed, as well as another matrix metalloproteinase *MMP16*, which was not identified from the microarray data.

Gene name	Gene type	Fold Change (OA/NOF)	Adjusted P-value	OA mean count	NOF mean count
CCL20	protein coding	-59.3	3.54E-25	5	1075
MMP1	protein coding	-29.0	3.94E-18	3	235
C5orf27	protein coding	-25.7	2.26E-51	35	1106
SAA1	protein coding	-25.0	5.14E-16	3	171
HMOX1	protein coding	-22.8	1.24E-97	192	4771
TNP1	protein coding	-21.6	3.77E-23	3	97
WTAPP1	pseudogene	-19.6	2.07E-39	10	228
MMP3	protein coding	-19.5	1.09E-46	3914	90144
RND1	protein coding	-18.8	3.77E-51	56	1215
NOS2	protein coding	-18.3	1.20E-36	291	6542
NOD2	protein coding	-18.2	1.21E-44	25	539
AC008592.8	lincRNA	-17.9	3.22E-17	6	184
G0S2	protein coding	-16.7	1.75E-20	117	2801
LIF	protein coding	-16.0	1.50E-12	29	927
PTGS2	protein coding	-15.7	5.73E-37	135	2516
ATF3	protein coding	-15.5	8.91E-31	31	583
BIRC3	protein coding	-14.3	6.45E-30	25	434
MYBPH	protein coding	-14.3	6.08E-23	38	697
BMP2	protein coding	-13.6	7.43E-66	453	6658
C11orf96	protein coding	-13.5	1.56E-28	1148	18649
LYVE1	protein coding	-12.6	4.93E-22	67	1056
TLR2	protein coding	-11.9	4.65E-107	102	1254
ORM1	protein coding	-11.6	1.82E-18	4	64
NR4A3	protein coding	-11.4	2.00E-30	22	286
SOD2	protein coding	-11.3	3.66E-15	4705	104828
GLRX	protein coding	-10.9	9.43E-23	240	3509
PTX3	protein coding	-9.8	2.37E-15	20	254
CCDC71L	protein coding	-9.6	1.94E-51	38	392
LCN2	protein coding	-9.6	1.91E-17	43	510
STEAP4	protein coding	-9.3	1.79E-11	428	5562
TSPAN11	protein coding	4.5	2.85E-11	395	80
AL645608.1	protein coding	4.5	3.20E-06	11	2
MMP16	protein coding	4.5	2.26E-07	294	55
SPTSSB	protein coding	4.5	7.58E-14	677	138
RP11-300E4.2	antisense	4.5	6.39E-05	20	3
RP1-39J2.1	lincRNA	4.6	1.82E-06	19	3
PART1	lincRNA	4.7	6.53E-20	595	120
NREP	protein coding	4.8	1.99E-21	326	65
MXRA5	protein coding	4.8	2.59E-09	3026	550
RNU5A-1	snRNA	5.0	0.000791459	6	0
TNNI2	protein coding	5.0	0.000219433	153	18
IFITM10	protein coding	5.1	4.28E-22	3865	723
RP11-231C14.5	pseudogene	5.2	5.50E-05	18	2
SERTAD4-AS1	antisense	5.2	3.30E-26	331	60
ZCCHC5	protein coding	5.2	2.75E-15	123	22
CAPN6	protein coding	5.3	6.07E-06	475	63
RP11-460I19.2	lincRNA	5.4	0.000120061	5	1
NCAM1	protein coding	5.7	3.23E-21	435	71
CTHRC1	protein coding	5.8	3.19E-10	1176	168
NFATC2	protein coding	6.0	5.43E-08	1516	194
SYT8	protein coding	6.0	4.87E-08	499	63
CRISPLD1	protein coding	6.3	1.75E-12	7544	1014
TPPP3	protein coding	6.4	5.67E-07	434	46
CCDC129	protein coding	6.7	2.29E-07	256	25
GDF10	protein coding	6.8	2.26E-23	15671	2104
COL2A1	protein coding	7.4	1.66E-08	679560	61785
SERTAD4	protein coding	7.8	2.07E-26	1829	214
ASPN	protein coding	8.1	1.15E-32	3122	354
KCNS1	protein coding	9.9	1.37E-18	187	15
MAMDC2	protein coding	11.0	1.73E-22	179	13

Table 5.3 Top 30 up and down regulated genes. The table shows the top 30 genes with highest fold change in down regulated genes and up regulated genes respectively. The down-regulated genes have generally smaller fold change than up regulated genes.

Over 200 GO terms were found significantly enriched in down regulated genes and around 110 terms for up regulated genes. (Additional Table A5.2 & A5.3) These terms were separated into the 3 main categories of GO terms: Biological Process (BP), Cellular Components (CC) and Molecular Functions (MF). Both up and down regulated genes are involved in various biological processes, such as cell proliferation, metabolic process, cell mobility, regulation of transcription from RNA polymerase II promoter and intracellular protein kinase cascade, and other 35 BP terms. Down regulated genes also showed strong association with cell death, inflammatory response, immune response, nitric oxide biosynthetic process and Estrogen Receptor-nucleus signaling pathway, while up regulated genes were found involved in several different biological processes, including: extracellular structure organization, prenylcysteine catabolic process, glycoprotein metabolic process, sulfur compound metabolic process and regulation of sequence-specific DNA binding transcription factor activity. In terms of CC, only 3 terms are specific to up regulated genes only, which are proteinaceous extracellular matrix, integral to plasma membrane and Golgi apparatus part. Extracellular region, extracellular space and extracellular matrix terms were found associated with both sets of genes as expected. Down regulated genes were also associated with cytosol, nucleus and I-kappaB/NF-kappaB complex. MF terms showed that up regulated genes have involved in extracellular matrix structural constituent, oxysterol 7-alpha-hydroxylase activity, Wnt-activated receptor activity and arylsulfatase activity, while down regulated genes are involved in 61 other molecular level activities. Polysaccharide binding, ion binding and sequence-specific DNA binding were found associated with both sets of differentially expressed genes.

5.4.4 Pathways of differentially expressed genes

The differentially expressed genes identified were uploaded to the Ingenuity Pathway Analysis to investigate their associated pathways. 81 canonical pathways were found to be significantly (P value ≤ 0.05) associated with the genes. Several of them are known to be associated with osteoarthritis, including PI3K/AKT Signalling, HIF1α Signalling, LXR/RXR Activation, Inhibition of Matrix Metalloproteinases and G-Protein Coupled Receptor Signalling (Woessner, 1991; Collins-Racie *et al.*, 2009; Kerkhof *et al.*, 2010) (Additional Table A5.4). A number of pathways containing genes that were previously shown to be differentially expressed in OA were also found associated with the gene set,

such as Sonic Hedgehog Signalling, VDR/RXR Activation, Estrogen-mediated S-phase Entry, Oncostatin M Signalling, Role of JAK family kinases in IL-6-type Cytokine Signalling and BMP signalling pathway, etc. (Hopwood *et al.*, 2007; Kapoor *et al.*, 2011; Beekhuizen *et al.*, 2013) Twelve cancer pathways also showed significant association largely based on several common genes contained in all of these pathways, including *MYC, CDKN1A, ABL1, TFDP1, NFKB2, FZD1, FZD2* and *FZD8.*

Almost half of the pathways (38/81) involves *RELA* and *NFKB2*, both genes were significantly down regulated in the OA samples (Table 5.4). *MMP1* and *MMP3* were also found to be more frequently involved in the associated pathways. There were other 17 genes listed in Table 5.4, all of which were involved in more than 10 associated pathways.

Gene Name	Frequency in the associated pathways
RELA	38
NFKB2	34
NFKBIE	22
PDGFC	17
VEGFA	15
PRKACB	14
MMP1	14
CDKN1A	14
PTGS2	13
TNFRSF1B	13
PRKAR2B	13
ABL1	13
MYC	13
MMP3	12
NOS2	12
CCL2	11
TFDP1	11
UTGA5	10
RHOJ	10
PDGFA	10
RHOG	10
RND3	10

Table 5.4 Differentially expressed genes involved in 10 or more associated pathways. The table shows the genes that are involved more than 10 OA associated pathways identified in our analysis. *RELA* and *NFKB2* are involved in almost half of the associated pathways (38/81). *MMP1* and *MMP3* were found to be involved in more than 12 pathways.

5.4.5 Differentially expressed transcripts

BitSeq was used to identify differentially expressed transcripts. The software does not require the assembling sequences of transcripts before testing differentially expression. This saves computing time but limits the analysis to the known transcripts within the available annotation. In total, 4352 transcripts were significantly differentially expressed between the NOF and OA patients. They represent 2488 different genes, 928 of which were also found as differentially expressed genes. (Additional Table A5.5) The transcripts encoding collagenases, including *COL1A2, COL2A1, COL3A1, COL5A1, COL5A2, COL8A2, COL9A1, COL9A2, COL9A3, COL11A1, COL11A2, COL16A1* and *COL27A1*, were up regulated at transcript level in OA patients, while *COL2A1*, *COL5A1, COL5A2* and *COL11A1* were also found up regulated in OA on gene level. Most of transcripts of *SOD2* were down regulated in OA. Comparing the protein-coding genes that were identified as differentially expressed genes to the transcripts, over 90% of the protein coding genes (745/812 genes) have differentially expressed transcripts, it is consistent as expected, since the transcript expression make up an observation of gene expression. In contrast, ~30% of genes that have differentially expressed transcripts were not detected as differentially expressed genes in the previous analysis.

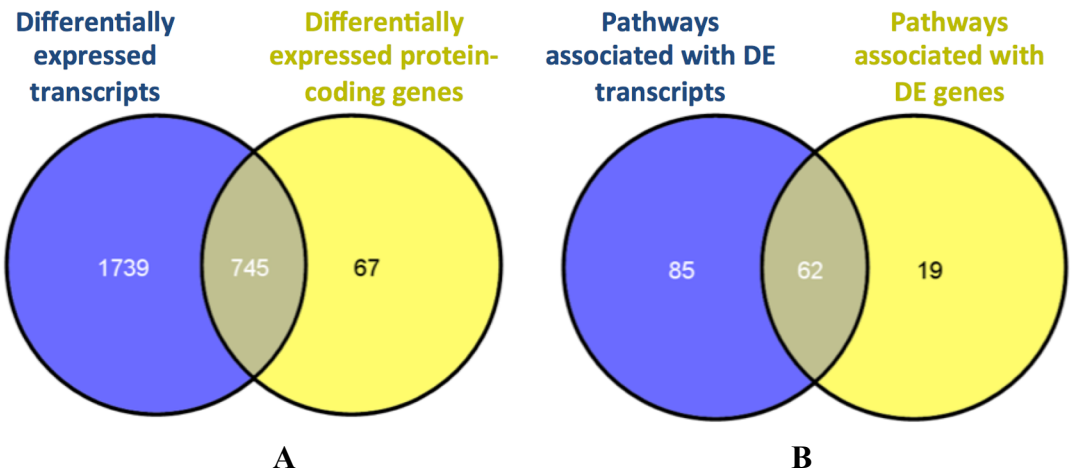


Figure 5.6 Comparison of differentially expressed transcripts and protein-coding genes. A): The Venn diagram shows the comparison between genes that have differentially expressed transcripts and differentially expressed protein-coding genes identified using DESeq. Over 90% of the genes have differentially expressed transcripts, while only ~30% of the differentially expressed transcripts have corresponding genes identified as differentially expressed. **B):** The Venn diagram shows the comparison between the pathways associated with the differentially expressed transcripts and the differentially expressed genes. The percentage of the overlapped pathways is 42% of the all pathways associated with the transcripts, which is greater than the percentage of the common genes in A.

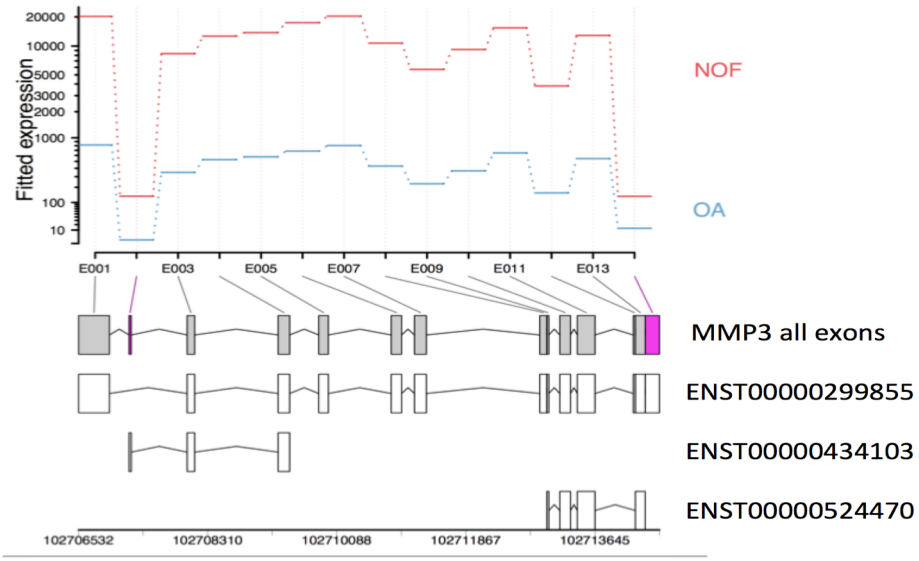
Pathways analysis using the same method as for differentially expressed genes showed that 147 pathways were significantly (P value ≤ 0.05) associated with the differentially expressed transcripts. (Additional Table A5.6) Comparing to the pathways that associated with differentially expressed genes, 85 pathways are unique to the transcripts sets (Figure 5.6B), including Inhibition of Angiogenesis by TSP1 and Wnt/β-catenin Signaling, which are known to be associated with OA (Giatromanolaki *et al.*, 2001b; Velasco *et al.*, 2010a). 62 out of 81 pathways associated with the differentially expressed genes were also found associated with the transcripts data set.

5.4.6 Alternative splicing events in OA cartilage

Exon usages were tested with DEXSeq, significantly differentially used exons (corrected P value ≤ 0.05 and fold change of the usage ≥ 2) were identified. The result reflects the alternative splicing events in OA comparing to the NOF samples. In total, 467 exons of 262 genes showed more usage in OA samples and 431 exons of 281 genes showed less usage. (Additional Table A5.7) 27 genes were found in both of the lists,

amongst which *SOD2, MMP3* and *ADAMTS4* were found. *SOD2* and *MMP3* are among the most down regulated genes. *SOD2* has 14 exons that have more usage in OA samples. However, considering the large fold change of down regulation of the gene and the number of total known exons of the gene (42 exons in ENSEMBL 74), the over usage of the exons could be caused by absent expression of the other exons. MMP3 has one exon with over usage in OA and another exon with less usage comparing to NOF samples (Figure 5.7 A). As there are only three known transcripts of the gene and the two exons belong to two different ones, this implies the composition of the transcripts in OA are different from NOF cartilage samples and transcript "ENST00000478394" could comprise less of the total expression of the gene in the OA samples. *ADAMTS4* have known alternatively spliced transcripts identified in the synovium of OA patients (Wainwright *et al.*, 2006). The alternative spliced transcript found in the study had a shorter exon 9 comparing to the longest transcript of the gene. In OA cartilage, we found the exon usage differences of two exons of the gene comparing to the NOF samples, the same change of the exon 9 (exon 5 in the Figure 5.7 B, as DEXSeq counts exons without the respect of the transcription direction) was identified.

A



B

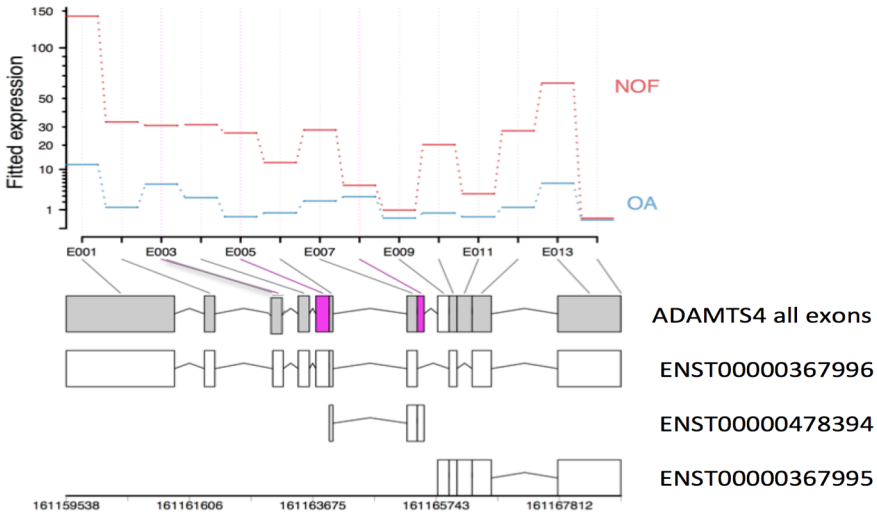


Figure 5.7 Alternative splicing isoforms of MMP3 and ADAMTS4. The alternative spliced exons of MMP3 and ADAMTS4 are shown in the figure A and B respectively. Normalized counts for each exon are shown with blue bars for OA samples and red bars for NOF samples. All known exons are listed with respect to their relative genomic positions. DEXSeq ignores the transcriptions direction of the transcripts and always count exons according to their coordinates on the forward strand. The known transcripts in ENSEMBL 74 are listed below. Exons whose usages were significantly changed (log2 fold change (OA/NOF) ≥ 1 or ≤ -1 with adjusted P value ≤ 0.05) were highlighted. **A):** The usage of the second exon of MMP3 is down changed in OA with log2 fold change (OA/NOF) = -3 and adjusted P value = 4.25e-4. The last exon is up changed with log2 fold change (OA/NOF) = 1.2 and adjusted P value = 2.55e-8. This indicates the isoform composition changed in OA and possibly the second transcript comprise more expression of the gene in OA. **B):** The usage of the eighth exon of ADAMPTS4 up changed in the OA samples with log2 fold change (OA/NOF) = 2.7 and adjusted P value = 2.78e-3, while the fifth exon down changed in the OA samples with log2 fold change (OA/NOF) = 2.43 and adjusted P value = 1.90e-2, which indicates the longest transcript comprise less of the total expression of the gene in OA samples. The change of the exon 5 is consistent with the splicing form reported in (Wainwright *et al.*, 2006).

DiffSplice uses a different approach to detect splicing patterns from DEXSeq and focuses on the coverage of the exon junctions. By comparing the normalized coverage, alternative splicing events can be identified. In order to test whether the approach can lead to novel findings, our RNAseq data was also analyzed using the software. In total 249 alternative splicing events of 196 genes that have significantly (False discovery rate ≤ 0.05 and more than two-fold change) different frequencies between OA and NOF samples. These events include 103 intron retention events, 32 alternative transcription starting or ending sites, 6 exon skipping events and 108 un-catalogued events. (Additional Table A5.8) Comparing the lists of genes that have alternative splicing events identified with the two software tools, only 26 genes were found in both datasets.

5.4.7 Transcripts expressed only in OA/NOF and novel transcripts

Transcripts expressed in cartilages were assembled using Cufflinks, with respect of the known gene annotation available in ENSEMBL. Comparing the assembled transcripts from the NOF and OA samples, there are 680 transcripts expressed only in OA group (cufflinks estimated coverage ≥ 1 in at least 80% of OA samples) and 1468 transcripts expressed only in NOF group (cufflinks estimated coverage ≥ 1 in at least 80% of NOF samples) (Additional Table A5.9 and A5.10). They represent 674 and 1435 genes in the OA and NOF samples respectively. Within these 15 genes have different transcripts unique to each condition, in other words, they have alternative spliced isoforms expressed in the OA samples compared to the NOF cartilage samples. One of the genes *SDC1* was previously reported up regulated in OA (Salminen-Mankonen *et al.*, 2005).

Novel transcripts of the cartilage samples were obtained by comparing all of the assembled transcripts to the transcripts in ENSEMBL. Transcripts that have novel sequences with acceptable abundances (cufflinks estimated coverage ≥ 1 in at least 80% of either OA or NOF samples) for NOF and OA patients were selected as novel transcripts, which are shown in the Additional Table A5.11 and A5.12. In total 56 transcripts of 55 known genes in OA samples and 151 transcripts of different known genes in NOF samples were identified as novel. These sequences were recorded in a text file in GTF format (Supplementary file S5.2).

5.4.8 Variants detected and Allelic expression analysis

Samtools was used to identify variants on the RNA level in the cartilage samples. On average, around 232,500 variants were detected for each sample. After selection of heterozygous locus that commonly exists in samples (heterozygous variants observed in at least 3 OA and 3 NOF samples), 12352 variants were used in the allelic expression analysis. The result showed that allele ratios on 114 loci of 85 genes are significantly different (P values ≤ 0.05) between OA and NOF samples (Additional Table A5.13). Only two genes, *ALPK3* and *SYN1*, have different allelic expressions in OA and also significantly down regulated in the OA cartilage. Six genes have more than 3 loci that have different allele imbalanced status, including *TRPV4, CDC27, CPSF3L, EDARADD, GPI* and *RP11-262H14.1*, the latter of which is a long non-coding RNA with no reported function. There is evidence showing that *TRPV4* may have a role in maintenance of the joint. (Clark *et al.*, 2010)

5.4.9 RNA-editing in OA cartilage comparing to NOF analysis

RNA-editing events are commonly observed in mammalians and are possible to detect only with RNAseq data (Levanon *et al.*, 2004; Danecek *et al.*, 2012). We used our RNAseq data to identify possible editing events and to test whether an event is associated with OA. Samtools was used to call variants from the RNAseq data and select heterozygous sequencing variants from the RNAseq data, then the number of reads supporting each base in the heterozygous variants was counted as the frequencies of an RNA editing event in the samples. A locus where a heterozygous variant with a base change consistent with the main form of RNA-editing changes (A-to-I, translated as A-to-G) in any of the 16 samples was selected. After quality filtering of the variants, the base counts for the filtered loci were tested for the differential editing frequencies between the OA and NOF samples using a generalized linear model in R. In total 128,067 loci where an A to G change was identified in at least one sample were used for the test. But no editing event was found to be associated with the disease.

5.4.10 Validation

Comparison of differentially expressed genes identified from this RNAseq study and the microarray study was made in Chapter 6. The strong and significant correlation (P-value < 2.2-e16 and $r = 0.74$) between the two data sets validated our expression profiles derived from the RNAseq.

5.5 Discussion

In this chapter, we demonstrated the application of RNAseq technology in comparing the OA cartilage transcriptome with that of NOF cartilage. The advantages of RNAseq compared to conventional expression profiling technologies are obvious: the high-throughput capability and ability to reveal the quantification, the structure and the sequence of transcripts in a single base resolution, which enable its application in the detection of novel gene isoforms. Furthermore, expression profiling using the microarray technology is only restricted to an organism with a known reference genome. The detailed comparison of the two technologies in terms of their ability in gene expression profiling is described in the next Chapter, which showed RNAseq identified two times more of differentially expressed genes than the microarray. In addition, using RNAseq we identified several characteristics of the molecular changes in OA cartilage comparing to the NOF cartilage, including differentially expressed genes, differentially expressed transcripts, alternative splicing events, novel transcripts and allelic expression in the OA transcriptomes. RNA editing events were identified but none of the events were found to be associated with the disease.

5.5.1 Findings about OA with the RNAseq data

With the RNAseq data, over a thousand genes were found differentially expressed. These include several genes that were previously associated with OA but were not detected in microarray experiment, such as *ADAMTS2* and *MMP16*. Up-regulation of *BMP2* was observed, which confirmed the similar finding from the microarray data, although the change is contradicted with other studies (Fukui *et al.*, 2003; Nakase *et al.*, 2003). These may suggest the different mechanism of hip OA from the other types of OA.

Several pathways that previously associated with OA were also identified, these include: 'iNOS Signaling' (Cheng *et al.*, 2011), 'Acute Phase Response Signaling' (Sipe, 1995), 'PI3K/AKT Signaling' (Huang *et al.*, 2011a), 'LXR/RXR Activation' (Collins-Racie *et al.*, 2009) and 'HIF1α Signaling'(Giatromanolaki *et al.*, 2003b). In the associated pathways, 22 genes were involved in more pathways than other genes, especially *RELA* and *NFKB2* were involved in almost half of the associated pathways (38/81). It indicates that the associated pathways were biased, as some of the

differentially expressed genes were better studied than the other genes. However, these genes could still be potentially interesting targets for studying the mechanism of the cartilage destruction in OA, as they play roles in a number of biological processes and their differentially expression could lead to significant changes to the metabolic status of chondrocytes. In addition to the differentially expressed genes, the expressions of the transcripts were estimated, their sequences and structures were also identified, and differentially expressed transcripts were also identified. All this information constructs the whole transcriptome of the OA and NOF cartilage samples. When comparing the transcriptomes, alternatively spliced transcripts were identified as well. Both *MMP3* and *ADAMTS4* were found significantly down regulated and also alternative spliced in OA, suggesting the down regulation of the genes are correlated with the transcripts composition changes. The previously reported splicing form of *AMDATS4* in synovium in OA patients was also identified in cartilage samples, suggesting the types of tissue have the same regulation mechanism in OA. While identifying genome-wide differential allelic gene expression, six loci of *TRPV4* showed differentially imbalanced status. There is evidence showing that *TRPV4* may have a roll in maintenance of joint health (Clark *et al.*, 2010). The imbalanced allelic expression of the gene may contribute to the OA susceptibility.

5.5.2 Issues related to the sequencing depth

Sequencing depth is an important parameter when designing an RNAseq study. The ability of detecting expressions of transcripts in an RNAseq experiment is dependent on the number of reads that can be mapped to the transcripts, while lowly expressed transcripts have less chance to be sequenced, so some genes may not be detected if insufficient sequencing depth is specified. The sequencing depth in our study is in line with the RNAseq guideline of the Encyclopedia of DNA Elements (ENCODE) Consortium (ENCODE, 2009) for expression profiling studies, but only half of the amount of the reads recommended for reliable detection of the relatively low expressed transcripts. In our RNAseq experiments, the sequencing depth achieved for the transcriptome is around 47 for each sample. With this depth, ~ 25,000 genes were found with RPKM $\geq$ 0.3 in 80% of either OA or NOF cartilage samples. In fact, a single read mapped to a transcript indicates the expression of the transcript, while RPKM is an ambiguous expression measurement of gene expressions. The threshold that I used in

the analysis is only a reflection of the confidence of the detection. The RPKM of 0.3 was suggested as a threshold when RNAseq was introduced in 2008 (Mortazavi *et al.*, 2008) and was followed in several other studies (Labaj *et al.*, 2011; Sam *et al.*, 2011), thus it was chosen in this study to filter genes before hierarchically clustering of samples. The RPKM were calculated after normalization with consideration of GC contents bias and gene length bias (Hansen *et al.*, 2012). However, genes with low counts can create difficulties in estimation and modeling gene expressions when identifying differentially expressed genes in RNAseq studies (Bullard *et al.*, 2010). Thus the expressed copy number of genes was calculated with the raw count data and only genes have at least 1 expressed copy in samples were considered as expressed, as this is the minimum requirement to indicate that a gene was sequenced at least once. In fact, with the read length and the average RNAseq library size (number of total reads) of our data, RPKM of 0.3 equates to ~1 expressed copy. However, when using 1 expressed copy as a threshold only ~14,500 genes were considered as expressed. The 10,000 genes difference can only be from the normalization, which implies their ambiguous detections in the samples and suggests insufficient sequencing depth in at least some of the samples.

The splicing events detected are also dependent on the sequencing depth. Because not all of the sequencing reads are from the junctions of exons, correct assembly of a transcript will require coverage of more than just 1 expressed copy. Our RNAseq data does not have the sufficient coverage recommended by the ENCODE to assemble transcriptomes. However, DEXSeq and DiffSplice were used to detect alternative splicing events without assembling the transcripts. The two algorithms are dependent on sufficient coverage of exons and exon junctions respectively in both comparing conditions, in our case OA versus NOF. But it can be problematic when the coverage in any of conditions is low. In my DEXSeq results, 14 exons of the gene *SOD2* were identified as alternatively spliced, which is likely the result of very low coverage of the exons in OA samples. Successful splicing event detections are not only dependent upon the sequencing depth but also several other aspects, such as the gene expression level in the tissue, complexity of the transcript structure and the performance of the software tool, which currently remains as a bottleneck in the assembly of the transcripts and alternative splicing events detection (Schliesky *et al.*, 2012).

5.5.3 Duplicates removal

In RNAseq reads, it is possible to observe duplicated reads due to PCR artifacts and optical duplication, which are from the same cluster on a flow cell but identified as from adjacent clusters. However, the necessity of removing these is still an open discussion in bioinformatics communities, as highly expressed genes could result in duplicated reads as well and these reads cannot be differentiated from the duplicated reads of other sources. Because this study was focused on the comparisons between OA and NOF RNA libraries, the noise of the duplicates should not affect our analysis. But when the RNAseq data was used to call variants on the mRNA level, the accuracy of the variants can be affected by the duplicated reads. This could lead to false positives even after applying filters on the variant call qualities and the coverage. In recent studies (Bahn *et al.*, 2012; Chen and Bundschuh, 2012) for variant calling of RNAseq data, the duplicates were recommended to be removed for accurate variant detection.

5.5.4 Aligners for RNAseq

In my analysis, Novoalign, GSNAP and Tophat were used to map reads to the reference genome and the reference transcriptome. Novoalign had better sensitivity (Chen and Bundschuh, 2012) but limited support of mapping RNAseq reads, thus it was used to map reads to the transcriptome. GSNAP can support the use of known SNPs and Tophat performs a two round alignment strategy with the support of the use of known transcript annotation, these features were particular useful for the RNAseq data mapping and unique to the software at the time, thus both of them were chosen to map reads to the reference genome. When using known SNPs, more reads can be mapped as mis-match on the SNP sites are tolerated. This can be helpful when assembly transcripts and identify alternative splicing events because more mapped reads can provide more evidences to support exon junctions and assembled transcripts. But the mapping accuracy is affected by the reliability of the SNPs as well, and it can be more problematic when identifying variants at the RNA level, because un-reliable SNPs can cause GSNAP to incorrectly score mis-matches and results false positive variants when calling variants from the alignment. Therefore, Tophat was used for the alignments, from which sequencing variants on the RNA level were called. As for the alignments for gene expression profiling, in my opinion, the choice of the software has little impact on

the results of differentially expressed genes, because all of the aligners treat the reads no differently between samples.

The mapping algorithm for RNAseq reads is a constant interest of the field in the recent years and several aligners were published over the years, such as STAR (Dobin *et al.*, 2013) and SMALT from the Sanger Institute. Comparison studies of the aligners (Engstrom *et al.*, 2013; Hatem *et al.*, 2013) showed their different strengths and weakness, and no single aligner outperformed the other ones. However, this gave end-users the freedom to choose the right aligner for their different needs.

5.5.5 The analysis of differentially expressed genes

Within the 14,507 genes found expressed in the cartilage samples, the majority of them are protein-coding genes with the expression of pseudo-genes, miRNAs, lincRNA and snRNAs were also observed, a number of which were also found as differentially expressed. However, the reads mapped to the snRNAs could be incorrectly mapped. As most of these genes are located in repetitive regions, the mapping could be incorrect. The read counts of them are therefore not as reliable as other type of genes. Compared to the small RNAs, the low read counts of those differentially expressed miRNAs are more reliable, plus there are no known exon overlapped with these genes, therefore the differences of their expression between OA and NOF samples could be real. But miRNAs are only ~22nt long and should have been lost in the sequencing library preparation, so the detection of miRNAs in my data could be pri-miRNA or pre-miRNAs per se. The software tool DESeq was used in this study to call differentially expressed genes. In a comparison of the software with other tools (Seyednasrollah *et al.*, 2013), DESeq detected more differentially expressed genes than other software tools. It also showed lowest false positive rate comparing to the other tools when number of biological replicates increased. In our study, it identified more than a thousand differentially expressed genes, which is twice of the genes identified in our microarray experiment when using the same fold change and p-value cut-offs (see in Chapter 6.3 Results).

5.5.6 The analysis of differentially expressed transcripts and splicing events

One of the advantages of RNAseq is that it can be used to assemble the sequences of the transcriptome, which is all of the expressed RNAs in the cartilage samples for our data. While in my analysis, de novo assembly was restricted because of the two factors: 1) none of our samples achieved the recommended sequencing depth suggested in the ENCODE best practice; 2) performing de novo assembly of the human transcriptome requires large amount of memory (up to 80Gb (Illumina, 2009)), which was not supported by our local set up at the time of this study. Therefore, Cufflinks was used to assemble transcripts, as it utilizes the existing reference genome and annotations to assemble transcripts. Consequently, it uses less memory and is easier to compare with the existing gene annotations. Compared to the de novo assembly strategies, Cufflinks performs better in several respects, including sensitivity, specificity and number of assembled full-length transcripts, but with increasing sequencing depth the difference has become smaller (Schulz *et al.*, 2012). Cuffdiff was developed by the same authors of Cufflinks and was designed to call differentially expressed genes, transcripts and alternative splicing events from the assembled transcriptome of Cufflinks. But in my analysis, Cuffdiff was used but did not report any significant hits of differentially expressed transcripts or alternative splicing events. However, when using BitSeq, DEXSeq and DiffSplice, considerable numbers of differentially expressed transcripts and alternative splicing events were identified. The overlap between the differentially expressed protein-coding genes and transcripts also verified each other. This suggests Cuffdiff did not work as expected. The cause of it is unclear, as the software was successfully used in many other studies (Wu *et al.*, 2012; Young *et al.*, 2012; Garzia *et al.*, 2013). Compared to Cufflinks, DEXSeq and DiffSplice use different algorithms and do not require assembled transcriptomes. The two tools also use different algorithms from Cufflinks to identify and test for alternative splicing events, which also explains that alternative splicing events of only 29 genes were commonly found in the results of the two tools. DEXSeq relies on the existing transcripts annotations, so it lacks the ability to detect novel splicing events. In the DEXSeq results of our RNAseq data, two exons of the *ADAMTS4* were found significantly differentially used between OA and NOF samples, indicating the transcript composition changes in OA. In fact, splicing event detection remains a challenge for RNAseq data. It is limited by the computational ability of current assembly software tools (Schliesky *et al.*, 2012), the quality of the

input RNAs, the targeted transcript lengths and also the GC contents of the transcripts (Rehrauer *et al.*, 2013).

5.5.7 Variant detection at the RNA level, allelic expression analysis and RNA-editing events identification

At the time of the study there were no published software or pipelines specifically designed to identify sequencing variants on RNA level. However, as the principle is the same as the identification of variants at the DNA level, Samtools was used to identify variants for our RNAseq data. Because insertions and deletions could result in incorrect mapping, in order to obtain accurate estimation of the coverage of the variants, only heterozygous single nucleotide variants were chosen to evaluate the allelic expressions of genes and frequencies of RNA-editing events. In my analysis, duplicated reads were not removed, which could lead to false positives in detection of variants but better accuracy in estimation of the coverage and the frequency of RNA-editing events. To my knowledge, there was no existing software at the time to test differentially allelic expression or differential frequency of RNA-editing events at the genome-widely, thus I implemented my own methods in Perl and R script for the tests. In fact, the RNA-editing events defined in the study were just changes that were consistent with the RNA-editing change (A to I). Unfortunately, there was no DNA data to verify the identifications.

Chapter 6 Comparison of the two platforms: RNAseq and microarray

6.1 Introduction

As a new technology to quantify gene expression RNAseq has several advantages compared to the microarray, including its single-base resolution, lower background noise and larger dynamic detection range (Wang *et al.*, 2009), In addition the RNAseq data can potentially provide other information of a transcriptome, such as alternative splicing events, etc. With the decreasing price of the NGS, RNAseq has become more popular with a trend for replacing microarray in the last few years. However, several studies have shown that not all genes detected using the microarray were also detected in RNAseq data (Illumina, 2011). In order to compare the performance of the microarray technology and the RNAseq in terms of gene expression profiling, we compared the two data sets produced from the cartilage samples in this study to the recently published quantitative analysis of gene expression in cartilage using RT-PCR (Swingler *et al.*, 2009b), in which over 500 genes were profiled.

6.2 Methods

6.2.1 Sources of the data sets for comparison

The data sets were obtained from different sources. The RT-PCR data was downloaded from the supplementary files of the published manuscript (Swingler *et al.*, 2009b). Our microarray data was processed and probes were filtered as described in the Chapter 4. Only the probes that passed the filtering criteria were considered as expressed in the cartilage samples. As there were duplicated probes for a same gene on the microarray chip, only the probe that has the lowest corrected P-values among the duplicates was selected for comparison. Expressed genes in our RNAseq data were defined as genes

that have at least 1 copy in 80% of samples in either group of the cartilages (described in Chapter 5). Fold changes and corrected P-values of the genes in RNAseq data were calculated as described in Chapter 5 for gene expression analysis. Genes that have more than 2 fold expression changes in OA with P-values (or corrected P-values) ≤ 0.05 in each data set were defined as differentially expressed genes.

6.2.2 Identifier conversion

In order to compare the expression and fold changes of the genes, the gene identifiers used in each data set were converted to ENSEMBL gene identifiers using the online tool DAVID (Huang da *et al.*, 2009).

6.2.3 Statistical analyses

The Pearson's correlations of fold changes between the data set were calculated in R. The P-values used in the RT-PCR study (Swingler *et al.*, 2009a) were used in this comparison. Differentially expressed genes in the microarray and the RNAseq data were analyzed as described in Chapter 4 and 5.

6.3 Results

6.3.1 Detected genes in the data sets

The number of expressed genes that can be corrected detected is the most important quality criterion for gene expression profiling technologies, thus we only compared the expressed genes identified in the three datasets. Different numbers of genes were investigated and identified as expressed in each data set, owing to the different technologies, samples sizes and the quality filtering criteria. (see Table 6.1) In the microarray data, there were almost 48 thousand probes used, but ~20% are designed to target the same genes, thus in total ~38 thousand unique genes were profiled, amongst which over 33 thousand genes had valid identifiers. In total 13202 genes were selected as expressed in the microarray data. Restricted by the library preparation protocol, our RNAseq data set in theory had the ability to investigate every transcript with a poly-A tail, thus every mRNA can be considered as investigated genes. Among them, 14507 genes were considered as expressed in the cartilage samples. The RT-PCR data

investigated 551 genes with valid identifier and detected 427 genes expressed in the cartilage samples. Comparison of the detected/expressed genes in the data sets revealed discrepancies of the two high-throughput platforms (Figure 6.1). The microarray data detected 3474 genes that were not determined being expressed in the RNAseq data. Seventy-two of them were also detected within the RT-PCR data. Among the genes expressed in the RNAseq data, 4779 genes were not detected in the microarray data, 24 of which were detected with RT-PCR. There are 56 genes only detected with the RT-PCR. In terms of the differentially expressed genes, only a fraction of such genes in the RT-PCR data was also identified as differentially expressed using the other two technologies. However, the RT-PCR data validated more differentially expressed genes from the RNAseq data than the microarray data. Furthermore, comparing the number of differentially expressed genes identified in the two high-throughput data sets, more than half of the up-regulated genes (104/203) and 85% of the down regulated genes (172/200) in the microarray were identified in the RNAseq data set, while the microarray data only identified no more than 30% of differentially expressed genes (272/1028) in the RNAseq data.

Data set /Technology	Sample size	Criteria to select expressed genes	Total Genes	Genes with valid ID	Expressed genes	Differenti-ally expressed genes
RT-PCR	OA=12 NOF=12	Median Ct <40	569	551	427	117
Microarray	OA=9 NOF=10	Probes with valid flag value in ≥ 80% of either group of samples	37838 (48784 probes)	33425	13202	403
RNAseq	OA=10 NOF=6	Genes have more than 1 molecule detected in ≥80% of either group of samples	All mRNAs	55765	14507	1028

Table 6.1 Number of genes in the 3 datasets. The table presents the technology used to generate the data, the sample size, the criteria to select expressed genes, the total number of genes interrogated, number of genes that have identifiers can be converted to ENSEMBL identifiers, number of genes expressed and number of genes which expressions were significantly changed (Fold change ≥ 2 or ≤ -2, P-values ≤ 0.05) in OA cartilage samples comparing to NOF for each data set. As there are no primers or probes used for RNAseq technology, we consider every mRNA was interrogated in the analysis. With the same fold change and p-value thresholds, RNAseq identified more than twice of the differentially expressed genes identified using microarray, despite the similar number of genes found expressed in the cartilages using the both technology.

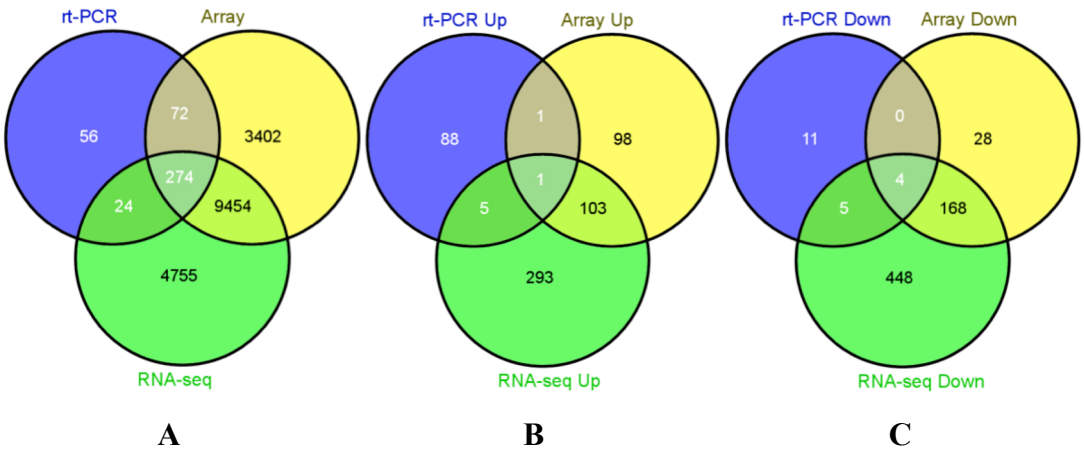


Figure 6.1: Comparison between detected genes. A. The Venn diagram of expressed genes in each data set. In the 427 expressed genes in the RT-PCR data, 56 of them were not found expressed in RNAseq or microarray data. 274 genes were found expressed in all of the 3 data sets. The overlapped expressed genes between the microarray and the RT-PCR data are more than the overlapping genes between RNAseq and RT-PCR data. **B.** The Venn diagram of up regulated genes in cartilages in each data set. Six genes identified in RT-PCR data were identified in RNAseq data, only 2 were identified in microarray data; **C.** The Venn diagram of down regulated genes in OA cartilages in each dataset. Nine genes identified in RT-PCR data were identified in RNAseq data, including the 4 genes identified in microarray data.

Among the 427 genes detected in the RT-PCR data, 274 genes were detected by all of the technologies. Their expressions are relatively higher than the other genes in the RT-PCR data (Figure 6.2), while the 56 genes detected by only the technology were expressed at the lowest levels compared to other genes. The genes that were not expressed in the RNAseq data had lower expression levels than the other genes in the RT-PCR data. The expressions of the 56 unique genes to the RT-PCR were also checked in the RNAseq data, as these genes could be detected using RNAseq but did not pass the criteria to be considered as expressed. None of the 56 genes had more than 0.5 molecule copy detected in the RNAseq data. Compared to the expressed genes, which had around or more than 20 copies, these genes expressed at very low levels (Table 6.2). Comparison of the Ct values of the 56 genes to all of the expressed genes in the RT-PCR confirmed the relative lower expression of the 56 genes.

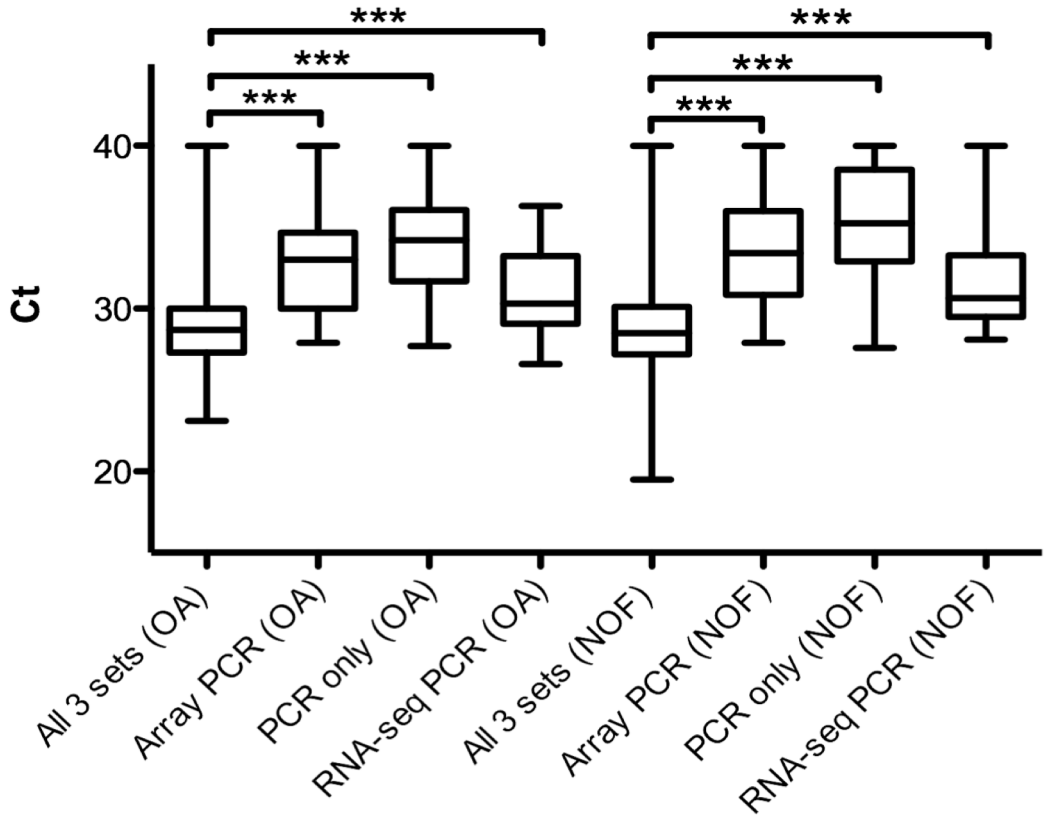


Figure 6.2 Relative expressions of genes in the rt-PCR data grouped by intersections of the 3 data sets. The figure shows the Ct values of genes in the RT-PCR data. The genes are separated into 8 groups: All 3 sets: genes that were commonly detected in the 3 data sets; Array PCR: genes that were detected in both the microarray and the RT-PCR data but not in the RNAseq data; PCR only: genes that were detected in the RT-PCR data only; RNAseq PCR: genes that were detected in both the RNAseq and the RT-PCR data but not in the microarrays data. Ct values for OA and NOF samples were separately denoted. The asterisks indicate the significances of the comparison between the groups. "***" means the p-value < 0.001. Comparing the genes commonly detected in all of the 3 data sets, genes detected by the RNAseq had lower Ct values meaning more expression than the other 2 groups of genes. Genes that were detected by RT-PCR only had highest Ct values, which meant they have lowest expression levels. It implies that the RNAseq data does not have enough coverage to detect the lowly expressed genes.

	RNAseq (copy of the molecule)		RT-PCR (Ct)	
	Genes expressed in RT-PCR only	**All expressed genes in RNAseq**	**Genes expressed in RT-PCR only**	**All expressed genes in RT-PCR**
OA	0.42 (±0.033, n=56)	19.75 (±1.59, n=14507)	33.87 (±0.15, n=56)	29.67 (±0.45, n=370)
NOF	0.38 (±0.031, n=56)	22.05 (±1.59, n=14507)	35.24 (±0.18, n=56)	29.97 (±0.50, n=370)

Table 6.2: Mean expressions of genes expressed in RT-PCR only and other expressed genes in the RNAseq and RT-PCR data set. For the RNAseq data, gene expressions were measured with approximate number of molecule copies of genes, which were calculated as: (number of reads mapped to a gene) * (read length) / (exonic length of the gene). Ct values were used as an approximate measure of gene expressions in RT-PCR data. The mean expression values for OA and NOF are listed in the table. The standard error of mean "±" and number of genes "n" are listed. The mean expression values of genes that were expressed in RT-PCR data were lower than other genes in both RNAseq and in RT-PCR data, which indicates that these genes were expressed at very lower levels comparing to the other genes.

6.3.2 Comparison of fold changes of genes in the datasets

The other quality criterion of a gene expression profiling technology is the accuracy of expression quantification, as none of these technologies can provide absolute expression levels of genes, fold changes of genes (OA/NOF) commonly detected between any two of the data sets were compared and the correlation of the fold changes were tested (Figure 6.3). There were 345 genes commonly expressed between the RT-PCR and the microarray data. The correlation of the fold changes is significant (P-value =1.54e-4) but is not strong (r = 0.19). Compared to the microarray data, RNAseq data showed better correlation with the RT-PCR data (P-value = 5.77e-12 and r=0.38). The correlation between the two high-throughput technologies is significant (P-value < 2.2e-16, note: the smallest P-value that R can produce) with an r = 0.74.

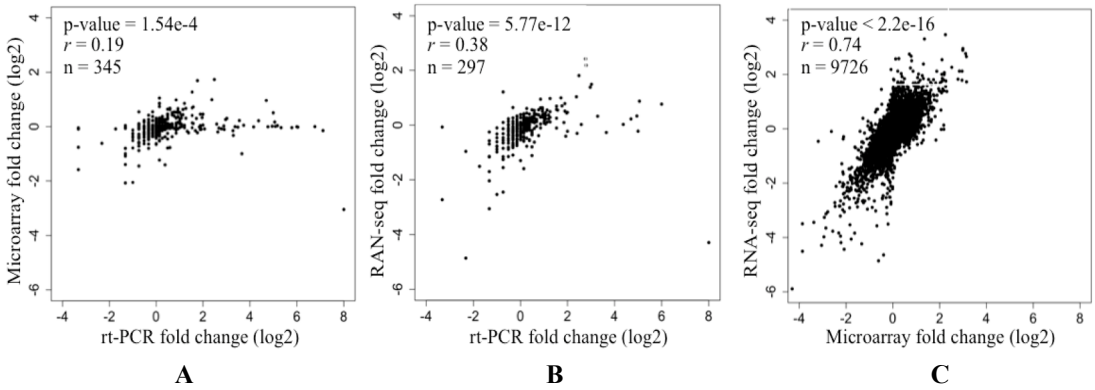


Figure 6.3: Comparison of the fold changes between data sets. A. The correlation of fold changes between microarray and RT-PCR data. There were 345 genes in total found expressed in cartilage in the both data sets. The p-value of the correlation is significant but the correlation is not as strong as between RNAseq and RT-PCR data; **B.** The correlation of fold changes between both RNAseq and RT-PCR data. Genes that could be used for the test are fewer comparing to A, but the correlation is stronger; **C.** The fold change correlation between microarray and RNAseq. There were 9726 genes found expressed in the two data sets. The correlation is significant (P-value < 2.2-e16, which is the smallest p-value that R can produce) and stronger (r =0.74).

6.4 Discussion

The very lowly expressed genes that were only detected by the RT-PCR data imply the insufficient coverage of our RNAseq data. As the protocol of the RNAseq library construction involves amplification of the cDNA sequences, highly expressed transcripts will have more chance to be sequenced than transcripts expressed at relatively low levels. In our experiment, the RNAseq has achieved 47-fold coverage of the whole transcriptome, which in fact exceed the requirement in the ENCODE guideline (ENCODE, 2009), while there were still genes expressed in the cartilage that did not receive enough reads to quantify them. This indicates that in order to investigate lowly expressed genes more sequencing depth is required and, depending on the expression levels of the genes, the requirement of the depth may increase exponentially.

To map the gene identifiers used in the different data sets, we tried to map all identifiers to ENSEMBL, while in the RT-PCR and the microarray data some of the genes could not be mapped. The reason could be the lack of updates of the annotations used in the two datasets. This led to the difficulties in comparison of genes. However, the RNAseq

has advantages in quantifying genes in latest annotation, because the generation of the RNAseq data does not depend on any transcriptome knowledge thus the data analysis can easily be adapted to the latest available transcriptome or genome annotation. In comparison, the ability of the microarray platforms is restricted by the number of the probes printed on the chips and the efficiency of these probes. The design of the probes often represents the knowledge of the transcriptome at the time, but such knowledge quickly becomes outdated, especially after the emergence of the next-generation sequencing.

Owing to the insufficient coverage of the RNAseq data, the microarray detected more expressed genes confirmed by rt-PCR data. This revealed an advantage of the microarray platform in profiling lowly expressed genes, which may not be cost-effective for RNAseq studies at the moment because of the uncertainty of how much depth is needed to be achieved for reliable quantifications of such genes (Tarazona *et al.*, 2011). However, in terms of differentially expressed genes, the RNAseq data identified more genes than the microarray platform, plus shared a better correlation with the RT-PCR. This could be due to the performance of the probes of the array chip. Several factors, such as cross-hybridization, can affect the performance of the probes on microarray chips (Chou *et al.*, 2004). Furthermore, the data analysis, especially the normalization, of RNAseq data is relatively simpler than the microarray data, as it has less background noise and single-base resolution. Although the number of sequencing reads derived from transcripts can be biased by their lengths, abundances and GC contents (Oshlack and Wakefield, 2009; Hansen *et al.*, 2012), when identifying differentially expressed genes the bias can be cancelled, as the same sequences are compared between samples. In the comparison of the fold change correlations, the RNAseq data also presented better correlation with the rt-PCR data than the microarray data, indicating the better accuracy of the technology. Herein the whole comparison of the two high-throughput technologies is based on gene expression profiling of human cartilage. When considering an organism without available complete gene annotation, such as most of plants and bacteria, the advantages of the RNAseq are even stronger. Furthermore the same data can also be used to determine splicing events, transcript sequences and other aspects of the transcriptome.

rt-PCR data showed worse fold change correlation with both of the microarray and

RNAseq data in this project, comparing to the correlation between the later two. This could largely because the cartilage samples used in the RNAseq experiment were just a subset of the microarray experiment, while cartilage from completely different patients were used in the rt-PCR study. The slight differences of RNA extraction protocol could contribute as well. The also implied that the sample size used in the three experiments were not sufficient to neutralize the genetic difference of individuals.

In fact, some of the discrepancies between the 3 datasets could also be from biological factors. For example, a recent study showed that the circadian clock of cartilage tissues regulated the gene expressions as well (Gossan *et al.*, 2013). In the study, 615 genes were found to have a circadian pattern. This could explain why some of the differentially expressed genes were found in one data set but not in the other two, and also the fold change differences. Our samples were collected from the same operation theatres and the isolation of RNA occurred at approximately the same time of data, but not such information was available from the PCR study. Furthermore, the age of the cartilage donors could have differed. Circadian rhythms are known to be altered by chronological age (Gossan *et al.*, 2013).

In conclusion, compared to the microarray, the RNAseq has better accuracy in terms of quantifying genes and identification of differentially expressed genes, but may present difficulties for lowly expressed genes when the sequencing depth is insufficient. The analysis of the RNAseq data is still not as mature as for the microarray data, but with the continually increasing interests in the technology, it will be as easy as microarray data analysis in the next several years.

Chapter 7 General Discussions

In order to understand the molecular changes in OA, in this project I investigated the OA transcriptome with two different technologies, microarray and RNAseq. With both technologies, more than a thousand genes were identified as differentially expressed. Both up-regulation of collagens and down-regulation of aggrecanases were found, including *ADAMTS2* that was only identified in the RNAseq data. When applied the same *P* value and fold change threshold to the both gene sets, the number of DE genes identified with RNAseq was around two-fold more than identified with microarray technologies. Although the cartilages samples used in the RNAseq experiment were a subset samples of the microarray datasets, the common DE genes of the two dataset were no more than three hundred. Nevertheless, there is still a significant correlation of fold changes between the two datasets.

With DE genes, over hundreds canonical pathways were found associated with OA. These include known OA associated pathways, such as Wnt/β-catenin Signaling, PI3K/AKT Signaling, HIF1α Signaling, LXR/RXR Activation, p38 MAPK Signaling, iNOS Signaling and Acute Phase Response Signaling, with later two only identified in the RNAseq data. Both of my experiments also associated more than 50 other canonical pathways with OA, such as Oncostatin M Signalling, TREM1 signalling and IL-17 Signalling. Three genes *CDKN1A*, *VEGFA* and *MYC* were also identified as keys genes(hubs) that linked the OA associated pathways and networks, among which MYC was reported to be associated with chondrocytes apoptosis.

Because of the advances of the technology, in the RNAseq experiment more findings of OA were discovered, in addition to the DE genes. As expected, most of protein-coding DE genes have differentially expressed transcripts as well, but there are also DE transcripts of genes with no expression change in OA. Most of collagens were found have up-regulated DE transcripts in OA, while *COL2A1*, *COL5A1, COL5A2* and *COL11A1* were also found up regulated in OA on gene level. A known alternatively spliced transcript of *ADAMTS4* was also found in the OA cartilage, which leads to the confidence of other alternative splicing events identified from the data, such as the transcripts composition change of *MMP3* in OA. Alternative splicing events have been

found to be associated with a number of diseases (Garcia-Blanco *et al.*, 2004), but it is difficult to determine whether a DE isoform can contribute to the destruction of chondrocytes without the overall gene expression change, however, these findings of the DE transcripts and isoforms can still be useful in terms of understanding the whole mechanism of OA progression.

The sequence and and the structure of the transcripts expressed in cartilage samples, and those uniquely expressed in normal and OA hip cartilage were also identified in this study. Due to the limitation of the analysis methods at the time, novel transcripts could be missed, but such results can still be very informative for future probe designs and other systematic studies of hip cartilages.

OA is a complex disease and have been studied for years from multiple different angles including genetics and epigenetics in recent years (Reynard and Loughlin, 2012). In this project, with the transcriptomes derived using the the two high-throughput technologies, I added more understanding to the disease mechanism, in the hope of providing interesting targets to conquer it in the future.

Chapter 8 Future Work

Since the emergence of RNAseq technology, it has started to show the trend of replacing the microarray technology. The knowledge of the RNAseq has expanded as well, including the understanding of the RNAseq data itself and the algorithms/tools for the analysis. All of the knowledge and the new algorithms/tools could be used to improve my analysis of the RNAseq data.

The sequencing depth was recognized as the critical parameter of RNAseq for identification of differential expressed genes (Tarazona *et al.*, 2011). Our RNAseq data has a sequencing depth in line with the recommended depth of ENCODE (ENCODE, 2009) for moderate gene expression profiling. In total, 14,507 genes were found expressed in the cartilage samples. With this depth, I managed to assemble transcripts, identify alternative splicing events and differential allelic expressions. However, in order to detect lowly expressed transcripts and increase the accuracy of expression estimation, the sequencing depths of the samples need to be increased. Using the new algorithms/tools published in the last several years, the whole analysis of the RNAseq data could be improved.

The mapping algorithm for RNAseq reads is a constant interest in the field and critical for accurate estimation of gene expression. Several aligners were published over the last several years since I undertook the work, such as STAR (Dobin *et al.*, 2013), which features better performance and significantly less computing time in comparison with other aligners (Engstrom *et al.*, 2013). The aligner was also recommended to be used as the best practice from the GATK developers for the detection of variants (GATK-Team, 2014). The use of STAR could improve both the gene expression estimation and the splicing event detection with less computing time. Besides the transcript assembly, mapping the reads to the reference genome usually is the most time consuming stage in an RNAseq data analysis pipeline, thus less computing time is also a desirable feature of an aligner.

In terms of de novo assembly of transcriptomes, several tool sets was published, such as Velvet-Oases (Schulz *et al.*, 2012), ABySS (Birol *et al.*, 2009) and Trinity (Grabherr *et al.*, 2011). Trinity has implementation to support the use of computing grid, which reduces the computing time and also the memory usage. With the improvement of our local computing setup, it is not feasible to use Trinity to assemble transcriptomes. Plus tools to estimate expression of transcripts are also provided in the software package, which eases the use of the software. Compared to the strategies of BitSeq and Cufflinks, which utilize the existing transcript annotations, Trinity can be used on organisms without a reference genome or comprehensive annotation. After transcripts assembly, alternative splicing events can also be easily identified.

Variant calling algorithms evolved from simple calling methods, like Samtools and Varscan, to more advanced methods, such as freebayes (Garrison and Marth, 2012) and HaplotypeCaller of GATK (McKenna *et al.*, 2010). Instead of searching along the genome for different bases from the reference, freebayes and HaplotypeCaller assemble the reads first then compare the assembly to the reference. With this method, false positives of mis-alignment caused by indels and low quality bases can be removed. Variants at the RNA level can be identified more accurately using these more advanced variant callers, thus allelic expression analysis and RNA-editing events identification would also be more accurate. In fact, the recent published tool 'REDI-tools' (Picardi and Pesole, 2013) provides the whole pipeline for RNA-editing event identification. It also has functions to compare RNAseq data to corresponded DNA data, which eases the RNA-editing analysis.

In the recent years, the methodology of the RNAseq library preparation and sequencing technology has also been developed. Strand specific RNAseq (Levin *et al.*, 2010) becomes more popular and is replacing the standard paired-end sequencing strategy. The strand specific RNAseq preserves the strand information of the transcripts. The strand information of the transcripts completes the transcriptome and also helps to differentiate reads from anti-sense transcripts of genes. Single cell RNAseq has also been introduced (Tang *et al.*, 2011). The technique allows researchers to study how a single cell response to the environmental stimulates at its different developmental stages.

My PhD project represents a pioneer and comprehensive transcriptome study of hip OA cartilages using the RNAseq. The technology has received extensive interests from the researchers since it was introduced because its irreplaceable advantages over the microarray analysis. Both the application and the analysis of the technology are continually improving, which not only extends the ability of the technology to suit different research themes but also eases the analysis towards customized research themes. Although the analysis is still not as straight forward as for the microarray data, in the forthcoming years, with the knowledge and experience accumulation, such analysis should become standard.

Appendices

Supplementary S5.1 R script of association between RNA-editing events and OA

```
x <- as.numeric(Sys.getenv("SGE_TASK_ID"))

testEdit<-function(j){
	Data20121011<-
read.table("Common_Heterozygous_with_info_filtered_4_regression_newVersion.tx
t_Editing_tested.txt_preprocessed", head=T, as.is=T, sep = "\t")
	Data20121011<-Data20121011[Data20121011$cRef!="-" &
Data20121011$cVar!="-",]
	Data20121011<-Data20121011[Data20121011$cRef!="0" |
Data20121011$cVar!="0",]
	VariantsID <- with(Data20121011, paste(Chromosome, Start, sep="."))
	Data20121011$VarID <- VariantsID
	Data20121011$cRef <- as.integer(Data20121011$cRef)
	Data20121011$cVar <- as.integer(Data20121011$cVar)
	Data20121011$Disease <- as.factor(Data20121011$Disease)
	Data20121011$Sample_name <- as.factor(Data20121011$Sample_name)

	NCov <-2 # what is this?????

	AllVariants <- unique(VariantsID)

	library(lme4)

	###Control how to split the file
	max <- 10000
	x<- seq_along(AllVariants)
	y<- x
	Split <- split(y, ceiling(x/max))

	Index <- Split[[j]]
	AllVariants.sub <- AllVariants[Index]
	###Control how to split the file ### end


	N <- length(AllVariants.sub)
	AllResults <- vector(length=N, mode="character")
	Mod0Warnings <- vector(length=N, mode="character")
	Mod1Warnings <- vector(length=N, mode="character")
	NA.test.type <- vector(length=N, mode="character")

	for (i in seq(N)){
		Indx<-Data20121011$VarID==AllVariants.sub[i] #chr15.100246942
		cat(i," ",AllVariants.sub[i],": ")
		YDat<-data.frame(Ref=Data20121011$cRef[Indx],
Var=Data20121011$cVar[Indx])

		if ((length(which(YDat$Ref>0))<2) ||
(length(which(YDat$Var>0))<2)) {# handle 0 count
			AllResults[i] <- "NA"
			NA.test.type[i] <- "0ReforVar"
```

```
                cat("0_NA","\n")
                next
            }

            XDat<-data.frame(Status=Data20121011$Disease[Indx],
SampleID=as.factor(1:length(Indx[Indx])))

        if(length(XDat$Status) == length(XDat$Status[XDat$Status == "NOF"]) |
length(XDat$Status) == length(XDat$Status[XDat$Status == "OA"])) {
               AllResults[i] <- "NA"
               NA.test.type[i] <- "1ConMissing"
               cat("MissCon_NA","\n")
               next
        }

      if(length( which(XDat$Status == "OA") ) < 6 &
length( which(XDat$Status == "NOF") ) <= 3 ) {
            AllResults[i] <- "NA"
            NA.test.type[i] <- "TooFewSamples"
            cat("TooFewSamples_NA","\n")
            next
        }

             TotCounts<-rowSums(YDat)
             #if (Trace) cat(i," ",as.character(Vars[i]),": ")
             Y<-unlist(apply(YDat,1,function(x){rep(c(0,1),x)}))
             X<-as.data.frame(matrix(unlist(apply(cbind(TotCounts,XDat),1,
function(x){rep(x[-1],as.integer(x[1]))})),byrow=T,ncol=NCov))
             WrkgDf<-cbind(Y,X)
             names(WrkgDf)<-c("Read",names(XDat))
             ConTab<-table(c(WrkgDf$Read,0,1),c(WrkgDf$Status,1,2) )-
diag(c(1,1))
             TotReads<-sum(ConTab)
             DiagCounts<- ConTab[1,1]+ConTab[2,2]

             if (DiagCounts==0 || DiagCounts==TotReads) {
                   P.val.fish <--fisher.test(ConTab)$p.value
                   AllResults[i] <- P.val.fish
                   NA.test.type[i] <- "fisher"
                   cat("fisher", " ", P.val.fish,"\n")
                   next
             }

             H0.glm<-tryCatch(lmer(Read~(1|SampleID),data=WrkgDf),
                      error=function(e){return(NA)},
                      warning=function(w){
                                           Mod0Warnings[i] <- "Y"
                                           return(NA)}
                                           )
             HA.glm<-tryCatch(lmer(Read~Status+(1|SampleID),data=WrkgDf),
                      error=function(e){return(NA)},
                      warning=function(w){
                                           Mod1Warnings[i] <- "Y"
                                           return(NA)}
                                           )
             if ((class(H0.glm)[1]=="mer") && (class(HA.glm)[1]=="mer")) {
                   LogLike1<-logLik(HA.glm)[[1]]
                   LogLike0<-logLik(H0.glm)[[1]]
```

```
                  P.val=pchisq(2*(LogLike1-LogLike0),1,lower=F)
                  NA.test.type[i] <- "glm"
                  AllResults[i] <- P.val
                  cat(P.val,"\n")
            } else {
                  AllResults[i] <- "NA"
                  NA.test.type[i] <- "Error_War"
                  cat("Error_War"," ","NA","\n")
            }

      }
      table.name <-
paste(paste("NewSlicingRestult/regression_array_test_slice", min(Index),
max(Index), sep="_"),"txt", sep = ".")
      result <- data.frame(Variants_id=AllVariants.sub, P_value=AllResults,
test_type = NA.test.type, Mod0_warnings = Mod0Warnings, Mod1_warnings =
Mod1Warnings)
      write.table(result, file=table.name, row.name=F, col.name=F,
quote=F,sep="\t")
}

testEdit(x)
```

Supplementary file S5.2 Novel transcripts in cartilage in GTF format (Not printed)

Reference

Adams, M.E., Huang, D.Q., Yao, L.Y. and Sandell, L.J. (1992) 'Extraction and isolation of mRNA from adult articular cartilage', *Anal Biochem*, 202(1), pp. 89-95.

Ahmed, A.S., Li, J., Erlandsson-Harris, H., Stark, A., Bakalkin, G. and Ahmed, M. (2012) 'Suppression of pain and joint destruction by inhibition of the proteasome system in experimental osteoarthritis', *Pain*, 153(1), pp. 18-26.

Aigner, T., Fundel, K., Saas, J., Gebhard, P.M., Haag, J., Weiss, T., Zien, A., Obermayr, F., Zimmer, R. and Bartnik, E. (2006a) 'Large-scale gene expression profiling reveals major pathogenetic pathways of cartilage degeneration in osteoarthritis', *Arthritis Rheum*, 54(11), pp. 3533-44.

Aigner, T., Fundel, K., Saas, J., Gebhard, P.M., Haag, J., Weiss, T., Zien, A., Obermayr, F., Zimmer, R. and Bartnik, E. (2006b) 'Large-scale gene expression profiling reveals major pathogenetic pathways of cartilage degeneration in osteoarthritis', *Arthritis and Rheumatism*, 54(11), pp. 3533-3544.

Aigner, T., Zien, A., Gehrsitz, A., Gebhard, P.M. and McKenna, L. (2001) 'Anabolic and catabolic gene expression pattern analysis in normal versus osteoarthritic cartilage using complementary DNA-array technology', *Arthritis Rheum*, 44(12), pp. 2777-89.

Anders, S. and Huber, W. (2010) 'Differential expression analysis for sequence count data', *Genome Biology*, 11(10), p. R106.

Anders, S., Pyl, P.T. and Huber, W. (2014) 'HTSeq - A Python framework to work with high-throughput sequencing data', *Bioinformatics*.

Anders, S., Reyes, A. and Huber, W. (2012) 'Detecting differential usage of exons from RNA-seq data', *Genome Research*, 22(10), pp. 2008-2017.

Andrews, S. *FastQC A Quality Control tool for High Throughput Sequence Data* Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Appleton, C.T., Pitelka, V., Henry, J. and Beier, F. (2007) 'Global analyses of gene expression in early experimental osteoarthritis', *Arthritis Rheum*, 56(6), pp. 1854-68.

Aris, V., Cody, M., Cheng, J., Dermody, J., Soteropoulos, P., Recce, M. and Tolias, P. (2004) 'Noise filtering and nonparametric analysis of microarray data underscores discriminating markers of oral, prostate, lung, ovarian and breast cancer', *BMC Bioinformatics*, 5(1).

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, M., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J.,

Ringwald, M., Rubin, G. and Sherlock, G. (2000) 'Gene Ontology: tool for the unification of biology', *Nature Genetics*, 25(1), pp. 25-29.

Bahn, J.H., Lee, J.H., Li, G., Greer, C., Peng, G. and Xiao, X. (2012) 'Accurate identification of A-to-I RNA editing in human by transcriptome sequencing', *Genome Res*, 22(1), pp. 142-50.

Barbazuk, W.B., Emrich, S.J., Chen, H.D., Li, L. and Schnable, P.S. (2007) 'SNP discovery via 454 transcriptome sequencing', *Plant J*, 51(5), pp. 910-8.

Bau, B., Gebhard, P.M., Haag, J., Knorr, T., Bartnik, E. and Aigner, T. (2002) 'Relative messenger RNA expression profiling of collagenases and aggrecanases in human articular chondrocytes in vivo and in vitro', *Arthritis Rheum*, 46(10), pp. 2648-57.

Beekhuizen, M., van Osch, G.J., Bot, A.G., Hoekstra, M.C., Saris, D.B., Dhert, W.J. and Creemers, L.B. (2013) 'Inhibition of oncostatin M in osteoarthritic synovial fluid enhances GAG production in osteoarthritic cartilage repair', *Eur Cell Mater*, 26, pp. 80-90; discussion 90.

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. Series B (Methodological)*, 57(1), pp. 289-300.

Berardi, S., Lang, A., Kostoulas, G., Horler, D., Vilei, E.M. and Baici, A. (2001) 'Alternative messenger RNA splicing and enzyme forms of cathepsin B in human osteoarthritic cartilage and cultured chondrocytes', *Arthritis Rheum*, 44(8), pp. 1819-31.

Birol, I., Jackman, S.D., Nielsen, C.B., Qian, J.Q., Varhol, R., Stazyk, G., Morin, R.D., Zhao, Y., Hirst, M., Schein, J.E., Horsman, D.E., Connors, J.M., Gascoyne, R.D., Marra, M.A. and Jones, S.J. (2009) 'De novo transcriptome assembly with ABySS', *Bioinformatics*, 25(21), pp. 2872-7.

Bolstad, B.M., Irizarry, R.A., Åstrand, M. and Speed, T.P. (2003) 'A comparison of normalization methods for high density oligonucleotide array data based on variance and bias', *Bioinformatics*, 19(2), pp. 185-193.

Brandt, K.D., Radin, E.L., Dieppe, P.A. and van de Putte, L. (2006) 'Yet more evidence that osteoarthritis is not a cartilage disease', *Ann Rheum Dis*, 65(10), pp. 1261-4.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S.R., Moon, K., Burcham, T., Pallas, M., DuBridge, R.B., Kirchner, J., Fearon, K., Mao, J. and Corcoran, K. (2000) 'Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays', *Nat Biotechnol*, 18(6), pp. 630-4.

Bullard, J.H., Purdom, E., Hansen, K.D. and Dudoit, S. (2010) 'Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments', *BMC Bioinformatics*, 11, p. 94.

Cagnard, N., Letourneur, F., Essabbani, A., Devauchelle, V., Mistou, S., Rapinat, A., Decraene, C., Fournier, C. and Chiocchia, G. 'Interleukin-32, CCL2, PF4F1 and GFD10 are the only cytokine/chemokine genes differentially expressed by in vitro cultured rheumatoid and osteoarthritis fibroblast-like synoviocytes'.

Campian, A.R. (2002) *Arthritis: The big picture*. Available at: https://www.ipsos-mori.com/Assets/Docs/Archive/Polls/arthritis.pdf.

Canales, R., Luo, Y., Willey, J., Austermiller, B., Barbacioru, C., Boysen, C., Hunkapiller, K., Jensen, R., Knight, C., Lee, K., Ma, Y., Maqsodi, B., Papallo, A., Peters, E.H., Poulter, K., Ruppel, P., Samaha, R., Shi, L., Yang, W., Zhang, L. and Goodsaid, F. (2006) 'Evaluation of DNA microarray results with quantitative gene expression platforms', *Nature biotechnology*, 24(9), pp. 1115-1122.

Casneuf, T., Van de Peer, Y. and Huber, W. (2007) 'In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation', *BMC Bioinformatics*, 8, p. 461.

Chapman, K., Takahashi, A., Meulenbelt, I., Watson, C., Rodriguez-Lopez, J., Egli, R., Tsezou, A., Malizos, K.N., Kloppenburg, M., Shi, D., Southam, L., van der Breggen, R., Donn, R., Qin, J., Doherty, M., Slagboom, P.E., Wallis, G., Kamatani, N., Jiang, Q., Gonzalez, A., Loughlin, J. and Ikegawa, S. (2008) 'A meta-analysis of European and Asian cohorts reveals a global role of a functional SNP in the 5' UTR of GDF5 with osteoarthritis susceptibility', *Hum Mol Genet*, 17(10), pp. 1497-504.

Chatterjee, A., Stockwell, P.A., Rodger, E.J. and Morison, I.M. (2012) 'Comparison of alignment software for genome-wide bisulphite sequence data', *Nucleic Acids Res*, 40(10), p. e79.

Chen, C. and Bundschuh, R. (2012) 'Systematic investigation of insertional and deletional RNA-DNA differences in the human transcriptome', *BMC Genomics*, 13, p. 616.

Cheng, A.W., Stabler, T.V., Bolognesi, M. and Kraus, V.B. (2011) 'Selenomethionine inhibits IL-1beta inducible nitric oxide synthase (iNOS) and cyclooxygenase 2 (COX2) expression in primary human chondrocytes', *Osteoarthritis Cartilage*, 19(1), pp. 118-25.

Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkaloglu, A., Ozen, S., Sanjad, S., Nelson-Williams, C., Farhi, A., Mane, S. and Lifton, R.P. (2009) 'Genetic diagnosis by whole exome capture and massively parallel DNA sequencing', *Proc Natl Acad Sci U S A*, 106(45), pp. 19096-101.

Chomczynski, P. and Sacchi, N. (1987) 'Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction', *Anal Biochem*, 162(1), pp. 156-9.

Chou, C.C., Chen, C.H., Lee, T.T. and Peck, K. (2004) 'Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression', *Nucleic Acids Res*, 32(12), p. e99.

Chou, C.H., Lee, C.H., Lu, L.S., Song, I.W., Chuang, H.P., Kuo, S.Y., Wu, J.Y., Chen, Y.T., Kraus, V.B., Wu, C.C. and Lee, M.T.M. (2013) 'Direct assessment of articular cartilage and underlying subchondral bone reveals a progressive gene expression change in human osteoarthritic knees', *Osteoarthritis and Cartilage*, 21(3), pp. 450-461.

Chu, T.-M., Weir, B. and Wolfinger, R. (2002) 'A systematic statistical linear modeling approach to oligonucleotide array experiments', *Mathematical Biosciences*, 176(1), pp. 35-51.

Clark, A.L., Votta, B.J., Kumar, S., Liedtke, W. and Guilak, F. (2010) 'Chondroprotective role of the osmotically sensitive ion channel transient receptor potential vanilloid 4: age- and sex-dependent progression of osteoarthritis in Trpv4-deficient mice', *Arthritis Rheum*, 62(10), pp. 2973-83.

Clements, D.N., Vaughan-Thomas, A., Peansukmanee, S., Carter, S.D., Innes, J.F., Ollier, W.E. and Clegg, P.D. (2006) 'Assessment of the use of RNA quality metrics for the screening of articular cartilage specimens from clinically normal dogs and dogs with osteoarthritis', *Am J Vet Res*, 67(8), pp. 1438-44.

Collins-Racie, L.A., Yang, Z., Arai, M., Li, N., Majumdar, M.K., Nagpal, S., Mounts, W.M., Dorner, A.J., Morris, E. and LaVallie, E.R. (2009) 'Global analysis of nuclear receptor expression and dysregulation in human osteoarthritic articular cartilage: reduced LXR signaling contributes to catabolic metabolism typical of osteoarthritis', *Osteoarthritis Cartilage*, 17(7), pp. 832-42.

Costigan, M., Belfer, I., Griffin, R.S., Dai, F., Barrett, L.B., Coppola, G., Wu, T., Kiselycznyk, C., Poddar, M., Lu, Y., Diatchenko, L., Smith, S., Cobos, E.J., Zaykin, D., Allchorne, A., Gershon, E., Livneh, J., Shen, P.H., Nikolajsen, L., Karppinen, J., Mannikko, M., Kelempisioti, A., Goldman, D., Maixner, W., Geschwind, D.H., Max, M.B., Seltzer, Z. and Woolf, C.J. (2010) 'Multiple chronic pain states are associated with a common amino acid-changing allele in KCNS1', *Brain*, 133(9), pp. 2519-27.

Cox, A. (2007) *ELAND: Efficient large-scale alignment of nucleotide databases* [Computer program]. Illumina.

Danecek, P., Nellaker, C., McIntyre, R.E., Buendia-Buendia, J.E., Bumpstead, S., Ponting, C.P., Flint, J., Durbin, R., Keane, T.M. and Adams, D.J. (2012) 'High levels of RNA-editing site conservation amongst 15 laboratory mouse strains', *Genome Biol*, 13(4), p. 26.

Davidson, R., Waters, J., Kevorkian, L., Darrah, C., Cooper, A., Donell, S. and Clark, I. (2006) 'Expression profiling of metalloproteinases and their inhibitors in synovium and cartilage', *Arthritis Research & Therapy*, 8(4), p. R124.

Dell'accio, F., De Bari, C., Eltawil, N.M., Vanhummelen, P. and Pitzalis, C. (2008) 'Identification of the molecular response of articular cartilage to injury, by microarray screening: Wnt-16 expression and signaling after injury and in osteoarthritis', *Arthritis Rheum*, 58(5), pp. 1410-21.

Dell'accio, F. and Vincent, T.L. (2010) 'Joint surface defects: clinical course and cellular response in spontaneous and experimental lesions', *European cells & materials*, 20, pp. 210-217.

Denko, C.W. and Malemud, C.J. (2005) 'Role of the Growth Hormone/Insulin-like Growth Factor-1 Paracrine Axis in Rheumatic Diseases', *Seminars in Arthritis and Rheumatism*, 35(1), pp. 24-34.

Dieguez-Gonzalez, R., Calaza, M., Shi, D., Meulenbelt, I., Loughlin, J., Tsezou, A., Dai, J., Malizos, K.N., Slagboom, E.P., Kloppenburg, M., Chapman, K., Jiang, Q., Kremer, D., Gomez-Reino, J.J., Nakajima, N., Ikegawa, S. and Gonzalez, A. (2009) 'Testing the druggable endothelial differentiation gene 2 knee osteoarthritis genetic factor for replication in a wide range of sample collections', *Ann Rheum Dis*, 68(6), pp. 1017-21.

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(1), pp. 15-21.

Du, P., Kibbe, W. and Lin, S. (2008) 'lumi: a pipeline for processing Illumina microarray', *Bioinformatics*, 24(13), pp. 1547-1548.

DuRaine, G.D., Chan, S.M. and Reddi, A.H. (2011) 'Effects of TGF-beta1 on alternative splicing of Superficial Zone Protein in articular cartilage cultures', *Osteoarthritis Cartilage*, 19(1), pp. 103-10.

Egli, R.J., Southam, L., Wilkins, J.M., Lorenzen, I., Pombo-Suarez, M., Gonzalez, A., Carr, A., Chapman, K. and Loughlin, J. (2009) 'Functional analysis of the osteoarthritis susceptibility-associated GDF5 regulatory polymorphism', *Arthritis Rheum*, 60(7), pp. 2055-64.

ENCODE (2009) *Standards, Guidelines and Best Practices for RNA-Seq: 2010/2011*. Available at: http://genome.ucsc.edu/ENCODE/protocols/dataStandards/RNA_standards_v1_2011_May.pdf.

Engstrom, P.G., Steijger, T., Sipos, B., Grant, G.R., Kahles, A., Ratsch, G., Goldman, N., Hubbard, T.J., Harrow, J., Guigo, R. and Bertone, P. (2013) 'Systematic evaluation of spliced alignment programs for RNA-seq data', *Nat Methods*, 10(12), pp. 1185-91.

Espina, V., Mueller, C., Edmiston, K., Sciro, M., Petricoin, E.F. and Liotta, L.A. (2009) 'Tissue is alive: New technologies are needed to address the problems of protein biomarker pre-analytical variability', *Proteomics Clin Appl*, 3(8), pp. 874-882.

Evangelou, E., Chapman, K., Meulenbelt, I., Karassa, F.B., Loughlin, J., Carr, A., Doherty, M., Doherty, S., Gómez-Reino, J.J., Gonzalez, A., Halldorsson, B.V., Hauksson, V.B., Hofman, A., Hart, D.J., Ikegawa, S., Ingvarsson, T., Jiang, Q., Jonsdottir, I., Jonsson, H., Kerkhof, H.J.M., Kloppenburg, M., Lane, N.E.,

Li, J., Lories, R.J., van Meurs, J.B.J., Näkki, A., Nevitt, M.C., Rodriguez-Lopez, J., Shi, D., Slagboom, P.E., Stefansson, K., Tsezou, A., Wallis, G.A., Watson, C.M., Spector, T.D., Uitterlinden, A.G., Valdes, A.M. and Ioannidis, J.P.A. (2009) 'Large-scale analysis of association between GDF5 and FRZB variants and osteoarthritis of the hip, knee, and hand', *Arthritis & Rheumatism*, 60(6), pp. 1710-1721.

Evans, C.H., Gouze, J.N., Gouze, E., Robbins, P.D. and Ghivizzani, S.C. (2004) 'Osteoarthritis gene therapy', *Gene Ther*, 11(4), pp. 379-89.

Fabio, A.d. (1998) *Supplement to The Art of Getting Well: Cartilage Replacement: The Polymer Age*. Available at: http://www.arthritistrust.org/wp-content/uploads/2012/10/Cartilage-Replacement-The-Polymer-Age.pdf.

Felson, D.T. (2006) 'Clinical practice. Osteoarthritis of the knee', *N Engl J Med*, 354(8), pp. 841-8.

Felson, D.T., Zhang, Y., Hannan, M.T., Naimark, A., Weissman, B., Aliabadi, P. and Levy, D. (1997) 'Risk factors for incident radiographic knee osteoarthritis in the elderly: the Framingham Study', *Arthritis Rheum*, 40(4), pp. 728-33.

Fernandez-Moreno, M., Rego, I., Carreira-Garcia, V. and Blanco, F.J. (2008) 'Genetics in osteoarthritis', *Curr Genomics*, 9(8), pp. 542-7.

Francis-West, P.H., Abdelfattah, A., Chen, P., Allen, C., Parish, J., Ladher, R., Allen, S., MacPherson, S., Luyten, F.P. and Archer, C.W. (1999) 'Mechanisms of GDF-5 action during skeletal development', *Development*, 126(6), pp. 1305-15.

Fukui, N., Zhu, Y., Maloney, W.J., Clohisy, J. and Sandell, L.J. (2003) 'Stimulation of BMP-2 expression by pro-inflammatory cytokines IL-1 and TNF-alpha in normal and osteoarthritic chondrocytes', *J Bone Joint Surg Am*, 85-A Suppl 3, pp. 59-66.

Garcia-Blanco, M.A., Baraniak, A.P. and Lasda, E.L. (2004) 'Alternative splicing in disease and therapy', *Nat Biotechnol*, 22(5), pp. 535-46.

Garrison, E. and Marth, G. (2012) *Haplotype-based variant detection from short-read sequencing*. Available at: http://arxiv.org/abs/1207.3907v2.

Garzia, A., Etxebeste, O., Rodriguez-Romero, J., Fischer, R., Espeso, E.A. and Ugalde, U. (2013) 'Transcriptional changes in the transition from vegetative cells to asexual development in the model fungus Aspergillus nidulans', *Eukaryot Cell*, 12(2), pp. 311-21.

GATK-Team (2014) *Calling variants in RNAseq*. Available at: http://gatkforums.broadinstitute.org/discussion/comment/14927.

Gautier, L., Cope, L., Bolstad, B.M. and Irizarry, R.A. (2004) 'affy—analysis of Affymetrix GeneChip data at the probe level', *Bioinformatics*, 20(3), pp. 307-315.

Geyer, M., Grassel, S., Straub, R.H., Schett, G., Dinser, R., Grifka, J., Gay, S., Neumann, E. and Muller-Ladner, U. (2009) 'Differential transcriptome analysis of intraarticular lesional vs intact cartilage reveals new candidate genes in osteoarthritis pathophysiology', *Osteoarthritis Cartilage*, 17(3), pp. 328-35.

Giatromanolaki, A., Sivridis, E., Athanassou, N., Zois, E., Thorpe, P.E., Brekken, R.A., Gatter, K.C., Harris, A.L., Koukourakis, I.M. and Koukourakis, M.I. (2001a) 'The angiogenic pathway 'vascular endothelial growth factor/flk-I (KDR)-receptor' in rheumatoid arthritis and osteoarthritis', *Journal of Pathology*, 194(1), pp. 101-108.

Giatromanolaki, A., Sivridis, E., Athanassou, N., Zois, E., Thorpe, P.E., Brekken, R.A., Gatter, K.C., Harris, A.L., Koukourakis, I.M. and Koukourakis, M.I. (2001b) 'The angiogenic pathway "vascular endothelial growth factor/flk-1(KDR)-receptor" in rheumatoid arthritis and osteoarthritis', *J Pathol*, 194(1), pp. 101-8.

Giatromanolaki, A., Sivridis, E., Maltezos, E., Athanassou, N., Papazoglou, D., Gatter, K., Harris, A. and Koukourakis, M. (2003a) 'Upregulated hypoxia inducible factor-1alpha and -2alpha pathway in rheumatoid arthritis and osteoarthritis', *Arthritis Res Ther*, 5(4), pp. R193-R201.

Giatromanolaki, A., Sivridis, E., Maltezos, E., Athanassou, N., Papazoglou, D., Gatter, K.C., Harris, A.L. and Koukourakis, M.I. (2003b) 'Upregulated hypoxia inducible factor-1alpha and -2alpha pathway in rheumatoid arthritis and osteoarthritis', *Arthritis Res Ther*, 5(4), pp. R193-201.

Glasson, S.S., Askew, R., Sheppard, B., Carito, B., Blanchet, T., Ma, H.L., Flannery, C.R., Peluso, D., Kanki, K., Yang, Z., Majumdar, M.K. and Morris, E.A. (2005) 'Deletion of active ADAMTS5 prevents cartilage degradation in a murine model of osteoarthritis', *Nature*, 434(7033), pp. 644-8.

Glaus, P., Honkela, A. and Rattray, M. (2012) 'Identifying differentially expressed transcripts from RNA-seq data with biological variation', *Bioinformatics*, 28(13), pp. 1721-8.

Goldring, M.B. and Goldring, S.R. (2010) 'Articular cartilage and subchondral bone in the pathogenesis of osteoarthritis', *Ann N Y Acad Sci*, 1192, pp. 230-7.

Gossan, N., Zeef, L., Hensman, J., Hughes, A., Bateman, J.F., Rowley, L., Little, C.B., Piggins, H.D., Rattray, M., Boot-Handford, R.P. and Meng, Q.J. (2013) 'The circadian clock in murine chondrocytes regulates genes controlling key aspects of cartilage homeostasis', *Arthritis Rheum*, 65(9), pp. 2334-45.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N. and Regev, A. (2011) 'Full-length transcriptome assembly from RNA-Seq data without a reference genome', *Nat Biotechnol*, 29(7), pp. 644-52.

Gu, H., Smith, Z.D., Bock, C., Boyle, P., Gnirke, A. and Meissner, A. (2011) 'Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling', *Nat Protoc*, 6(4), pp. 468-81.

Hai, T., Wolford, C.C. and Chang, Y.S. (2010) 'ATF3, a hub of the cellular adaptive-response network, in the pathogenesis of diseases: Is modulation of inflammation a unifying component?', *Gene Expression*, 15(1), pp. 1-11.

Hamel, M.B., Toth, M., Legedza, A. and Rosen, M.P. (2008) 'Joint replacement surgery in elderly patients with severe osteoarthritis of the hip or knee: decision making, postoperative recovery, and clinical outcomes', *Arch Intern Med*, 168(13), pp. 1430-40.

Hansen, K.D., Brenner, S.E. and Dudoit, S. (2010) 'Biases in Illumina transcriptome sequencing caused by random hexamer priming', *Nucleic Acids Research*, 38(12), p. e131.

Hansen, K.D., Irizarry, R.A. and Wu, Z. (2012) 'Removing technical variability in RNA-seq data using conditional quantile normalization', *Biostatistics*, 13(2), pp. 204-16.

Harbers, M. and Carninci, P. (2005) 'Tag-based approaches for transcriptome research and genome annotation', *Nat Methods*, 2(7), pp. 495-502.

Hatem, A., Bozdag, D., Toland, A.E. and Catalyurek, U.V. (2013) 'Benchmarking short sequence mapping tools', *BMC Bioinformatics*, 14, p. 184.

Hatzis, C., Sun, H., Yao, H., Hubbard, R.E., Meric-Bernstam, F., Babiera, G.V., Wu, Y., Pusztai, L. and Symmans, W.F. (2011) 'Effects of tissue handling on RNA integrity and microarray measurements from resected breast cancers', *J Natl Cancer Inst*, 103(24), pp. 1871-83.

Heap, G.A., Yang, J.H., Downes, K., Healy, B.C., Hunt, K.A., Bockett, N., Franke, L., Dubois, P.C., Mein, C.A., Dobson, R.J., Albert, T.J., Rodesch, M.J., Clayton, D.G., Todd, J.A., van Heel, D.A. and Plagnol, V. (2010) 'Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing', *Hum Mol Genet*, 19(1), pp. 122-34.

Hirota, K., Yoshitomi, H., Hashimoto, M., Maeda, S., Teradaira, S., Sugimoto, N., Yamaguchi, T., Nomura, T., Ito, H., Nakamura, T., Sakaguchi, N. and Sakaguchi, S. (2007) 'Preferential recruitment of CCR6-expressing Th17 cells to inflamed joints via CCL20 in rheumatoid arthritis and its animal model', *Journal of Experimental Medicine*, 204(12), pp. 2803-2812.

Hopwood, B., Tsykin, A., Findlay, D.M. and Fazzalari, N.L. (2007) 'Microarray gene expression profiling of osteoarthritic bone suggests altered bone remodelling, WNT and transforming growth factor-beta/bone morphogenic protein signalling', *Arthritis Res Ther*, 9(5), p. R100.

Hu, Y., Huang, Y., Du, Y., Orellana, C.F., Singh, D., Johnson, A.R., Monroy, A., Kuan, P.F., Hammond, S.M., Makowski, L., Randell, S.H., Chiang, D.Y., Hayes, D.N., Jones, C., Liu, Y., Prins, J.F. and Liu, J. (2013) 'DiffSplice: the genome-wide detection of differential splicing events with RNA-seq', *Nucleic Acids Res*, 41(2), p. e39.

Huang, D., Sherman, B. and Lempicki, R. (2008) 'Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources', *Nat. Protocols*, 4(1), pp. 44-57.

Huang, D., Sherman, B. and Lempicki, R. (2009) 'Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists', *Nucleic Acids Research*, 37(1), pp. 1-13.

Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) 'Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources', *Nat Protoc*, 4(1), pp. 44-57.

Huang, J.G., Xia, C., Zheng, X.P., Yi, T.T., Wang, X.Y., Song, G. and Zhang, B. (2011a) '17beta-Estradiol promotes cell proliferation in rat osteoarthritis model chondrocytes via PI3K/Akt pathway', *Cell Mol Biol Lett*, 16(4), pp. 564-75.

Huang, J.G., Xia, C., Zheng, X.P., Yi, T.T., Wang, X.Y., Song, G. and Zhang, B. (2011b) '17β-Estradiol promotes cell proliferation in rat osteoarthritis model chondrocytes via PI3K/Akt pathway', *Cellular and Molecular Biology Letters*, 16(4), pp. 564-575.

Ijiri, K., Zerbini, L.F., Peng, H., Otu, H.H., Tsuchimochi, K., Otero, M., Dragomir, C., Walsh, N., Bierbaum, B.E., Mattingly, D., Van Flandern, G., Komiya, S., Aigner, T., Libermann, T.A. and Goldring, M.B. (2008) 'Differential expression of GADD45β in normal and osteoarthritic cartilage: Potential role in homeostasis of articular chondrocytes', *Arthritis and Rheumatism*, 58(7), pp. 2075-2087.

Iliopoulos, D., Malizos, K.N. and Tsezou, A. (2007) 'Epigenetic regulation of leptin affects MMP-13 expression in osteoarthritic chondrocytes: possible molecular target for osteoarthritis therapeutic intervention', *Ann Rheum Dis*, 66(12), pp. 1616-21.

Illumina (2009) *De Novo Assembly Using Illumina Reads*. Available at: http://www.ucl.ac.uk/cancer/supportservices/SupportDocs/DeNovoAssembly.pdf.

Illumina (2011) *RNA-seq Data Comparison with Gene Expression Microarrays*. Available at: http://eh.uc.edu/genomics/files/Illumina_Whitepaper_RNASeq_to_arrays_comparison.pdf.

Jacinto, F.V., Ballestar, E. and Esteller, M. (2008) 'Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome', *Biotechniques*, 44(1), pp. 35, 37, 39 passim.

Jakob, M., Demarteau, O., Schafer, D., Stumm, M., Heberer, M. and Martin, I. (2003) 'Enzymatic digestion of adult human articular cartilage yields a small fraction of the total available cells', *Connect Tissue Res*, 44(3-4), pp. 173-80.

Johnson, W.E., Li, C. and Rabinovic, A. (2007) 'Adjusting batch effects in microarray expression data using empirical Bayes methods', *Biostatistics*, 8(1), pp. 118-127.

Kao, W.C., Stevens, K. and Song, Y.S. (2009) 'BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing', *Genome Res*, 19(10), pp. 1884-95.

Kapoor, M., Martel-Pelletier, J., Lajeunesse, D., Pelletier, J.P. and Fahmi, H. (2011) 'Role of proinflammatory cytokines in the pathophysiology of osteoarthritis', *Nat Rev Rheumatol*, 7(1), pp. 33-42.

Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P. and Gingeras, T.R. (2002) 'Large-scale transcriptional activity in chromosomes 21 and 22', *Science*, 296(5569), pp. 916-9.

Karlsson, C., Dehne, T., Lindahl, A., Brittberg, M., Pruss, A., Sittinger, M. and Ringe, J. (2010a) 'Genome-wide expression profiling reveals new candidate genes associated with osteoarthritis', *Osteoarthritis Cartilage*, 18(4), pp. 581-92.

Karlsson, C., Dehne, T., Lindahl, A., Brittberg, M., Pruss, A., Sittinger, M. and Ringe, J. (2010b) 'Genome-wide expression profiling reveals new candidate genes associated with osteoarthritis', *Osteoarthritis and Cartilage*, 18(4), pp. 581-592.

Katz, J.N. (2006) 'Total joint replacement in osteoarthritis', *Best Pract Res Clin Rheumatol*, 20(1), pp. 145-53.

Kennedy, S. and Moran, M. (2010) 'Pharmacological treatment of osteoarthritis of the hip and knee', *BCMJ*, 52(8).

Kerkhof, H.J., Lories, R.J., Meulenbelt, I., Jonsdottir, I., Valdes, A.M., Arp, P., Ingvarsson, T., Jhamai, M., Jonsson, H., Stolk, L., Thorleifsson, G., Zhai, G., Zhang, F., Zhu, Y., van der Breggen, R., Carr, A., Doherty, M., Doherty, S., Felson, D.T., Gonzalez, A., Halldorsson, B.V., Hart, D.J., Hauksson, V.B., Hofman, A., Ioannidis, J.P., Kloppenburg, M., Lane, N.E., Loughlin, J., Luyten, F.P., Nevitt, M.C., Parimi, N., Pols, H.A., Rivadeneira, F., Slagboom, E.P., Styrkarsdottir, U., Tsezou, A., van de Putte, T., Zmuda, J., Spector, T.D., Stefansson, K., Uitterlinden, A.G. and van Meurs, J.B. (2010) 'A genome-wide association study identifies an osteoarthritis susceptibility locus on chromosome 7q22', *Arthritis Rheum*, 62(2), pp. 499-510.

Kevorkian, L., Young, D.A., Darrah, C., Donell, S.T., Shepstone, L., Porter, S., Brockbank, S.M., Edwards, D.R., Parker, A.E. and Clark, I.M. (2004a) 'Expression profiling of metalloproteinases and their inhibitors in cartilage', *Arthritis Rheum*, 50(1), pp. 131-41.

Kevorkian, L., Young, D.A., Darrah, C., Donell, S.T., Shepstone, L., Porter, S., Brockbank, S.M.V., Edwards, D.R., Parker, A.E. and Clark, I.M. (2004b) 'Expression Profiling of Metalloproteinases and Their Inhibitors in Cartilage', *Arthritis and Rheumatism*, 50(1), pp. 131-141.

Khatri, P., Sirota, M. and Butte, A. (2012) 'Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges', *PLoS Comput Biol*, 8(2), p. e1002375.

Kijowski, R., Blankenbaker, D., Stanton, P., Fine, J. and De Smet, A. (2006) 'Arthroscopic Validation of Radiographic Grading Scales of Osteoarthritis of the Tibiofemoral Joint', *American Journal of Roentgenology*, 187(3), pp. 794-799.

Kim, D.W., Nam, S.H., Kim, R.N., Choi, S.H. and Park, H.S. (2010) 'Whole human exome capture for high-throughput sequencing', *Genome*, 53(7), pp. 568-74.

Kim, T.-H., Choi, S.J., Lee, Y.H., Song, G.G. and Ji, J.D. (2012) 'Soluble triggering receptor expressed on myeloid cells-1 as a new therapeutic molecule in rheumatoid arthritis', *Medical Hypotheses*, 78(2), pp. 270-272.

Klesney-Tait, J., Turnbull, I.R. and Colonna, M. (2006) 'The TREM receptor family and signal integration', *Nature Immunology*, 7(12), pp. 1266-1273.

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L. and Wilson, R.K. (2012) 'VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing', *Genome Res*, 22(3), pp. 568-76.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., Taillon, B.E., Chen, Z., Tanzer, A., Saunders, A.C., Chi, J., Yang, F., Carter, N.P., Hurles, M.E., Weissman, S.M., Harkins, T.T., Gerstein, M.B., Egholm, M. and Snyder, M. (2007) 'Paired-end mapping reveals extensive structural variation in the human genome', *Science*, 318(5849), pp. 420-6.

Kotake, S., Sato, K., Kim, K.J., Takahashi, N., Udagawa, N., Nakamura, I., Yamaguchi, A., Kishimoto, T., Suda, T. and Kashiwazaki, S. (1996) 'Interleukin-6 and soluble interleukin-6 receptors in the synovial fluids from rheumatoid arthritis patients are responsible for osteoclast-like cell formation', *J Bone Miner Res*, 11(1), pp. 88-95.

Krueger, F. *Trim Galore!* Available at: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

Kuai, J., Gregory, B., Hill, A., Pittman, D.D., Feldman, J.L., Brown, T., Carito, B., O'Toole, M., Ramsey, R., Adolfsson, O., Shields, K.M., Dower, K., Hall, J.P., Kurdi, Y., Beech, J.T., Nanchahal, J., Feldmann, M., Foxwell, B.M., Brennan, F.M., Winkler, D.G. and Lin, L.L. (2009) 'TREM-1 expression is increased in the synovium of rheumatoid arthritis patients and induces the expression of pro-inflammatory cytokines', *Rheumatology (Oxford, England)*, 48(11), pp. 1352-1358.

Kwan Tat, S., Pelletier, J.-P., Amiable, N., Boileau, C., Lavigne, M. and Martel-Pelletier, J. (2009) 'Treatment with ephrin B2 positively impacts the abnormal metabolism of human osteoarthritic chondrocytes', *Arthritis Research & Therapy*, 11(4), p. R119.

Labaj, P.P., Leparc, G.G., Linggi, B.E., Markillie, L.M., Wiley, H.S. and Kreil, D.P. (2011) 'Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling', *Bioinformatics*, 27(13), pp. i383-91.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001) 'Initial sequencing and analysis of the human genome', *Nature*, 409(6822), pp. 860-921.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) 'Ultrafast and memory-efficient alignment of short DNA sequences to the human genome', *Genome Biol*, 10(3), p. R25.

Leung, Y.F. and Cavalieri, D. (2003) 'Fundamentals of cDNA microarray data analysis', *Trends in Genetics*, 19(11), pp. 649-659.

Levanon, E.Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z.Y., Shoshan, A., Pollock, S.R., Sztybel, D., Olshansky, M., Rechavi, G. and Jantsch, M.F. (2004) 'Systematic identification of abundant A-to-I editing sites in the human transcriptome', *Nat Biotechnol*, 22(8), pp. 1001-5.

Levin, J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A. and Regev, A. (2010) 'Comprehensive comparative analysis of strand-specific RNA sequencing methods', *Nat Methods*, 7(9), pp. 709-15.

Li, C. and Wong, W.H. (2001) 'Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection', *Proceedings of the National Academy of Sciences*, 98(1), pp. 31-36.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, 25(14), pp. 1754-60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics (Oxford, England)*, 25(16), pp. 2078-2079.

Li, H. and Homer, N. (2010) 'A survey of sequence alignment algorithms for next-generation sequencing', *Brief Bioinform*, 11(5), pp. 473-83.

Li, H., Ruan, J. and Durbin, R. (2008) 'Mapping short DNA sequencing reads and calling variants using mapping quality scores', *Genome Res*, 18(11), pp. 1851-8.

Li, M., Wang, I.X., Li, Y., Bruzel, A., Richards, A.L., Toung, J.M. and Cheung, V.G. (2011a) 'Widespread RNA and DNA sequence differences in the human transcriptome', *Science*, 333(6038), pp. 53-8.

Li, X., Li, J., Cheng, K., Lin, Q., Wang, D., Zhang, H., An, H., Gao, M. and Chen, A. (2011b) 'Effect of Low-Intensity Pulsed Ultrasound on MMP-13 and MAPKs Signaling Pathway in Rabbit Knee Osteoarthritis', *Cell Biochemistry and Biophysics*, 61(2), pp. 427-434.

Lisignoli, G., Manferdini, C., Codeluppi, K., Piacentini, A., Grassi, F., Cattini, L., Filardo, G. and Facchini, A. (2009) 'CCL20/CCR6 chemokine/receptor expression in bone tissue from osteoarthritis and rheumatoid arthritis patients: Different response of osteoblasts in the two groups', *Journal of Cellular Physiology*, 221(1), pp. 154-160.

Loeser, R.F. (2008) 'Molecular mechanisms of cartilage destruction in osteoarthritis', *J Musculoskelet Neuronal Interact*, 8(4), pp. 303-6.

Loeser, R.F. (2009) 'Aging and osteoarthritis: the role of chondrocyte senescence and aging changes in the cartilage matrix', *Osteoarthritis Cartilage*, 17(8), pp. 971-9.

Love, M., Anders, S. and Huber, W. (2014) 'Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2', *bioRxiv*.

Madry, H. and Cucchiarini, M. (2013) 'Advances and challenges in gene-based approaches for osteoarthritis', *J Gene Med*, 15(10), pp. 343-55.

Maher, C.A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., Sam, L., Barrette, T., Palanisamy, N. and Chinnaiyan, A.M. (2009) 'Transcriptome sequencing to detect gene fusions in cancer', *Nature*, 458(7234), pp. 97-101.

Mahr, S., Burmester, G.R., Hilke, D., Gobel, U., Grutzkau, A., Haupl, T., Hauschild, M., Koczan, D., Krenn, V., Neidel, J., Perka, C., Radbruch, A., Thiesen, H.J. and Muller, B. (2006) 'Cis- and trans-acting gene regulation is associated with osteoarthritis', *Am J Hum Genet*, 78(5), pp. 793-803.

Malfait, A.M., Arner, E.C., Song, R.H., Alston, J.T., Markosyan, S., Staten, N., Yang, Z., Griggs, D.W. and Tortorella, M.D. (2008) 'Proprotein convertase activation of aggrecanases in cartilage in situ', *Arch Biochem Biophys*, 478(1), pp. 43-51.

Martin, M. (2011) 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *EMBnet.journal*, 17(1), p. 10.

Matthews, B.F. (1953) 'Composition of articular cartilage in osteoarthritis; changes in collagen/chondroitin-sulphate ratio', *Br Med J*, 2(4837), pp. 660-1.

Maurer, P., Hohenadl, C., Hohenester, E., Gohring, W., Timpl, R. and Engel, J. (1995) 'The C-terminal portion of BM-40 (SPARC/osteonectin) is an autonomously folding and crystallisable domain that binds calcium and collagen IV', *J Mol Biol*, 253(2), pp. 347-57.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M.A. (2010) 'The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data', *Genome Res*, 20(9), pp. 1297-303.

McKenna, L.A., Gehrsitz, A., Soder, S., Eger, W., Kirchner, T. and Aigner, T. (2000) 'Effective isolation of high-quality total RNA from human adult articular cartilage', *Anal Biochem*, 286(1), pp. 80-5.

Medvedev, P., Stanciu, M. and Brudno, M. (2009) 'Computational methods for discovering structural variation with next-generation sequencing', *Nat Methods*, 6(11 Suppl), pp. S13-20.

Meissner, A., Gnirke, A., Bell, G.W., Ramsahoye, B., Lander, E.S. and Jaenisch, R. (2005) 'Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis', *Nucleic Acids Res*, 33(18), pp. 5868-77.

Menendez, M.I., Clark, D.J., Carlton, M., Flanigan, D.C., Jia, G., Sammet, S., Weisbrode, S.E., Knopp, M.V. and Bertone, A.L. (2011) 'Direct delayed human adenoviral BMP-2 or BMP-6 gene therapy for bone and cartilage regeneration in a pony osteochondral model', *Osteoarthritis Cartilage*, 19(8), pp. 1066-75.

Meng, J., Ma, X., Ma, D. and Xu, C. (2005) 'Microarray analysis of differential gene expression in temporomandibular joint condylar cartilage after experimentally induced osteoarthritis', *Osteoarthritis Cartilage*, 13(12), pp. 1115-25.

'The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements', (2006) *Nat Biotech*, 24(9), pp. 1151-1161.

Miklos, G.L.G. and Maleszka, R. (2004) 'Microarray reality checks in the context of a complex disease', *Nature Biotechnology*, 22(5), pp. 615-621.

Milner, J.M., Patel, A., Davidson, R.K., Swingler, T.E., Desilets, A., Young, D.A., Kelso, E.B., Donell, S.T., Cawston, T.E., Clark, I.M., Ferrell, W.R., Plevin, R., Lockhart, J.C., Leduc, R. and Rowan, A.D. (2010) 'Matriptase is a novel initiator of cartilage matrix degradation in osteoarthritis', *Arthritis Rheum*, 62(7), pp. 1955-66.

Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P., Ingle, C., Nisbett, J., Guigo, R. and Dermitzakis, E.T. (2010) 'Transcriptome genetics using second generation sequencing in a Caucasian population', *Nature*, 464(7289), pp. 773-7.

Morgan, M., Anders, S., Lawrence, M., Aboyoun, P., Pages, H. and Gentleman, R. (2009) 'ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data', *Bioinformatics*, 25(19), pp. 2607-8.

Morin, R.D., O'Connor, M.D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M., Eaves, C.J. and Marra, M.A. (2008) 'Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells', *Genome Res*, 18(4), pp. 610-21.

Morko, J.P., Soderstrom, M., Saamanen, A.M., Salminen, H.J. and Vuorio, E.I. (2004) 'Up regulation of cathepsin K expression in articular chondrocytes in a transgenic mouse model for osteoarthritis', *Ann Rheum Dis*, 63(6), pp. 649-55.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) 'Mapping and quantifying mammalian transcriptomes by RNA-Seq', *Nat Methods*, 5(7), pp. 621-8.

Mototani, H., Iida, A., Nakajima, M., Furuichi, T., Miyamoto, Y., Tsunoda, T., Sudo, A., Kotani, A., Uchida, A., Ozaki, K., Tanaka, Y., Nakamura, Y., Tanaka, T., Notoya, K. and Ikegawa, S. (2008) 'A functional SNP in EDG2 increases susceptibility to knee osteoarthritis in Japanese', *Hum Mol Genet*, 17(12), pp. 1790-7.

Mutter, G.L., Zahrieh, D., Liu, C., Neuberg, D., Finkelstein, D., Baker, H.E. and Warrington, J.A. (2004) 'Comparison of frozen and RNALater solid tissue storage methods for use in RNA expression microarrays', *BMC Genomics*, 5, p. 88.

Nakamura, S., Kamihagi, K., Satakeda, H., Katayama, M., Pan, H., Okamoto, H., Noshiro, M., Takahashi, K., Yoshihara, Y., Shimmei, M., Okada, Y. and Kato, Y. (1996) 'Enhancement of SPARC (osteonectin) synthesis in arthritic cartilage. Increased levels in synovial fluids from patients with rheumatoid arthritis and regulation by growth factors and cytokines in chondrocyte cultures', *Arthritis Rheum*, 39(4), pp. 539-51.

Nakamura, T., Shi, D., Tzetis, M., Rodriguez-Lopez, J., Miyamoto, Y., Tsezou, A., Gonzalez, A., Jiang, Q., Kamatani, N., Loughlin, J. and Ikegawa, S. (2007) 'Meta-analysis of association between the ASPN D-repeat and osteoarthritis', *Hum Mol Genet*, 16(14), pp. 1676-81.

Nakase, T., Miyaji, T., Tomita, T., Kaneko, M., Kuriyama, K., Myoui, A., Sugamoto, K., Ochi, T. and Yoshikawa, H. (2003) 'Localization of bone morphogenetic protein-2 in human osteoarthritic cartilage and osteophyte', *Osteoarthritis Cartilage*, 11(4), pp. 278-84.

Nanba, Y., Nishida, K., Yoshikawa, T., Sato, T., Inoue, H. and Kuboki, Y. (1997) 'Expression of osteonectin in articular cartilage of osteoarthritic knees', *Acta Med Okayama*, 51(5), pp. 239-43.

Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., Shendure, J. and Bamshad, M.J. (2010) 'Exome sequencing identifies the cause of a mendelian disorder', *Nat Genet*, 42(1), pp. 30-5.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., Bamshad, M., Nickerson, D.A. and Shendure, J. (2009) 'Targeted capture and massively parallel sequencing of 12 human exomes', *Nature*, 461(7261), pp. 272-6.

Nyren, P., Pettersson, B. and Uhlen, M. (1993) 'Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay', *Anal Biochem*, 208(1), pp. 171-5.

Oldham, M., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S. and Geschwind, D. (2008) 'Functional organization of the transcriptome in human brain', *Nature neuroscience*, 11(11), pp. 1271-1282.

Onishi, R.M. and Gaffen, S.L. (2010) 'Interleukin-17 and its target genes: Mechanisms of interleukin-17 function in disease', *Immunology*, 129(3), pp. 311-321.

Oshlack, A. and Wakefield, M.J. (2009) 'Transcript length bias in RNA-seq data confounds systems biology', *Biol Direct*, 4, p. 14.

Ozsolak, F., Goren, A., Gymrek, M., Guttman, M., Regev, A., Bernstein, B.E. and Milos, P.M. (2010) 'Digital transcriptome profiling from attomole-level RNA samples', *Genome Res*, 20(4), pp. 519-25.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J. and Blencowe, B.J. (2008) 'Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing', *Nat Genet*, 40(12), pp. 1413-5.

Parker, A.E., Boutell, J., Carr, A. and Maciewicz, R.A. (2002) 'Novel cartilage-specific splice variants of fibronectin', *Osteoarthritis Cartilage*, 10(7), pp. 528-34.

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H. and Soldatov, A. (2009) 'Transcriptome analysis by strand-specific sequencing of complementary DNA', *Nucleic Acids Res*, 37(18), p. e123.

Pelletier, J.P., Martel-Pelletier, J. and Abramson, S.B. (2001) 'Osteoarthritis, an inflammatory disease: potential implication for the selection of new therapeutic targets', *Arthritis Rheum*, 44(6), pp. 1237-47.

Peng, Z., Cheng, Y., Tan, B.C., Kang, L., Tian, Z., Zhu, Y., Zhang, W., Liang, Y., Hu, X., Tan, X., Guo, J., Dong, Z., Liang, Y., Bao, L. and Wang, J. (2012) 'Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome', *Nat Biotechnol*, 30(3), pp. 253-60.

Pepke, S., Wold, B. and Mortazavi, A. (2009) 'Computation for ChIP-seq and RNA-seq studies', *Nat Methods*, 6(11 Suppl), pp. S22-32.

Perkins, T.T., Kingsley, R.A., Fookes, M.C., Gardner, P.P., James, K.D., Yu, L., Assefa, S.A., He, M., Croucher, N.J., Pickard, D.J., Maskell, D.J., Parkhill, J., Choudhary, J., Thomson, N.R. and Dougan, G. (2009) 'A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhi', *PLoS Genet*, 5(7), p. e1000569.

Picardi, E., Horner, D.S., Chiara, M., Schiavon, R., Valle, G. and Pesole, G. (2010) 'Large-scale detection and analysis of RNA editing in grape mtDNA by RNA deep-sequencing', *Nucleic Acids Res*, 38(14), pp. 4755-67.

Picardi, E. and Pesole, G. (2013) 'REDItools: high-throughput RNA editing detection made easy', *Bioinformatics*, 29(14), pp. 1813-4.

Quackenbush, J. (2002) 'Microarray data normalization and transformation', *Nat Genet*, 32 Suppl, pp. 496-501.

Quinlan, A.R. and Hall, I.M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26(6), pp. 841-842.

R Core Team (2012) *R: A Language and Environment for Statistical Computing* [Computer program]. R Foundation for Statistical Computing. Available at: http://www.R-project.org/.

Radwan, M., Gavriilidis, C., Robinson, J.H., Davidson, R., Clark, I.M., Rowan, A.D. and Young, D.A. (2013) 'Matrix metalloproteinase 13 expression in response to double-stranded RNA in human chondrocytes', *Arthritis Rheum*, 65(5), pp. 1290-301.

Raha, D., Hong, M. and Snyder, M. (2010) 'ChIP-Seq: a method for global identification of regulatory elements in the genome', *Curr Protoc Mol Biol*, Chapter 21, pp. Unit 21.19.1-14.

Raine, E.V., Dodd, A.W., Reynard, L.N. and Loughlin, J. (2013) 'Allelic expression analysis of the osteoarthritis susceptibility gene COL11A1 in human joint tissues', *BMC Musculoskelet Disord*, 14, p. 85.

Ramaswami, G., Zhang, R., Piskol, R., Keegan, L.P., Deng, P., O'Connell, M.A. and Li, J.B. (2013) 'Identifying RNA editing sites using RNA sequencing data alone', *Nat Methods*, 10(2), pp. 128-32.

Ratnayake, M., Reynard, L.N., Raine, E.V., Santibanez-Koref, M. and Loughlin, J. (2012) 'Allelic expression analysis of the osteoarthritis susceptibility locus that maps to MICAL3', *BMC Med Genet*, 13, p. 12.

Rehrauer, H., Opitz, L., Tan, G., Sieverling, L. and Schlapbach, R. (2013) 'Blind spots of quantitative RNA-seq: the limits for assessing abundance, differential expression, and isoform switching', *BMC Bioinformatics*, 14, p. 370.

Reynard, L.N. and Loughlin, J. (2012) 'Genetics and epigenetics of osteoarthritis', *Maturitas*, 71(3), pp. 200-4.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics*, 26(1), pp. 139-40.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. and Nyren, P. (1996) 'Real-time DNA sequencing using detection of pyrophosphate release', *Anal Biochem*, 242(1), pp. 84-9.

Roughley, P.J. (2001) 'Articular cartilage and changes in arthritis: noncollagenous proteins and proteoglycans in the extracellular matrix of cartilage', *Arthritis Res*, 3(6), pp. 342-7.

Rowan, A.D., Litherland, G.J., Hui, W. and Milner, J.M. (2008) 'Metalloproteases as potential therapeutic targets in arthritis treatment', *Expert Opinion on Therapeutic Targets*, 12(1), pp. 1-18.

Rowan, A.D. and Young, D.A. (2007) 'Collagenase gene regulation by pro-inflammatory cytokines in cartilage', *Frontiers in Bioscience*, 12(2), pp. 536-550.

Ruettger, A., Neumann, S., Wiederanders, B. and Huber, R. (2010) 'Comparison of different methods for preparation and characterization of total RNA from cartilage samples to uncover osteoarthritis in vivo', *BMC Res Notes*, 3, p. 7.

Ruffalo, M., LaFramboise, T. and Koyuturk, M. (2011) 'Comparative analysis of algorithms for next-generation sequencing read alignment', *Bioinformatics*, 27(20), pp. 2790-6.

Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., Hunt, S.E., Cole, C.G., Coggill, P.C., Rice, C.M., Ning, Z., Rogers, J., Bentley, D.R., Kwok, P.Y., Mardis, E.R., Yeh, R.T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R.H., McPherson, J.D., Gilman, B., Schaffner, S., Van Etten, W.J., Reich, D., Higgins, J., Daly, M.J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M.C., Linton, L., Lander, E.S. and Altshuler, D. (2001) 'A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms', *Nature*, 409(6822), pp. 928-33.

Salminen-Mankonen, H., Saamanen, A.M., Jalkanen, M., Vuorio, E. and Pirila, L. (2005) 'Syndecan-1 expression is upregulated in degenerating articular cartilage in a transgenic mouse model for osteoarthritis', *Scand J Rheumatol*, 34(6), pp. 469-74.

Sam, L.T., Lipson, D., Raz, T., Cao, X., Thompson, J., Milos, P.M., Robinson, D., Chinnaiyan, A.M., Kumar-Sinha, C. and Maher, C.A. (2011) 'A comparison of single molecule and amplification based sequencing of cancer transcriptomes', *PLoS One*, 6(3), p. e17305.

Sampson, T.G. (2011) 'Arthroscopic treatment for chondral lesions of the hip', *Clin Sports Med*, 30(2), pp. 331-48.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M. and Smith, M. (1977) 'Nucleotide sequence of bacteriophage phi X174 DNA', *Nature*, 265(5596), pp. 687-95.

Sato, T., Konomi, K., Yamasaki, S., Aratani, S., Tsuchimochi, K., Yokouchi, M., Masuko-Hongo, K., Yagishita, N., Nakamura, H., Komiya, S., Beppu, M., Aoki, H., Nishioka, K. and Nakajima, T. (2006) 'Comparative analysis of gene expression profiles in intact and damaged regions of human osteoarthritic cartilage', *Arthritis Rheum*, 54(3), pp. 808-17.

Schena, M., Shalon, D., Davis, R. and Brown, P. (1995) 'Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray', *Science*, 270(5235), pp. 467-470.

Schliesky, S., Gowik, U., Weber, A.P. and Brautigam, A. (2012) 'RNA-Seq Assembly - Are We There Yet?', *Front Plant Sci*, 3, p. 220.

Schmidt, D., Stark, R., Wilson, M.D., Brown, G.D. and Odom, D.T. (2008) 'Genome-scale validation of deep-sequencing libraries', *PLoS One*, 3(11), p. e3713.

Schones, D.E., Cui, K., Cuddapah, S., Roh, T.Y., Barski, A., Wang, Z., Wei, G. and Zhao, K. (2008) 'Dynamic regulation of nucleosome positioning in the human genome', *Cell*, 132(5), pp. 887-98.

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S., Menzel, W., Granzow, M. and Ragg, T. (2006) 'The RIN: an RNA integrity number for assigning integrity values to RNA measurements', *BMC Mol Biol*, 7, p. 3.

Schulz, M.H., Zerbino, D.R., Vingron, M. and Birney, E. (2012) 'Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels', *Bioinformatics*, 28(8), pp. 1086-92.

Scott, J.L., Gabrielides, C., Davidson, R.K., Swingler, T.E., Clark, I.M., Wallis, G.A., Boot-Handford, R.P., Kirkwood, T.B., Taylor, R.W. and Young, D.A. (2010a) 'Superoxide dismutase downregulation in osteoarthritis progression and end-stage disease', *Ann Rheum Dis*, 69(8), pp. 1502-10.

Scott, J.L., Gabrielides, C., Davidson, R.K., Swingler, T.E., Clark, I.M., Wallis, G.A., Boot-Handford, R.P., Kirkwood, T.B.L., Talyor, R.W. and Young, D.A. (2010b) 'Superoxide dismutase downregulation in osteoarthritis progression and end-stage disease', *Annals of the Rheumatic Diseases*, 69(8), pp. 1502-1510.

Serre, D., Lee, B.H. and Ting, A.H. (2010) 'MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome', *Nucleic Acids Res*, 38(2), pp. 391-9.

Seyednasrollah, F., Laiho, A. and Elo, L.L. (2013) 'Comparison of software packages for detecting differential expression in RNA-seq studies', *Brief Bioinform*.

Shendure, J. and Ji, H. (2008) 'Next-generation DNA sequencing', *Nat Biotechnol*, 26(10), pp. 1135-45.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D. and Church, G.M. (2005) 'Accurate multiplex polony sequencing of an evolved bacterial genome', *Science*, 309(5741), pp. 1728-32.

Sipe, J.D. (1995) 'Acute-phase proteins in osteoarthritis', *Semin Arthritis Rheum*, 25(2), pp. 75-86.

Southam, L., Rodriguez-Lopez, J., Wilkins, J.M., Pombo-Suarez, M., Snelling, S., Gomez-Reino, J.J., Chapman, K., Gonzalez, A. and Loughlin, J. (2007) 'An SNP in the 5'-UTR of GDF5 is associated with osteoarthritis susceptibility in Europeans and with in vivo differences in allelic expression in articular cartilage', *Hum Mol Genet*, 16(18), pp. 2226-32.

Stecher, R.M. and Hersh, A.H. (1944) 'HEBERDEN'S NODES: THE MECHANISM OF INHERITANCE IN HYPERTROPHIC ARTHRITIS OF THE FINGERS', *J Clin Invest*, 23(5), pp. 699-704.

Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E. and Mesirov, J. (2005a) 'Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles', *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), pp. 15545-15550.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. (2005b) 'Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles', *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), pp. 15545-15550.

Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H. and Yaspo, M.L. (2008) 'A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome', *Science*, 321(5891), pp. 956-60.

Swingler, T., Waters, J., Davidson, R., Pennington, C., Puente, X., Darrah, C., Cooper, A., Donell, S., Guile, G., Wang, W. and Clark, I. (2009a) 'Degradome expression profiling in human articular cartilage', *Arthritis Research & Therapy*, 11(3), p. R96.

Swingler, T.E., Waters, J.G., Davidson, R.K., Pennington, C.J., Puente, X.S., Darrah, C., Cooper, A., Donell, S.T., Guile, G.R., Wang, W. and Clark, I.M. (2009b) 'Degradome expression profiling in human articular cartilage', *Arthritis Res Ther*, 11(3), p. R96.

Syddall, C.M., Reynard, L.N., Young, D.A. and Loughlin, J. (2013) 'The identification of trans-acting factors that regulate the expression of GDF5 via the osteoarthritis susceptibility SNP rs143383', *PLoS Genet*, 9(6), p. e1003557.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L. and von Mering, C. (2011) 'The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored', *Nucleic acids research*, 39(Database issue), pp. D561-D568.

Takada, K., Hirose, J., Senba, K., Yamabe, S., Oike, Y., Gotoh, T. and Mizuta, H. (2011) 'Enhanced apoptotic and reduced protective response in chondrocytes following endoplasmic reticulum stress in osteoarthritic cartilage', *Int J Exp Pathol*, 92(4), pp. 232-42.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., Lao, K. and Surani, M.A. (2009) 'mRNA-Seq whole-transcriptome analysis of a single cell', *Nat Methods*, 6(5), pp. 377-82.

Tang, F., Lao, K. and Surani, M.A. (2011) 'Development and applications of single-cell transcriptome analysis', *Nat Methods*, 8(4 Suppl), pp. S6-11.

Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A. and Conesa, A. (2011) 'Differential expression in RNA-seq: a matter of depth', *Genome Res*, 21(12), pp. 2213-23.

Termine, J.D., Kleinman, H.K., Whitson, S.W., Conn, K.M., McGarvey, M.L. and Martin, G.R. (1981) 'Osteonectin, a bone-specific protein linking mineral to collagen', *Cell*, 26(1 Pt 1), pp. 99-105.

Tew, S.R., McDermott, B.T., Fentem, R.B., Peffers, M.J. and Clegg, P.D. (2014) 'Transcriptome-wide analysis of mRNA decay in normal and osteoarthritic human articular chondrocytes', *Arthritis Rheumatol*.

Thu, K.L., Pikor, L.A., Kennett, J.Y., Alvarez, C.E. and Lam, W.L. (2010) 'Methylation analysis by DNA immunoprecipitation', *J Cell Physiol*, 222(3), pp. 522-31.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L. and Pachter, L. (2013) 'Differential analysis of gene regulation at transcript resolution with RNA-seq', *Nat Biotechnol*, 31(1), pp. 46-53.

Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) 'TopHat: discovering splice junctions with RNA-Seq', *Bioinformatics*, 25(9), pp. 1105-11.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L. (2012) 'Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks', *Nat Protoc*, 7(3), pp. 562-78.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. (2010) 'Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation', *Nat Biotechnol*, 28(5), pp. 511-5.

Valdes, A.M. and Spector, T.D. (2010) 'The clinical relevance of genetic susceptibility to osteoarthritis', *Best Pract Res Clin Rheumatol*, 24(1), pp. 3-14.

Valdes, A.M., Spector, T.D., Tamm, A., Kisand, K., Doherty, S.A., Dennison, E.M., Mangino, M., Tamm, A., Kerna, I., Hart, D.J., Wheeler, M., Cooper, C., Lories, R.J., Arden, N.K. and Doherty, M. (2010) 'Genetic variation in the SMAD3 gene is associated with hip and knee osteoarthritis', *Arthritis Rheum*, 62(8), pp. 2347-52.

Velasco, J., Zarrabeitia, M.T., Prieto, J.R., Perez-Castrillon, J.L., Perez-Aguilar, M.D., Perez-Nunez, M.I., Sanudo, C., Hernandez-Elena, J., Calvo, I., Ortiz, F., Gonzalez-Macias, J. and Riancho, J.A. (2010a) 'Wnt pathway genes in osteoporosis and osteoarthritis: differential expression and genetic association study', *Osteoporos Int*, 21(1), pp. 109-18.

Velasco, J., Zarrabeitia, M.T., Prieto, J.R., Perez-Castrillon, J.L., Perez-Aguilar, M.D., Perez-Nuñez, M.I., Sañudo, C., Hernandez-Elena, J., Calvo, I., Ortiz, F., Gonzalez-Macias, J. and Riancho, J.A. (2010b) 'Wnt pathway genes in osteoporosis and osteoarthritis: Differential expression and genetic association study', *Osteoporosis International*, 21(1), pp. 109-118.

Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) 'Serial analysis of gene expression', *Science*, 270(5235), pp. 484-7.

Vijay, N., Poelstra, J.W., Kunstner, A. and Wolf, J.B. (2013) 'Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments', *Mol Ecol*, 22(3), pp. 620-34.

Wainwright, S.D., Bondeson, J. and Hughes, C.E. (2006) 'An alternative spliced transcript of ADAMTS4 is present in human synovium from OA patients', *Matrix Biol*, 25(5), pp. 317-20.

Walsh, T., Shahin, H., Elkan-Miller, T., Lee, M.K., Thornton, A.M., Roeb, W., Abu Rayyan, A., Loulus, S., Avraham, K.B., King, M.C. and Kanaan, M. (2010) 'Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82', *Am J Hum Genet*, 87(1), pp. 90-4.

Wang, Z., Gerstein, M. and Snyder, M. (2009) 'RNA-Seq: a revolutionary tool for transcriptomics', *Nat Rev Genet*, 10(1), pp. 57-63.

Werner, T. (2008) 'Bioinformatics applications for pathway analysis of microarray data', *Current Opinion in Biotechnology*, 19(1), pp. 50-54.

Woessner, J.F., Jr. (1991) 'Matrix metalloproteinases and their inhibitors in connective tissue remodeling', *Faseb j*, 5(8), pp. 2145-54.

Wu, C., Carta, R. and Zhang, L. (2005) 'Sequence dependence of cross-hybridization on short oligo microarrays', *Nucleic acids research*, 33(9), pp. e84-e84.

Wu, S., Li, C., Huang, W., Li, W. and Li, R.W. (2012) 'Alternative splicing regulated by butyrate in bovine epithelial cells', *PLoS One*, 7(6), p. e39182.

Wu, T.D. and Nacu, S. (2010) 'Fast and SNP-tolerant detection of complex variants and splicing in short reads', *Bioinformatics*, 26(7), pp. 873-81.

Xu, C., Houck, J.R., Fan, W., Wang, P., Chen, Y., Upton, M., Futran, N.D., Schwartz, S.M., Zhao, L.P., Chen, C. and Mendez, E. (2008) 'Simultaneous isolation of DNA and RNA from the same cell population obtained by laser capture microdissection for genome and transcriptome profiling', *J Mol Diagn*, 10(2), pp. 129-34.

Xue, Y., Wang, Q., Long, Q., Ng, B.L., Swerdlow, H., Burton, J., Skuce, C., Taylor, R., Abdellah, Z., Zhao, Y., MacArthur, D.G., Quail, M.A., Carter, N.P., Yang, H. and Tyler-Smith, C. (2009) 'Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree', *Curr Biol*, 19(17), pp. 1453-7.

Yatsugi, N., Tsukazaki, T., Osaki, M., Koji, T., Yamashita, S. and Shindo, H. (2000) 'Apoptosis of articular chondrocytes in rheumatoid arthritis and osteoarthritis: Correlation of apoptosis with degree of cartilage destruction and expression of apoptosis-related proteins of p53 and c-myc', *Journal of Orthopaedic Science*, 5(2), pp. 150-156.

Young, R.S., Marques, A.C., Tibbit, C., Haerty, W., Bassett, A.R., Liu, J.L. and Ponting, C.P. (2012) 'Identification and properties of 1,119 candidate lincRNA loci in the Drosophila melanogaster genome', *Genome Biol Evol*, 4(4), pp. 427-42.

Yudoh, K., van Trieu, N., Nakamura, H., Hongo-Masuko, K., Kato, T. and Nishioka, K. (2005) 'Potential involvement of oxidative stress in cartilage senescence and development of osteoarthritis: oxidative stress induces chondrocyte telomere instability and downregulation of chondrocyte function', *Arthritis Res Ther*, 7(2), pp. R380-R391.

Zeggini, E., Panoutsopoulou, K., Southam, L., Rayner, N.W., Day-Williams, A.G., Lopes, M.C., Boraska, V., Esko, T., Evangelou, E., Hoffman, A., Houwing-Duistermaat, J.J., Ingvarsson, T., Jonsdottir, I., Jonnson, H., Kerkhof, H.J., Kloppenburg, M., Bos, S.D., Mangino, M., Metrustry, S., Slagboom, P.E., Thorleifsson, G., Raine, E.V., Ratnayake, M., Ricketts, M., Beazley, C., Blackburn, H., Bumpstead, S.,

Elliott, K.S., Hunt, S.E., Potter, S.C., Shin, S.Y., Yadav, V.K., Zhai, G., Sherburn, K., Dixon, K., Arden, E., Aslam, N., Battley, P.K., Carluke, I., Doherty, S., Gordon, A., Joseph, J., Keen, R., Koller, N.C., Mitchell, S., O'Neill, F., Paling, E., Reed, M.R., Rivadeneira, F., Swift, D., Walker, K., Watkins, B., Wheeler, M., Birrell, F., Ioannidis, J.P., Meulenbelt, I., Metspalu, A., Rai, A., Salter, D., Stefansson, K., Stykarsdottir, U., Uitterlinden, A.G., van Meurs, J.B., Chapman, K., Deloukas, P., Ollier, W.E., Wallis, G.A., Arden, N., Carr, A., Doherty, M., McCaskie, A., Willkinson, J.M., Ralston, S.H., Valdes, A.M., Spector, T.D. and Loughlin, J. (2012) 'Identification of new susceptibility loci for osteoarthritis (arcOGEN): a genome-wide association study', *Lancet*, 380(9844), pp. 815-23.

Zerbino, D.R. and Birney, E. (2008) 'Velvet: algorithms for de novo short read assembly using de Bruijn graphs', *Genome Res*, 18(5), pp. 821-9.

Zhang, H., Marshall, K.W., Tang, H., Hwang, D.M., Lee, M. and Liew, C.C. (2003) 'Profiling genes expressed in human fetal cartilage using 13,155 expressed sequence tags', *Osteoarthritis Cartilage*, 11(5), pp. 309-19.

Zhang, Q., Hui, W., Litherland, G.J., Barter, M.J., Davidson, R., Darrah, C., Donell, S.T., Clark, I.M., Cawston, T.E., Robinson, J.H., Rowan, A.D. and Young, D.A. (2008) 'Differential Toll-like receptor-dependent collagenase expression in chondrocytes', *Ann Rheum Dis*, 67(11), pp. 1633-41.