Comprehensive NGS analysis and integration of high throughput data in ageing and disease

Anela Tosevska

September 30, 2016

Date Performed:	January 1, 2012
Project Supervisor:	Matteo Pellegrini, Ph.D
Thesis Supervisor:	Karl-Heinz Wagner, PhD

Abstract

Ageing is a combination of multifactorial changes that occur to an organism throughout time. Large number of studies have tried to define the exact mechanisms of the process, dealing with isolated phenomena that occur during ageing. However, most analyses are very limiting, either focusing on a specific pathway or quantitative trait, while missing a multitude of interactions. The era of next generation techniques has brought us the promised step forward in overcoming these limitations, allowing the complete catalogue of a cell or tissues macromolecules to be considered for analysis.

Here I present a framework for analysis of transcriptomic data obtained by next generation RNA sequencing of muscle biopsies obtained within the Vienna Active Ageing Study (VAAS), a large collaborative study conducted from 2011 to 2013 in Vienna.

We have obtained a set of results that can be directly used, in combination with other parameters, to test our hypothesis that a lifestyle intervention can improve the transcriptome signature in the ageing muscle. This will be useful in assessing whether regular exercise might delay the onset of sarcopenia and improve the general fitness of an individual. The knowledge can be then translated into creating guidelines for creating training protocols specifically for this age group in order to prevent sarcopenia and improve the overall quality of life.

Contents

1 Introduction

Ageing is defined as a progressive loss of body functions with decreasing fertility and increasing morbidity and mortality. As the organisms age their normal functions decline, mainly as a result of accumulated molecular damage in their cells and tissues. Up to now, there have been numerous attempts to explain the molecular mechanisms of ageing. What is common in all of them is the increased production of ROS with decreased defensive capacity, telomere attrition and mitochondrial dysfunction.

Furthermore, ageing is generally not a devastating condition by itself; however it is highly associated to number of diseases, such as metabolic syndrome, type 2 diabetes mellitus, cardiovascular diseases, cancer, cognitive decline, sarcopenia, osteoporosis etc. These conditions are generally referred to as age-related diseases. It is known that the susceptibility of an individual for developing one of these conditions is not only genetically determined, but depends on number of environmental factors. Lifestyle factors, such as nutrition, exercise, smoking etc, are implied to have the highest influence on an individuals healthy ageing. However, there is no firm evidence on how these factors are involved in the ageing process and to which extent they control the molecular mechanisms of ageing [McLean and Le Couteur, 2004].

Next Generation Sequencing (NGS) has emerged as a powerful technique with a great potential for uncovering characteristics of organisms potentially related to accelerated ageing or higher life expectancy, such as susceptibility to disease or resistance to certain external influences etc. This information can be revealed upon analysis of the genome of different organisms and variations in the genetic code of individuals or subgroups in the population, called genetic variants, but also by assessing the transcriptional activity of different cells and tissues. The transcriptional activity of the cell is very dynamic and specific for a certain cell type, developmental stage and condition. As such, capturing the transcriptional changes upon different conditions can unravel mechanisms of both cell function and dysfunction. Qualitative and quantitative analysis of mRNA has started with the labour intensive and low throughput Nortern Blotting , to quantitative polimerase chain reaction for relative expression analysis and large scale microarray assays [Ding et al., 2007]. Ever since, RNA sequencing has developed as a faster and more reliable technique, all in order to produce the same or improved result output using a less time-consuming process with lower error rates. In fact, with the possibility to sequence and assemble complete genomes or transcriptomes, NGS has allowed for detection of even minor variations in the genome of different individuals of the same species, such as single-nucleotide polymorphisms (SNPs), which were often found to be associated with a number of diseases [Hutchison, 2007]. In the last years, due to the higher availability

and affordability, RNAseq has grown into a standard method for transcriptome analysis, opening up new possibilities for large-scale data integration and interpretation.

An interesting aspect of dealing with RNA extracted from tissues is the presence of several different cell types in a certain tissue. These cell types often have different set of genes expressed at a higher level whereas other genes are either not expressed or downregulated. In addition, different conditions and developmental stages can "turn on" and "turn off" different genes. This can often lead to errors in interpretation of the obtained results as apparent changes in gene expression often just reflect a shift in cell populations or difference in sample quality. Tissue deconvolution can be performed using RNAseq data unraveling how signatures differ between cell types, and hence discriminating between real changes in expression and population shifts. With various protocols it is possible to sequence whole exomes, to find protein-DNA interaction sites (ChIP-seq) and most importantly RNA-sequencing sheds light on transcriptional patterns (reviewed by Buermans and Den Dunnen [2014], van Dijk et al. [2014]).

The most successful among Next generation sequencers are the products of Roche and Illumina or Solexa, which dominated the markets with their platforms. The methods are roughly based on the detection of a light signal upon the incorporation of a nucleotide during a PCR-like amplification step. The main benefits from NGS are the possibility to obtain millions of sequence reads for each run in a short time, along with relatively little costs (reviewed by van Dijk et al. [2014]). After the filtering and trimming steps of the sequences according to their quality, the first step of a workflow, only working with low computational power, aiming the calculation of a genome is usually the indexing of the reference genome as it is compressing its size by creating patters being present a few times in the genome. That is necessary as it allows a faster alignment later on. Such indexing can be done either by Bowtie 2 [Langmead and Salzberg, 2012], which can do both, the indexing by using an algorithm based on the Burrow-Wheeler-Algorithm and the alignment or only with the BWA algorithm and another tool of choice [Li and Durbin, 2009]. After the alignment, the data still has to be processed in order to get rid of e.g. alignment artifacts. These tasks can be conveniently performed by samtools [Li et al., 2009] or the Genome Analysis Toolkit (GATK - [McKenna et al., 2010]).

A typical protocol for analysis of RNAseq datasets starts with mapping the sequence reads to a reference genome. A very fast and memory-saving alignment program is e.g. Bowtie, with the main downside that it does not allow for big gaps (deriving e.g. from introns) in the alignment. Therefore, TopHat was

developed to break down unmappable reads into smaller parts, thus rendering them mappable and in addition it discovers possible splice sites. The main reasons for choosing TopHat instead of the, more widely used for genomic DNA, Burrows-Wheeler Aligner in RNAseq is the fact that BWA does not consider splicing events [Borozan et al., 2013], whereas TopHat is a spliced read aligner [Trapnell et al., 2009], thus more suitable for RNAseq analyses. Finally, once the reads have been aligned to the prospective positions, it is time to evaluate whether differences in expression between different conditions can be detected. The entire RNAseq workflow and the latest updates have been thoroughly explained on RNA-seqlopedia [1] as well as by Griffith et al. [2015].

The biological interpretation of RNA seq data depends highly on the problem being investigated, as well as the design and statistical approach. Unlike clinical data where most tests are performed routinely in blood samples, tissue biopsies can pose an increased challenge, as they are more difficult to obtain, require adaptation of experimental protocols and there is often scarce information in the published literature. Hence, one of the aims of this project was to bring big data analysis with NGS and data interpretation on the effect of physical activity on muscle transciptome.

With regards to physical activity, it has been shown previously, that even moderate resistance training can reduce morbidity and mortality and prevent from development sarcopenia and dynapenia. Muscle strength can be improved even at very advanced age. For instance, one study showed that 24 weeks of strength training can lead to a `15.6%` increase in muscle strength in 70 year-olds, whereas the control lost `0.6%` of their baseline strength [Rabelo et al., 2011]. Another study reported of a —verb—41-47after 12 weeks of resistance training. A systematic review evaluated that an evidence-based training protocol, consisting of balance training and progressive resistance training at mid to high intensity, can influence physical fitness, functional health and quality of life in institutionalized elderly individuals in a positive manner [Weening-Dijksterhuis et al., 2011]. Similar results were observed employing a meta-analysis, showing that the efficiency of strength training is more meaningful with higher training intensity [Peterson et al., 2010]. One common assumption, is that a combination of protein and micromolecular-based nutrients combined with strength training could be beneficial. One study, for instance, has investigating the effect of strength training in a combination with with a carbohydrate/ amino acid supplementation on functional parameters and physical performance in elderly participants. The participants were provided the carbohydrate drink before training and the amino acid supplement after. However, in a lack of a control group, the improvementcould not be atributed to a certain type of intervention [Onambélé-Pearson

[1] `www.rnaseq.uoregon.edu`

et al., 2010]. No additional effect of ingesting 10g of protein before and also after resistance training was observed in elderly men (722 years) after 12 weeks of training [Verdijk et al., 2009]. Another positive effect of strength training, besides increased muscle protein synthesis, could be a decrease of oxidative and inflammatory stress, all of which are strongly associated with the aging process.

The aim of the present study is to investigate the effect of progressive resistance training with or without consuming a commercially available nutritional supplement for the elderly on clinical functional parameter (muscle mass and function, physical performance) institutionalized elderly. The data used within the current project originate from the Vienna Active Ageing Study (VAAS). It was a collaborative randomized intervention trial that span over a period of 18 months aiming to assess the effect of regular exercise on biomarkers of ageing and physical fitness in institutionalized elderly (6598 years of age). The study participants were predominantly females. A number of parameters have already been analyzed individually and the outcomes have been reported [Franzke et al., 2014, Oesen et al., 2015]. In addition to numerous anthropological and functional measurements, blood, urine and saliva samples collected within the study, a subset of muscle biopsies have been obtained for whole genome RNA sequencing. The resulting dataset has been approached from two directions. First, by looking at a baseline set of coding or non-coding RNAs in the ageing muscle. Second, looking at the effect of resistance training and supplementation on changes in the transcriptome over 6 months. This approach allows us to dissect the changes occurring in the muscle during ageing and in respect to certain medical conditions often present in the elderly.

2 Methodology

2.1 Study design, sample and library preparation and sequencing

In the following section I will briefly describe the background study design, initial sample preparation and sequencing. This part of the work was not directly part of the project but a brief description is essential for interpretation of the results. The workflow is presented schematically in 3.

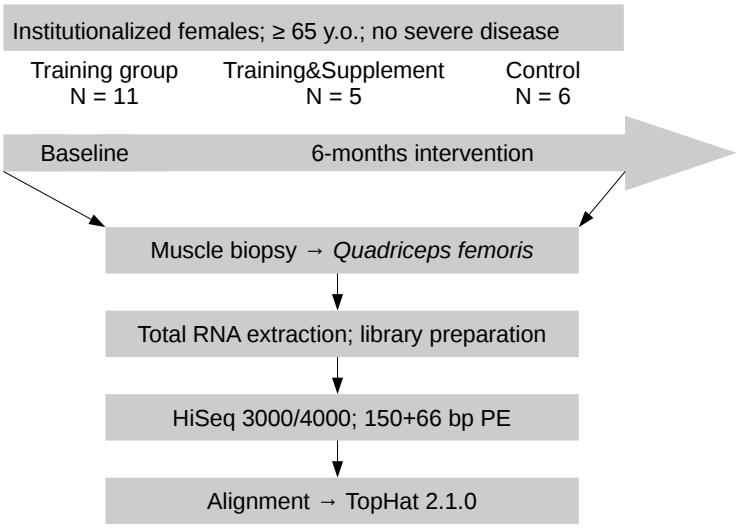


Figure 3: Study design of the Vienna Active Ageing Study. The workflow prior to data analysis is shown. Only those samples on which RNAseq was performed are shown.

2.1.1 Study design

The recruitment characteristics of the participants in the Vienna Active Ageing Study (VAAS) have been published previously [Franzke et al., 2014, Oesen et al., 2015]. Briefly, 14 males and 103 females study participants, residents of 5 different senior residencies in Vienna, gave their written informed consent prior to inclusion in the study. The VAAS was conducted in accordance with the Declaration of Helsinki and approved by the ethical committee of the City of Vienna (EK11-151-0811). The study was designed as a prospective randomised

and controlled intervention study and further details can be found under ClinicalTrials.gov, project number NCT01775111.

Study inclusion criteria:

- Women and Men aged 65 or more
- Adequate mental condition in order to follow the instructions and to perform the resistance exercise independently (Mini-Mental-State higher than 23)
- Ability to walk 10 meters independently (without orthopaedic devices)
- at least 4 points at the Short Physical Performance Battery (SPPB)

Study exclusion criteria:

- Chronic diseases, which contraindicate a medical training therapy
- Serious cardiovascular diseases (congestive chronic heart failure, severe or symptomatic aortic stenosis, unstable angina pectoris, untreated arterial hypertension, cardiac arrhythmias)
- Diabetic retinopathy
- Manifest osteoporosis
- Anticoagulants (example: Marcomar)
- Regular use of cortisone-containing drugs
- Regular strength training (more than once per week) in the last 6 months before inclusion
- Lack of written declaration of consent for testing physical fitness

All inclusion criteria had to be fulfilled in order to take part in the study, whereas only one exclusion criterion was enough for disqualification.

2.1.2 Intervention

The study participants were randomly assigned to three groups: resistance training (RT), resistance training + supplementation (RTS) and control (CT).

- Training (RT): 2x/week progressive resistance training using easy accessible Therabands in order to increase sustainability. One unit consists of 10 minutes warm-up, 35-40min strength training of the major muscle groups and 10min cool down. In the habituation phase (4 weeks, M1) 15 repetitions of the easiest exercise are performed (increase to heavier exercise only when strongly underutilized). From the fifth week intensity and volume increase progressively to two sets of light exercise, if easily done to one set of heavy exercise and one set of easy exercise and further to two sets of heavy exercise. Easily done means that 15 repetitions in the 2nd set is within reach without any problems (2 more repetitions in the 2nd set would be possible) (training program in annex). Within the first 6 months (M1-M6) the guided training was performed 2x/week on non-consecutive days; M7-M12: guided 1x/week + self-organized 1x/week; M13-M24; self-organized 2x/week.

- Training + Supplementation (RTS): 1 daily dose of 1 portion FortiFit (Nutricia) = 150kcal; 20.7g protein (56 `En%`, 19.7g whey protein, 3g leucine, more than 10g essential amino acids); 9.3g carbohydrates (25 `En%`; 0.8 BE); 3.0g fat (18 `En%`); 1.2 g fiber (2 `En%`); 800 IU (20microg) vitamin D; 250mg calcium; vitamins B6 and B12 folate; magnesium. The supplement was consumed every morning with breakfast and additionally after every training session.

- Control (CT): Maintaining the current exercise and eating habits, with additional supervised units for training of cognition, fine motor skills and relaxation exercises twice a week.

For the current analysis, resting skeletal muscle biopsies at baseline (before the beginning of the intervention) and six months post intervention have been obtained from a subgroup of the study population. After quality assessment a subgroup of samples from 22 women was used for RNA extraction and downstream analyses.

2.1.3 Skeletal muscle biopsies

Biopsies were obtained from the middle portion of the muscle *vastus lateralis*, the largest and most powerful part of the *quadriceps femoris*. The procedure was performed using a percutaneous needle biopsy technique according to Bergstrm. After dissecting the muscle samples from blood, adipose and connective tissues, the samples were flash-frozen in liquid nitrogen and stored at `-80C` until further analyses. The weight of the obtained muscle tissue was at least 7 mg up to a maximum of 120 mg (44 +/- 24 mg).

2.1.4 Isolation of mRNA

Total mRNA was extracted from the muscle biopsy samples using Qiazol Lysis Reagent (QIAGEN) according to the protocol given by the manufacturer. The sample quality was assessed using the Agilent 2100 Bioanalyzer, and were rendered suitable for downstream application.

2.1.5 Library preparation

The sequencing library was prepared using a commercial NEBNext Ultra Directional RNA Library Prep Kit (New England Biolabs) following the protocol for polyA enrichment for preparation of the mRNA libraries. The library fragment size ranged between 200 and 1000 bp, with the highest abundance at around 350 bp. All samples produced libraries of acceptable quality.

2.1.6 Sequencing

The libraries were run in four lanes (batches) on an Illumina HiSeq 3000/4000. The output generated 150+66 bp paired end (PE) reads with 30x nominal coverage.

2.2 RNA-seq data analyses

The analyses of RNAseq data starts with collecting the reads and demultiplexing them. As mentioned previously, muscle biopsies from the active ageing study group have been collected from 22 individuals at two time-points: pre and post intervention, yielding a total number of 44 data files. The initial extension of the files is FASTA, and after quality check and trimming the reads are stored into

a FASTQ file, which contains the quality scores in addition to the nucleotide sequence. The initial step in the analysis is mapping fragments to a reference genome, counting the number of reads aligning to a feature, sample comparison and quality check and differential expression.

2.2.1 Read alignment with TopHat 2.1.0

TopHat is built on the ultrafast short read mapping program Bowtie2 and is specifically designed to align RNAseq reads while taking in account exon-exon junctions. It runs on Linux and OS X. TopHat can use both FASTA and FASTQ formats as input.

Typical usage of TopHat:

```
tophat [options] <genome_index_base> PE_reads_1.fq.gz PE_reads_2.fq.gz
```

Here, `<genome_index_base>` corresponds to the basename of the genome index to be searched; `PE_reads_1 and _2` are separate comma-delimited lists corresponding to the paired end reads, and the files in the list must have the same file order in both `*_1 and *_2`.

TopHat reports splice junctions on the basis of RNA-Seq read alignments in UCSC BED track format. Each junction consists of two connected BED blocks, where each block is as long as the maximal overhang of any read spanning the junction. The score is the number of alignments spanning the junction. UCSC BED tracks of insertions and deletions are also reported by TopHat. Before comparing samples and calling differential expression the individual-sample specific GTF files are merged. The merged file defines the comparison run-specific XLOC (gene) and TCOMNS (transcript) identifiers that are used throughout the differential expression tables. The reference transcriptome GTF file of Ensembl gene, transcript and exon annotation from release e75 on GRCh37. Depending on the actual read data some transcript clusters (i.e. genes) may be merged, for example, because the data suggests a read-through event.

The main output file from read alignment with TopHat is a SAM file, or its compressed counterpart, a BAM file. A BAM file contains the same information

as a SAM file, in a binary form. A useful package for working with SAM/BAM files is samtools [2]. Although some minor features of samtools have been used for this work, I will not describe the package here. The file required for the next step is an aligned BAM file in combination with a corresponding index file with the extension BAI.

A newer version of TopHat, 2.1.1, has been released since the initial analyses, however, at this point TopHat has entered a low maintenance and low support stage, due to its replacement with the more efficient HISAT2. Another alternative to TopHat is the Spliced Transcripts Alignment to a Reference (STAR) software, described as an ultrafast universal sequence aligner.

2.2.2 Counting reads with HTSeq count 0.6.1p2

In the next step HTSeq version 0.6.1p2 was used to estimate the read counts from HTSeq is a Python-based framework that allows for efficient analyses of high-throuput sequencing data [Anders et al., 2014]. The command "count" is used to determine the number of reads from an aligned genome map to a certain feature. Depending on the question and method for generating reads, the feature can be a gene, a transcript (exon) or a region. In the case of RNA seq the features are normally genes; exons can be used as features in the case when alternative splicing is under investigation.

The function can be called using the following code:

```
python -m HTSeq.scripts.count [options] <alignment_file> <gff_file>
```

Here, according to the options described in the HTSeq documentation [3] The options used are:

"`-f <format>, --format=<format>`

Format of the input data. Possible values are sam (for text SAM files) and bam (for binary BAM files). Default is sam.

[2] `http://samtools.github.io/hts-specs/SAMv1.pdf`
[3] `http://www-huber.embl.de/users/anders/HTSeq/doc/count.html`

`-r <order>, --order=<order>`

For paired-end data, the alignment have to be sorted either by read name or by alignment position. If your data is not sorted, use the samtools sort function of samtools to sort it. Use this option, with name or pos for order to indicate how the input data has been sorted. The default is name.

If name is indicated, htseq-count expects all the alignments for the reads of a given read pair to appear in adjacent records in the input data. For pos, this is not expected; rather, read alignments whose mate alignment have not yet been seen are kept in a buffer in memory until the mate is found. While, strictly speaking, the latter will also work with unsorted data, sorting ensures that most alignment mates appear close to each other in the data and hence the buffer is much less likely to overflow.

`-s <yes/no/reverse>, --stranded=<yes/no/reverse>`

whether the data is from a strand-specific assay (default: yes)

For stranded=no, a read is considered overlapping with a feature regardless of whether it is mapped to the same or the opposite strand as the feature. For stranded=yes and single-end reads, the read has to be mapped to the same strand as the feature. For paired-end reads, the first read has to be on the same strand and the second read on the opposite strand. For stranded=reverse, these rules are reversed.

`-a <minaqual>, --a=<minaqual>`

skip all reads with alignment quality lower than the given minimum value (default: 10 Note: the default used to be 0 until version 0.5.4.)

`-t <feature type>, --type=<feature type>`

feature type (3rd column in GFF file) to be used, all features of other type are ignored (default, suitable for RNA-Seq analysis using an Ensembl GTF file: exon)

`-i <id attribute>, --idattr=<id attribute>`

GFF attribute to be used as feature ID. Several GFF lines with the same feature ID will be considered as parts of the same feature. The feature ID is used to identity the counts in the output table. The default, suitable for RNA-Seq analysis using an Ensembl GTF file, is gene-id

`-m <mode>, --mode=<mode>`

Mode to handle reads overlapping more than one feature. Possible values for mode are union, intersection-strict and intersection-nonempty (default: union)

`-o <samout>, --samout=<samout>`

write out all SAM alignment records into an output SAM file called samout, annotating each line with its assignment to a feature or a special counter (as an optional field with tag XF)"

2.2.3 Data normalization and differential expression analyses with DESeq2

The count output files from HTSeq can be used further directly as input files for DESeq. A designated function, *DESeqDataSetFromHTSeqCount*, is designed in order to create a matrix directly from HTSeq count files. The default DESeq input is a count matrix that contains features such as genes as rows and samples as columns. The count matrix consists of integers X_ij where i corresponds to the gene and j to the sample. These integers represent non-normalized counts of sequencing reads in the case of a single-end RNAseq experiment or fragments in paired-end experiment. In order to avoid over-normalization, normalized counts should not be used as the DESeq2 model corrects for library size.

According to the extensive and easy to follow DESeq2 manual [4] the design formula expresses the variables which will be used in modeling. The formula is written with a tilde (`~`) followed by the variables or factors of interest connected by a plus sign. An intercept is included, representing the base mean of the counts. In order to benefit from the default settings of the package, you should put the variable of interest at the end of the formula and make sure the control level is the first level.

To test for differentially expressed genes throughout condition and treatment, we have used two approaches. One was looking at both time and condition as factors in the same formula and looked at the differential gene expression post intervention, taking in account the group differences that might have been present at baseline. The second approach takes each group separately and compares the post-intervention condition to the pre-intervention condition.

[4] `http://www.bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf`

The following code has been used for the current analysis, with some modifications regarding the desired output:

```
# set a working directory where the HTSeq ouput files are located

dir<- "/source_directory/Data"

# load the DESeq2 package

library(DESeq2)

# first, from the working directory select all HTSeq output files using
    the command "grep" and a common name excerpt

sampleFiles <- grep("*samout.txt",list.files(dir),value=TRUE)

# next, set the conditions in a data.frame corresponding to each sample;
    at this point the sample names are listed alphabetically
pd <- data.frame(group=rep(c('Cntrl',
    'Cntrl','Cntrl','Cntrl','Cntrl','Cntrl',
      'Cntrl','Cntrl','Cntrl','Cntrl', 'Training', 'Training',
               'Training','Training','Training','Training','Training',
               'Training','Training','Training','Training','Training',
               'Training','Training','Training','Training','Training',
               'Training','Training','Training','Training','Training',
               'TrSuplement', 'TrSuplement','TrSuplement','TrSuplement',
               'TrSuplement','TrSuplement','TrSuplement','TrSuplement',
               'TrSuplement','TrSuplement')),
               time=rep(c( 'post','pre'), 21))

# alternatively prepare a csv file containing the condition variables
    (it can contain other variables such as age, clinical parameters
    etc.

pd <- read.csv("/source_directory/Data/PD_file.csv")

# generate the sample table

sampleTable <- data.frame(sampleName = sampleFiles, fileName =
    sampleFiles, condition = pd)


# generate the DESeq data set; the design formula can contain more than
    one factor, however, DESeq2 calculates differential expression for
    the last listed factor and the rest are used as covariates
```

```
ddsHTSeq <- DESeqDataSetFromHTSeqCount(sampleTable = sampleTable,
    directory = dir, design= ~ condition.group + condition.time)

# remove all features that have less than 1 read per sample (42 samples)

dds <- ddsHTSeq[ rowSums(counts(ddsHTSeq)) > 42, ]

# estimate size factors; this feature is useful in order to normalize
    the data

geoMeans <- exp(rowMeans(log(counts(dds))))
dds <- estimateSizeFactors(dds,geoMeans=geoMeans)
sizeFactors(dds)

# export the normalized counts as a matrix-like table in csv or tsv
    format; this table can be used then to estimate outliers by PCA or
    SaVant
norm<- counts(dds, normalized=TRUE)
write.table(norm, file="/source_directory/Data/DESeq_matrix_norm.tsv",
    quote=FALSE, sep='\t', col.names = NA)

# set the reference condition, in this case "pre"

dds$condition <- relevel(dds$time, ref="pre")

# run DESeq and collect results in a tabelar form

dds <- DESeq(dds)
res <- results(dds)
res
write.table(res, file="/source_directory/Data/DEG_all.tsv", quote=FALSE,
    sep='\t', col.names = NA)

# examine the results

sum(res$padj < 0.05, na.rm=TRUE)
sum(res$padj < 0.1, na.rm=TRUE)
sum(res$pvalue < 0.05, na.rm=TRUE)

# plot the results

svg("/source_directory/Data/Plot_DEG.svg")
plotMA(res, main="DESeq2", ylim=c(-2,2))
dev.off()

# to view results related to a different condition

results(dds, contrast=c("condition","Cntrl","Training"))
```

```
# to look at the cumulative effect of two conditions and compare
    subgroups (eg. Cntrl post vs. Cntrl pre) independently of the other
    groups

dds$cumulative <- factor(paste0(dds$group, dds$time))
design(dds) <- ~ cumulative
dds <- DESeq(dds)
resultsNames(dds)
results(dds, contrast=c("cumulative", "Cntrlpost", "Cntrlpre"))
```

2.2.4 Signature Visualization with SaVant

In order to estimate the predominant tissue and cell type through all samples we used SavAnt (Lopez et. al, unpublished, [5]), a web service aimed at estimating specific tissue and cell signatures from RNAseq data. [6]
Using SaVant is relatively simple, it requires a table containing Gene symbols in the first column and samples in the other columns; the rows represent normalized read counts of a certain feature. At The user can choose a signature category (at the moment between human and mouse). Additional features are described in [7]. The output is a heat map matrix showing the abundance of signatures per tissue or cell type for the given sample.

2.2.5 PCA analysis and clustering with COVAIN

Principal Component analysis and bi-clustering were utilized in order to identify outliers among the tested samples. For this purpose, a Matlab toolbox, COVAIN, was used [Sun and Weckwerth, 2012]. This toolbox allows for input of up to 500 features across all samples. The input matrix uses normalized counts from DESeq2 followed by log transformation. The 500 genes with the highest coefficient of variation across all samples were used for this purpose.

2.2.6 Pathway analyses with ConsensusPathDB

In order to identify pathways that are up- or downregulated upon intervention in the groups, after running differential expression analyses, a pathway analyses can be employed. There are several tools that can be utilized for this purpose,

[5] http://pellegrini.mcdb.ucla.edu/Lab/pellegrinilabscps/SCP_SaVant.pdf
[6] http://pathways.mcdb.ucla.edu/savant/
[7] http://pellegrini.mcdb.ucla.edu/Lab/pellegrinilabscps/SCP_SaVant.pdf

including Ingenuity Pathway Analysis (IPA) or Gene Set Enrichment Analysis (GSEA). A commonly used database is The Database for Annotation, Visualization and Integrated Discovery (DAVID), that provides gene-annotation enrichment analysis and functional annotation clustering. However, this database has not been regularly maintained, and the current version v6.7 will be replaced in October 2016 with the updated version v6.8. Hence, ConsensusPathDB is used here as an alternative with very similar interphase and output. ConsensusPathDB is a database integrating interactions between genes or transcripts on a functional level.Cur The information on the interactions comes from 30 public resources and is integrated into a common network. Gene regulatory interactions are mapped and grouped together according to similarity. This database content is updated every three months, to make sure it contains the latest information.

The statistical approach used to analyze user-specified lists of genes by ConsensusPathDB is over-representation analysis, where predefined lists of functionally associated genes (pathways, Gene Ontology (GO) categories and neighborhood-based entity sets) are tested for over-representation in the user-specified list based on the hypergeometric test. The over-representation functionality takes as input a relatively short, non-weighted list of differentially expressed genes. The input identifiers are then mapped to physical entities and over-represented sets are searched among three categories of predefined gene sets: network neighborhood-based sets, pathway-based sets and Gene Ontology-based sets. The user can select which functionality is most informative for the specific experiment [8]. As an input we used differentially expressed genes identified by DESeq2, with adjusted p values lower than 0.1. As the list of such genes was very short, we have additionally used a list of genes where the non-adjusted p-value was lower than 0.05. This set was further divided in two subsets: genes that are upregulated after intervention and genes that are downregulated.

[8] `http://cpdb.molgen.mpg.de/`

3 Results and Conclusions

3.1 Count features

The output from HTSeq count is a table containing raw reads per feature in a given sample. HTSeq count outputs the data as a list of features (genes or transcripts) with their Ensembl identifiers and raw count of reads per feature. In addition, as an output HTSeq count gives a statistical evaluation on the number of reads that correspond to different features.

eg. `MG_24_pre`

- `__no_feature` 1228833
- `__ambiguous` 472619
- `__too_low_aQual` 0
- `__not_aligned` 0
- `__alignment_not_unique` 916119

These features were similar in all samples rendering around `6%` of the reads not aligning to a feature, `2%` being ambiguous and approximately `5%` of all reads aligning to more than one feature. A large number of features were not covered by reads, yielding a count result of zero. Such features could be problematic for downstream analyses and can be filtered out at this point or in the next step.

3.2 Sample signature visualtisation with SaVant, PCA analyses and clustering

Using SaVant for signature visualisation on log-transformed, normalized counts and using the Human Body atlas signature database. We could confirm that the predominant cell type in all samples was skeletal muscle (figure 4). However, one of the samples showed a very distinct signature, indicative of potential infiltration of immune cells (figure 4).

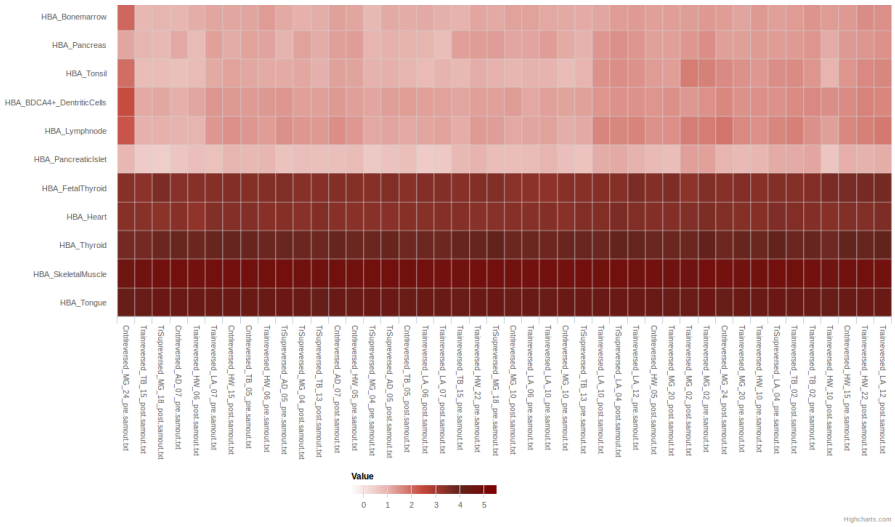


Figure 4: Signature Visualisation with Savant revealed a consistent and strong skeletal muscle signature in all samples. One of the samples contains signals corresponding to immune cells. Only the most relevant tissues from the human body atlas are shown.

Validation using PCA on the 500 most variable genes across all samples showed that this sample does not cluster closely with all other samples (figure 5). This lead to exclusion of the sample from all further analyses, in order to minimize potential discrepancies in the results due to inadequate sample quality.

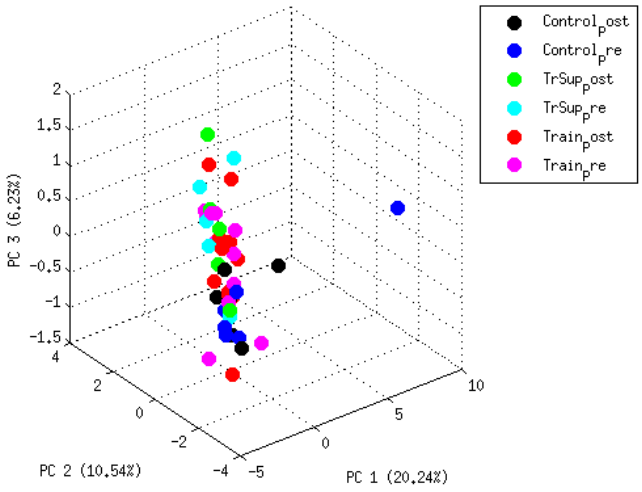


Figure 5: Principal component analysis reveals how samples cluster together. One sample deviates significantly from the rest.

3.3 Differential gene expression with DESeq2

Using the `condition.group + condition.time` formula in the DESeq design we could identify genes that were differentially expressed post intervention compared to pre intervention, taking in account the intervention group of the study participant. The output data frame has 6 columns: baseMean, log2FoldChange, lfcSE, stat, pvalue and padj. The log2FoldChange shows the size of the effect, and whether it is positive or negative. The padj corresponds to an adjusted p-value that takes in account multiple testing artefacts present in big data analyses. The MA plot in figure 6 shows the number of genes with padj value lower than 0.1, marked with red. As observed, only a small number of genes were differentially expressed, mostly downregulated, after intervention, and the overall effect was not very large.

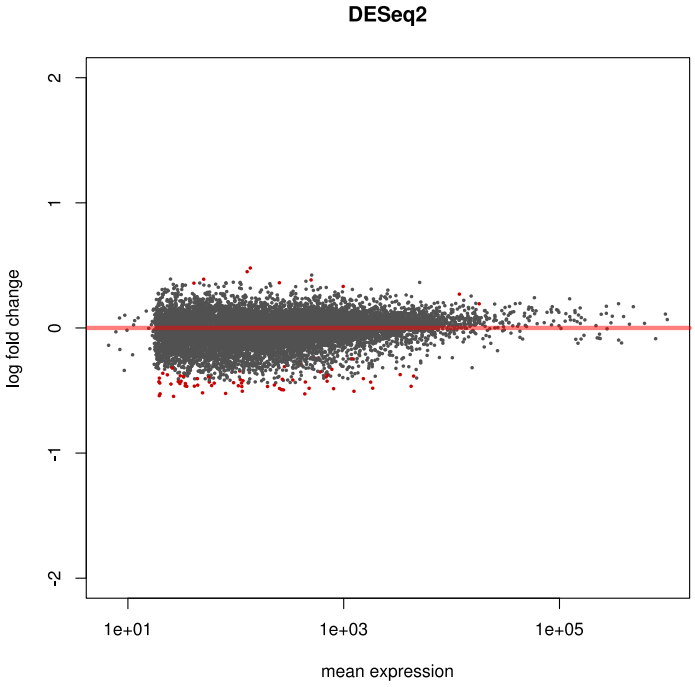
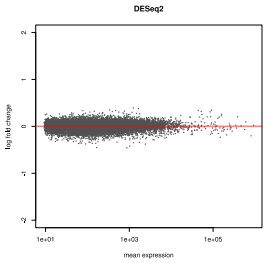
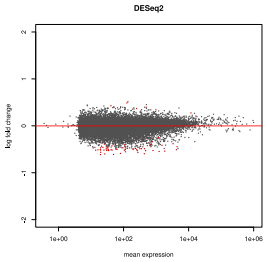


Figure 6: MA-plot of log2fold values between post and pre-treatment, taking in account the intervention group. Red dots represent differentially expressed genes, p(adjusted) ≤ 0.1.

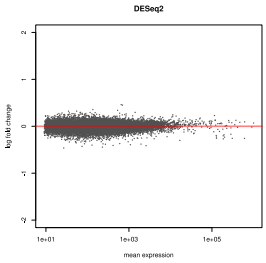
We next compared each of the three intervention groups separately, using the interaction feature from DESeq and using different group contrasts. Figure 7 shows the MA plots of differentially expressed genes after intervention in all three groups. As evident from the plots, there were no differentially expressed genes in neither the control group nor the traaining and supplementation group at a padj level below 0.1. In contrast, the training group showed a modest level of differentially expressed genes. This might indicate that protein supplementation diminishes the changes occuring in skeletal muscle during exercise training.



(a) MA-plot of log2fold values between post and pre-treatment in the control group



(b) MA-plot of log2fold values between post and pre-treatment in the training group.



(c) MA-plot of log2fold values between post and pre-treatment in the training and supplementation group

Figure 7: Red dots represent differentially expressed genes, p(adjusted) ≤ 0.1.

We further explored the number of genes that showed a modest change with an unadjusted p-value below 0.05 in the three intervention groups. We were interested in the overlap of genes that have a potential to be affected in all three groups. Figure 8 shows the number of features with a p-value below 0.05 in all groups separately, as well as the genes commonly changed among two or more groups. There was only a very small overlap between each two groups and almost no common features changed in all three groups at the same time.

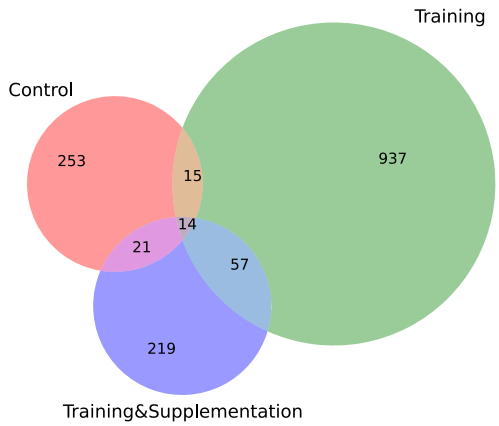


Figure 8: Venn diagram representing differentially expressed genes at p-value ≤ 0.05 in all three groups separately, and overlapping differentially expressed genes between groups.

3.4 Pathway enrichment analyses

In the previous section we could establish a modest effect of exercise training alone on skeletal muscle, due to the number of differentially expressed genes between pre and post training intervention. In order to identify which cell function might be altered and to reduce the number of candidate genes for quantitative PCR verification, we run a pathway analysis based on gene enrichment sets using the ConsensusPathDB. We used the pathway analysis option and the significantly enriched pathways are shown in table 1. The gene IDs are given as ENSAMBLE gene IDs. These can be easily translated into different identifiers using conversion tools such as biodbnet [9].

[9] `https://biodbnet-abcc.ncifcrf.gov/db/db2db.php`

p-value	q-value	pathway	source	members_input_overlap
0.000	0.000	Vascular smooth muscle contraction - Homo sapiens (human)	KEGG	ENSG00000101335; ENSG00000065534; ENSG00000145936; ENSG00000173175; ENSG00000107796; ENSG00000164116; ENSG00000122786; ENSG00000163017; ENSG00000133392; ENSG00000151067; ENSG00000072952; ENSG00000067900; ENSG00000151617; ENSG00000125503; ENSG00000167641
0.000	0.002	ECM proteoglycans	Reactome	ENSG00000105664; ENSG00000130702; ENSG00000157766; ENSG00000259207; ENSG00000122176; ENSG00000041982; ENSG00000182492
0.000	0.002	Smooth Muscle Contraction	Reactome	ENSG00000101335; ENSG00000065534; ENSG00000107796; ENSG00000122786; ENSG00000163017; ENSG00000133392
0.000	0.006	Platelet degranulation	Reactome	ENSG00000196924; ENSG00000137801; ENSG00000259207; ENSG00000130402; ENSG00000120885; ENSG00000072110; ENSG00000174175; ENSG00000131236
0.000	0.006	Focal adhesion - Homo sapiens (human)	KEGG	ENSG00000196924; ENSG00000105664; ENSG00000065534; ENSG00000101335; ENSG00000169398; ENSG00000130702; ENSG00000125503; ENSG00000259207; ENSG00000130402; ENSG00000137801; ENSG00000072110; ENSG00000041982; ENSG00000067900
0.000	0.006	Response to elevated platelet cytosolic Ca2+	Reactome	ENSG00000196924; ENSG00000120885; ENSG00000259207; ENSG00000130402; ENSG00000137801; ENSG00000072110; ENSG00000174175; ENSG00000131236

0.000	0.007	Endothelin Pathways	Wikipathways	ENSG00000078401; ENSG00000164128; ENSG00000065534; ENSG00000130176; ENSG00000151617
0.000	0.017	erk and pi-3 kinase are necessary for collagen binding in corneal epithelia	BioCarta	ENSG00000169398; ENSG00000130402; ENSG00000065534; ENSG00000067900; ENSG00000072110
0.000	0.017	Syndecan interactions	Reactome	ENSG00000072110; ENSG00000137801; ENSG00000041982; ENSG00000259207
0.000	0.019	pkc-catalyzed phosphorylation of inhibitory phosphoprotein of myosin phosphatase	BioCarta	ENSG00000141837; ENSG00000065534; ENSG00000067900; ENSG00000167641
0.001	0.025	Non-integrin membrane-ECM interactions	Reactome	ENSG00000041982; ENSG00000137801; ENSG00000130702; ENSG00000072110; ENSG00000259207
0.001	0.025	Sympathetic Nerve Pathway (Neuroeffector Junction)	PharmGKB	ENSG00000141837; ENSG00000164128; ENSG00000151067
0.001	0.026	TGF beta Signaling Pathway	Wikipathways	ENSG00000162772; ENSG00000169398; ENSG00000259207; ENSG00000137801; ENSG00000125740; ENSG00000185359; ENSG00000170345; ENSG00000041982; ENSG00000067900
0.001	0.028	RHO GTPases activate PAKs	Reactome	ENSG00000196924; ENSG00000065534; ENSG00000133392; ENSG00000101335
0.001	0.028	Sema4D induced cell migration and growth-cone collapse	Reactome	ENSG00000164050; ENSG00000101335; ENSG00000067900; ENSG00000133392

0.001	0.039	cGMP-PKG signaling pathway - Homo sapiens (human)	KEGG	ENSG00000101335; ENSG00000065534; ENSG00000145936; ENSG00000173175; ENSG00000164116; ENSG00000151067; ENSG00000072952; ENSG00000067900; ENSG00000151617
0.001	0.039	Syndecan-4-mediated signaling events	PID	ENSG00000169398; ENSG00000041982; ENSG00000137801; ENSG00000072110
0.001	0.039	Sema4D in semaphorin signaling	Reactome	ENSG00000164050; ENSG00000101335; ENSG00000067900; ENSG00000133392
0.002	0.040	Muscle contraction	Reactome	ENSG00000101335; ENSG00000065534; ENSG00000107796; ENSG00000122786; ENSG00000163017; ENSG00000133392
0.002	0.042	Focal Adhesion	Wikipathways	ENSG00000196924; ENSG00000105664; ENSG00000065534; ENSG00000169398; ENSG00000130702; ENSG00000259207; ENSG00000137801; ENSG00000072110; ENSG00000041982; ENSG00000067900
0.002	0.042	Extracellular matrix organization	Reactome	ENSG00000105664; ENSG00000130702; ENSG00000157766; ENSG00000259207; ENSG00000137801; ENSG00000072110; ENSG00000169436; ENSG00000122176; ENSG00000041982; ENSG00000182871; ENSG00000182492
0.002	0.043	EPHA-mediated growth cone collapse	Reactome	ENSG00000169398; ENSG00000101335; ENSG00000067900; ENSG00000133392

Table 1: ConsensusPathDB output of significantly enriched pathways affected by intervention training in the elderly. Input genes that yielded p-value lower than 0.05 with DESeq analyses were used for this analysis. Only pathways with q-value lower than 0.05 are shown.

As observed from table 1, there are numerous pathways that might be of interest for functional and mechanistic evaluation One of them, the TGF beta signalling pathway has been evaluated in blood and blood cells in this study population and hasbeen published previously [Halper et al., 2015, Hofmann et al., 2015]. However, other interesting candidate pathways are yet to be evaluated either in the context of the present study or using a new in vivo or in vitro model.

4 Discussion and Future prospects

The analysis of RNA sequencing data is relatively complex, and there are many different packages that can specifically fulfill a certain task or requirement. Conesa et al. [2016] have recently summarized the best practices for RNAseq analyses. The choice of pipeline for RNAseq analysis will greatly influence the final results as many of the different packages use different statistical methods or normalization procedures. However, taking in account the current study design, DESeq2 was the method of choice as it allows for a model that includes several factors at the same time and in addition it controls for confounders such as batch differences.

The next step in the analysis and at the same time an important goal of the Active Ageing Study would be to investigate effects of blood-based and functional biomarkers, as well as medication or underlying conditions on the change in transcription in this study group. DESeq2 can be used with continuous variables as factors in addition to categorical. This approach will be useful for building a multivariable model for prediction of transriptional changes. However, a more comprehensive view of the complex interactions between gene expression and clinical or functional parameters can be obtained by network analysis. To that end, future plans include weighted correlation network analysis (WGCNA) for finding clusters of highly correlated genes, and relating these to one another and to external sample traits [Langfelder and Horvath, 2016]. A similar approach has been utilized recently [Ponsuksili et al., 2015], however, the Vienna Active Ageing Study provides a unique and interesting study group of elderly institutionalized individuals, where the effects of training on the skeletal muscle transcriptome has not as yet been evaluated.

In order to get more meaningful interpretation of the data obtained, apart from more extensive bioinformatics analysis, it is of great importance to employ a highly specialized interpretation of the output. Especially in the case of large data, the evidence can be often too overwhelming and misinterpreted. Studies on the interphase of several disciplines such as molecular biology, physiology, bioinformatics and statistics are essential for advancement of health-related biomedical translational research.

Bibliography

Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. Htseq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, page btu638, 2014.

Ivan Borozan, Stuart N Watt, and Vincent Ferretti. Evaluation of alignment algorithms for discovery and identification of pathogens using rna-seq. *PloS one*, 8(10):e76935, 2013.

HPJ Buermans and JT Den Dunnen. Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1842(10):1932–1941, 2014.

Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):1, 2016.

Yongzeng Ding, Li Xu, Borko D Jovanovic, Irene B Helenowski, David L Kelly, William J Catalona, Ximing J Yang, Michael Pins, and Raymond C Bergan. The methodology used to measure differential gene expression affects the outcome. *Journal of Biomolecular Techniques*, 18(5):321, 2007.

Bernhard Franzke, Barbara Halper, Marlene Hofmann, Stefan Oesen, Heidemarie Peherstorfer, Klemens Krejci, Birgit Koller, Karin Geider, Andreas Baierl, Anela Tosevska, et al. The influence of age and aerobic fitness on chromosomal damage in austrian institutionalised elderly. *Mutagenesis*, 29 (6):441–445, 2014.

Malachi Griffith, Jason R Walker, Nicholas C Spies, Benjamin J Ainscough, and Obi L Griffith. Informatics for rna sequencing: a web resource for analysis on the cloud. *PLoS Comput Biol*, 11(8):e1004393, 2015.

Barbara Halper, Marlene Hofmann, Stefan Oesen, Bernhard Franzke, Petra Stuparits, Claudia Vidotto, Harald Tschan, Norbert Bachl, Eva-Maria Strasser, Michael Quittan, et al. Influence of age and physical fitness on mirna-21, tgf-ß and its receptors in leukocytes of healthy women. *Exercise immunology review*, 21, 2015.

Marlene Hofmann, Barbara Halper, Stefan Oesen, Bernhard Franzke, Petra Stuparits, Harald Tschan, Norbert Bachl, Eva-Maria Strasser, Michael Quittan, Martin Ploder, et al. Serum concentrations of insulin-like growth factor-1, members of the tgf-beta superfamily and follistatin do not reflect different stages of dynapenia and sarcopenia in elderly women. *Experimental gerontology*, 64:35–45, 2015.

Clyde A Hutchison. Dna sequencing: bench to bedside and beyond. *Nucleic acids research*, 35(18):6227–6237, 2007.

Peter Langfelder and Steve Horvath. Tutorial for the wgcna package for r ii. consensus network analysis of liver expression data, female and male mice 2. b step-by-step network construction and module detection. 2016.

Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.

Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

Allan J McLean and David G Le Couteur. Aging biology and geriatric clinical pharmacology. *Pharmacological reviews*, 56(2):163–184, 2004.

Stefan Oesen, Barbara Halper, Marlene Hofmann, Waltraud Jandrasits, Bernhard Franzke, Eva-Maria Strasser, Alexandra Graf, Harald Tschan, Norbert Bachl, Michael Quittan, et al. Effects of elastic band resistance training and nutritional supplementation on physical performance of institutionalised elderlya randomized controlled trial. *Experimental gerontology*, 72:99–108, 2015.

Gladys Leopoldine Onambélé-Pearson, Leigh Breen, and Claire E Stewart. Influences of carbohydrate plus amino acid supplementation on differing exercise intensity adaptations in older persons: skeletal muscle and endocrine responses. *Age*, 32(2):125–138, 2010.

Mark D Peterson, Matthew R Rhea, Ananda Sen, and Paul M Gordon. Resistance exercise for muscular strength in older adults: a meta-analysis. *Ageing research reviews*, 9(3):226–237, 2010.

Siriluck Ponsuksili, Puntita Siengdee, Yang Du, Nares Trakooljul, Eduard Murani, Manfred Schwerin, and Klaus Wimmers. Identification of common regulators of genes in co-expression networks affecting muscle and meat properties. *PloS one*, 10(4):e0123678, 2015.

Heloisa T Rabelo, Lídia A Bezerra, Denize F Terra, Ricardo M Lima, Maria AF Silva, Tailce K Leite, and Ricardo J de Oliveira. Effects of 24 weeks of progressive resistance training on knee extensors peak torque and fat-free mass in older women. *The Journal of Strength & Conditioning Research*, 25 (8):2298–2303, 2011.

Xiaoliang Sun and Wolfram Weckwerth. Covain: a toolbox for uni-and multivariate statistics, time-series and correlation network analysis and inverse estimation of the differential jacobian from metabolomics covariance data. *Metabolomics*, 8(1):81–93, 2012.

Cole Trapnell, Lior Pachter, and Steven L Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.

Erwin L van Dijk, Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. Ten years of next-generation sequencing technology. *Trends in genetics*, 30(9): 418–426, 2014.

Lex B Verdijk, Richard AM Jonkers, Benjamin G Gleeson, Milou Beelen, Kenneth Meijer, Hans HCM Savelberg, Will KWH Wodzig, Paul Dendale, and Luc JC van Loon. Protein supplementation before and after exercise does not further augment skeletal muscle hypertrophy after resistance training in elderly men. *The American journal of clinical nutrition*, 89(2):608–616, 2009.

Elizabeth Weening-Dijksterhuis, Mathieu HG de Greef, Erik JA Scherder, Joris PJ Slaets, and Cees P van der Schans. Frail institutionalized older persons: A comprehensive review on physical exercise, physical fitness, activities of daily living, and quality-of-life. *American Journal of Physical Medicine & Rehabilitation*, 90(2):156–168, 2011.