Institute for Biostatistics and Informatics in Medicine and Ageing Research (IBIMA),
Rostock University Medical Center
Director: Prof. Dr. Georg Fuellen

Microarray transcriptomics analysis of Hutchinson-Gilford progeria genetic disease and the exploitation of dual RNA-seq technology to unveil host-pathogen interaction

Inaugural Dissertation

to obtain the academic degree
Doctor rerum humanarum (Dr. rer. hum.)
of Rostock University Medical Center

Submitted by
Bioinformatician, Salem Oduro Beffi Sueto
Born November 2nd 1988,
in Accra (Ghana)

Rostock, 24.04.2023

https://doi.org/10.18453/rosdok_id00004514

DOKTORARBEIT
UNIVERSITÄTSMEDIZIN ROSTOCK

Reviewers:

Prof. Dr. Georg Fuellen, Universitätsmedizin Rostock (Rostock, Germany), Institut für Biostatistik und Informatik in Medizin und Alternsforschung

Prof. Dr. Micheal Walter, Universitätsmedizin Rostock (Rostock, Germany), Institut für Klinische Chemie und Laboratoriumsmedizin

Prof. Dr. Dirk Repsilber, Örebro University (Örebro, Sweden), Functional Bioinformatics (Department of Medical Science)

Date of submission: 24.04.2023

Date of oral defense: 07.11.2023

Personal declaration

I hereby officially declare that I have written this dissertation independently. Any help and assistance in creating this work are clearly indicated in the acknowledgments. In addition, I affirm that I have cited all publications and other sources used in the preparation of this academic work in the appropriate place. I further confirm that my work has been accomplished in accordance with the "Rules to ensure good scientific practice and to avoid scientific misconduct" of Rostock University Medical Center.

Ich versichere eidesstattlich durch eigenhändige Unterschrift, dass ich die Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen sind, habe ich als solche kenntlich gemacht.

Die Arbeit ist noch nicht veröffentlicht und ist in gleicher oder ähnlicher Weise noch nicht als Studienleistung zur Anerkennung oder Bewertung vorgelegt worden. Ich weiß, dass bei Abgabe einer falschen Versicherung die Prüfung als nicht bestanden zu gelten hat.

Salem Oduro Beffi Sueto

“Do not quench the Spirit. Do not
despise prophecies, but test everything; hold fast
what is good.”

1 Thessalonians 5:19-21
Holy Bible (New International Version)

Acknowledgements

The present dissertation is the physical manifestation of the sacrifice of multiple people throughout my life. I would like to show my gratitude to each one of them for their support and encouragement.

First of all, I sincerely want to thank Dr. Prof. Georg Fuellen for giving me the opportunity to be part of the IBIMA group and for the constant support throughout the years. I would like to thank my direct supervisor Israel Barrantes for the constant guidance, vision, and constructive criticisms he provided me during my IBIMA experience.

I would like to express my appreciation to our collaborative partners: Dr. Prof. Bernd Kreikemeyer, Nadja Patenge, and Kevin Strey from the Institute for Medical Microbiology, Virology, and Hygiene (University of Rostock, Rostock, Germany) for their help and support for the human infection study.

Moreover, I would like to thank our second collaborative partners Prof. Dr. Michael Walter and Kathrin Jäger from the Institute of Clinical Chemistry and Laboratory Medicine (University of Rostock, Rostock, Germany) for their help in the Progeria study.

Furthermore, I would like to thank all former and current members of the IBIMA research group: Almut Brauer, Yvonne Gladbach, Sarah Fischer, Mohammed Fahmy, Steffen Möller, Axel Kowald, Daniel Palmer, Franziska Meiners, and Riccardo Secci. Thank you for creating a supportive, collaborative, and constructive work environment. It was my honor to be part of the group.

Special appreciation to all my friends that I made in Rostock for their gentle and constructive help throughout my residency in Rostock, Germany.

Finally, a special thanks to my family and God for giving me this opportunity to aspire toward my dreams.

Table of Contents

Abbreviations

CAGE	Cap Analysis of Gene Expression
CAM	Cell Adhesion Molecule
CD	Characteristic Direction
cDNA	complementary DNA
ChIP-seq	Chromatin Immune Precipitation Sequencing
CMAP	Connectivity Map
DE	Differentially Expressed
DEG	Differentially Expressed Gene
DR	Drug repurposing
ECM	Extracellular Matrix
ER	Endoplasmic reticulum
ESTs	Expressed Sequence Tags
FDR	False Discovery Rate
FPKM	Fragment Per Kilobase per Million mapped reads
GAS	Group A Streptococci (Streptococcus pyogenes)
GAS M1-AP1	S. pyogenes serotype M1 strain AP1
GAS M49-591	S. pyogenes serotype M49 strain 591
GEO	Gene Expression Omnibus
GO	Gene Ontology
GOA	Gene Ontology Annotation
GO:BP	Gene Ontology Biological Process
GO:CC	Gene Ontology Cellular Component
GO:MF	Gene Ontology Molecular Function
HGPS	Hutchinson-Gilford Progeria syndrome
IAV	Influenza A virus
iNOS	inducible NO Synthase
LFC	Log_2 fold change
LINCS	Library of Integrated Network-based Cellular Signatures
LMNA	Lamin A
MAMs	Membrane-associated mucins
MDS	Multidimensional Scaling
MOA	Mechanism Of Action
PCA	Principal Component Analysis
RMA	Robust Multi-array Average
RNA-Seq	RNA sequencing
RPKM	Reads Per Kilobase per Million mapped reads
SAGE	Serial Analysis of Gene Expression
SNP	Single Nucleotide Polymorphism
TPM	Transcripts per Kilobase Million
WIPO	World Intellectual Property Organization

Summary

The transcriptomic analysis provides a snapshot of the entire repertoire of RNA molecules (transcripts) exhibited by the cells in a specific biological moment. In the last decade, RNA-seq and microarray platforms have become the standard platforms due to their ability to capture thousands of different transcripts in a single experiment. The present work presents the versatility of the two transcriptomic platforms for human healthcare studies.
The microarray platform was used to analyse the gene expression of fibroblast human cells from patients affected by the Hutchinson-Gilford progeria syndrome (HGPS). The disease is caused by a single genetic mutation in the LMNA gene that causes several premature ageing symptoms in children. The analysis uses fibroblast cells of different sources: fibroblast from HGPS patients, fibroblast from healthy people treated with UV-B light, fibroblast from HGPS patients treated for telomere elongation, and fibroblast from healthy people as the control group. The aim of the study is the transcriptomics profiling of the HGPS, i.e. the identification of the significantly expressed genes and their biological regulation; and, its comparison to other cell senescence processes such as the UV-B treated cells and the telomere elongated treated cells. The results show HGPS affects several genes related to the cell cycle checkpoints, histone modification, and telomerase maintenance. The healthy UV-treated cells exhibit genes related to the cell cycle regulation during G1/S and G2/S phase transition, and cellular response to DNA damage. The effect of telomere elongation on HGPS fibroblast cells affects different stages of the transcription-translation process such as the RNA polymerase I, RNA polyadenylation, and mRNA 3'-end processing. Furthermore, the analysis shows that HGPS and the UV-B treated cells have different senescence profiles. The telomere elongation in HGPS does not provide a decisive conclusion on its ability to solve HGPS deficiencies.

An extensive literature review was conducted to identify drug repurposing tools that use gene expression data to identify potential drug treatments. The results of the review was then applied in the second study of the dissertation to identify potential drugs against bacterial and viral infection in human.

The RNA-seq platform was used to unveil the host-pathogen interaction through the use of the dual RNA-seq approach. This method permits the simultaneous examination of gene expression in two interacting species without a physical separation step during the library preparation but a subsequent separation of the reads from each specie in silico with bioinformatics tools. In the present work, dual RNA-seq methodology was applied to study the interaction between human cells and two S. pyogenes (GAS) bacterial strains (AP1 and NZ131) and the Influenza A virus (IAV). The goal of the study is to unveil the mechanisms which render IAV-infected patients more susceptible to S. pyogenes, the biological reasons why IAV-GAS dual-infected patients show severe symptoms when compared to single-infected ones, and the identification of potential drug treatment against the two pathogens. The analysis show all three pathogens exhibit an up-regulation of ATP production and genes related to mitochondria complexes. The S. pyogenes NZ131 infected cells show a higher infection rate compared to S. pyogenes AP1. The results also show that IAV infection impacts secondary S. pyogenes infection by affecting the stability of the extracellular matrix through cell-cell adhesion and down-regulation of several histone genes, known for their extracellular antimicrobial properties. At last, drug repurposing analysis identified 40 potential drugs for the treatments of an IAV-GAS co-infection in human cells.

Zusammenfassung

Die transkriptomische Analyse liefert eine Momentaufnahme des gesamten Repertoires an RNA-Molekülen (Transkripten), die von den Zellen in einem bestimmten biologischen Moment gezeigt werden. In den letzten zehn Jahren haben sich RNA-seq- und Microarray-Plattformen aufgrund ihrer Fähigkeit, Tausende verschiedener Transkripte in einem einzigen Experiment zu erfassen, zu Standardplattformen entwickelt. In der vorliegenden Arbeit wird die Vielseitigkeit der beiden transkriptomischen Plattformen für Studien im Bereich der menschlichen Gesundheit vorgestellt.

Die Microarray-Plattform wurde verwendet, um die Genexpression von menschlichen Fibroblastenzellen von Patienten zu analysieren, die vom Hutchinson-Gilford-Progerie-Syndrom (HGPS) betroffen sind. Die Krankheit wird durch eine einzige genetische Mutation im LMNA-Gen verursacht, die bei Kindern verschiedene Symptome vorzeitiger Alterung hervorruft. Für die Analyse werden Fibroblastenzellen unterschiedlicher Herkunft verwendet: Fibroblasten von HGPS-Patienten, Fibroblasten von Gesunden, die mit UV-B-Licht behandelt wurden, Fibroblasten von HGPS-Patienten, die auf Telomerverlängerung behandelt wurden, und Fibroblasten von Gesunden als Kontrollgruppe. Ziel der Studie ist die Erstellung eines Transkriptomik-Profils von HGPS, d. h. die Identifizierung der transkribierten Gene und ihrer Regulierung, sowie der Vergleich mit anderen Zellseneszenzprozessen, wie z. B. den mit UV-B behandelten Zellen und der Gruppe der telomerelongierten Zellen. Die Ergebnisse zeigen, dass HGPS mehrere Gene beeinflusst, die mit den Kontrollpunkten des Zellzyklus, der Histonmodifikation und der Telomerase-Erhaltung zusammenhängen. Die gesunden UV-behandelten Zellen weisen Gene auf, die mit der Regulierung des Zellzyklus während der G1/S- und G2/S-Phasenübergänge und der zellulären Reaktion auf DNA-Schäden zusammenhängen. Die Auswirkung der Telomerverlängerung auf HGPS-Fibroblastenzellen betrifft verschiedene Stadien des Transkriptions-Translationsprozesses wie die RNA-Polymerase I, die RNA-Polyadenylierung und die mRNA-3'-Endverarbeitung. Außerdem zeigt die Analyse, dass HGPS- und UV-B-behandelte Zellen unterschiedliche Seneszenzprofile aufweisen. Die Telomerverlängerung bei HGPS lässt keinen entscheidenden Schluss auf die Fähigkeit zu, HGPS-Mängel zu beheben.

Es wurde eine umfassende Literaturrecherche durchgeführt, um Instrumente für das Repurposing von Arzneimitteln zu identifizieren, die Genexpressionsdaten zur Ermittlung potenzieller Arzneimittelbehandlungen nutzen. Die Ergebnisse der Überprüfung wurden dann

in der zweiten Studie der Dissertation angewandt, um potenzielle Medikamente gegen bakterielle und virale Infektionen beim Menschen zu identifizieren.

Die RNA-seq-Plattform wurde verwendet, um die Wirt-Pathogen-Interaktion durch den Einsatz des dualen RNA-seq-Ansatzes zu enthüllen. Diese Methode ermöglicht die gleichzeitige Untersuchung der Genexpression in zwei interagierenden Spezies ohne einen physischen Trennungsschritt während der Bibliotheksvorbereitung, sondern eine anschließende Trennung der Reads von jeder Spezies in silico mit bioinformatischen Tools. In der vorliegenden Arbeit wurde die duale RNA-seq-Methode angewandt, um die Interaktion zwischen menschlichen Zellen und zwei S. pyogenes (GAS) Bakterienstämmen (AP1 und NZ131) und dem Influenza A Virus (IAV) zu untersuchen. Ziel der Studie ist es, die Mechanismen aufzudecken, die IAV-infizierte Patienten anfälliger für S. pyogenes machen, die biologischen Gründe zu ermitteln, warum IAV-GAS-doppelinfizierte Patienten im Vergleich zu einfach infizierten Patienten schwerere Symptome zeigen, und mögliche medikamentöse Behandlungen gegen die beiden Erreger zu identifizieren. Die Analyse zeigt, dass alle drei Erreger eine Hochregulierung der ATP-Produktion und der Gene für Mitochondrienkomplexe aufweisen. Die mit S. pyogenes NZ131 infizierten Zellen weisen im Vergleich zu S. pyogenes AP1 eine höhere Infektionsrate auf. Die Ergebnisse zeigen auch, dass eine IAV-Infektion die sekundäre S. pyogenes-Infektion beeinflusst, indem sie die Stabilität der extrazellulären Matrix durch Zell-Zell-Adhäsion und die Herunterregulierung mehrerer Histon-Gene, die für ihre extrazellulären antimikrobiellen Eigenschaften bekannt sind, beeinträchtigt. Schließlich wurden im Rahmen einer Analyse zum Repurposing von Arzneimitteln 40 potenzielle Arzneimittel für die Behandlung einer IAV-GAS-Koinfektion in menschlichen Zellen ermittelt.

Chapter 1 - Introduction

1.1. Omics Technology

The last decade has been defined by the development of new scientific branches known informally as omics whose names end with the suffix -omics. The term “omics” is a derivation of the Greek word “ome” which means “complete”, “all”, or “whole”. The main concept of omics studies is a holistic view of all the molecules that make up a cell, tissue, or organism in a specific biological sample (Dai and Shen 2022). Omics technologies permit a better understanding of the complex system through the integration of all its constituents. These advancements in technology and analysis differ from traditional approaches, which are known to provide hypothesis-driven data or reductionistic. On the other hand, omics platforms are considered hypothesis-generating where the study has no prior hypothesis but all available data is attained and its study can suggest a new hypothesis that can be further tested (Iyer 2022).

The omics platforms can be divided into four main groups: genomics, transcriptomics, proteomics, and metabolomics (Iyer 2022). Genomics is the study of the entire genome of an organism. The genome is built upon the four nucleotides and its order reveals the information encoded in the DNA. One of the main components of the genome is the gene which represents a specific unit of DNA that is, in the most simple way, defined to hold the information to produce a specific functional unit called protein. The advent of high-throughput sequencing platforms has facilitated the analysis of variations between individuals at the genomics level for both disease-related mutations and characterizations of different human populations. Transcriptomics research studies the transcriptome that involves the entire collection of RNA molecules, called transcripts, in a cell or a specific sample. The next-generation RNA sequencing (RNA-Seq) technologies give deeper information concerning gene variations and their expression in different conditions. The third branch is Proteomics which studies the proteome, which describes the whole set of expressed proteins; and the interacting protein family networks and biochemical pathways used by the cell, tissue, or at the organism level. The last main sub-division is the Metabolomics branch which studies the metabolome generated within cells, tissues, or biofluids.

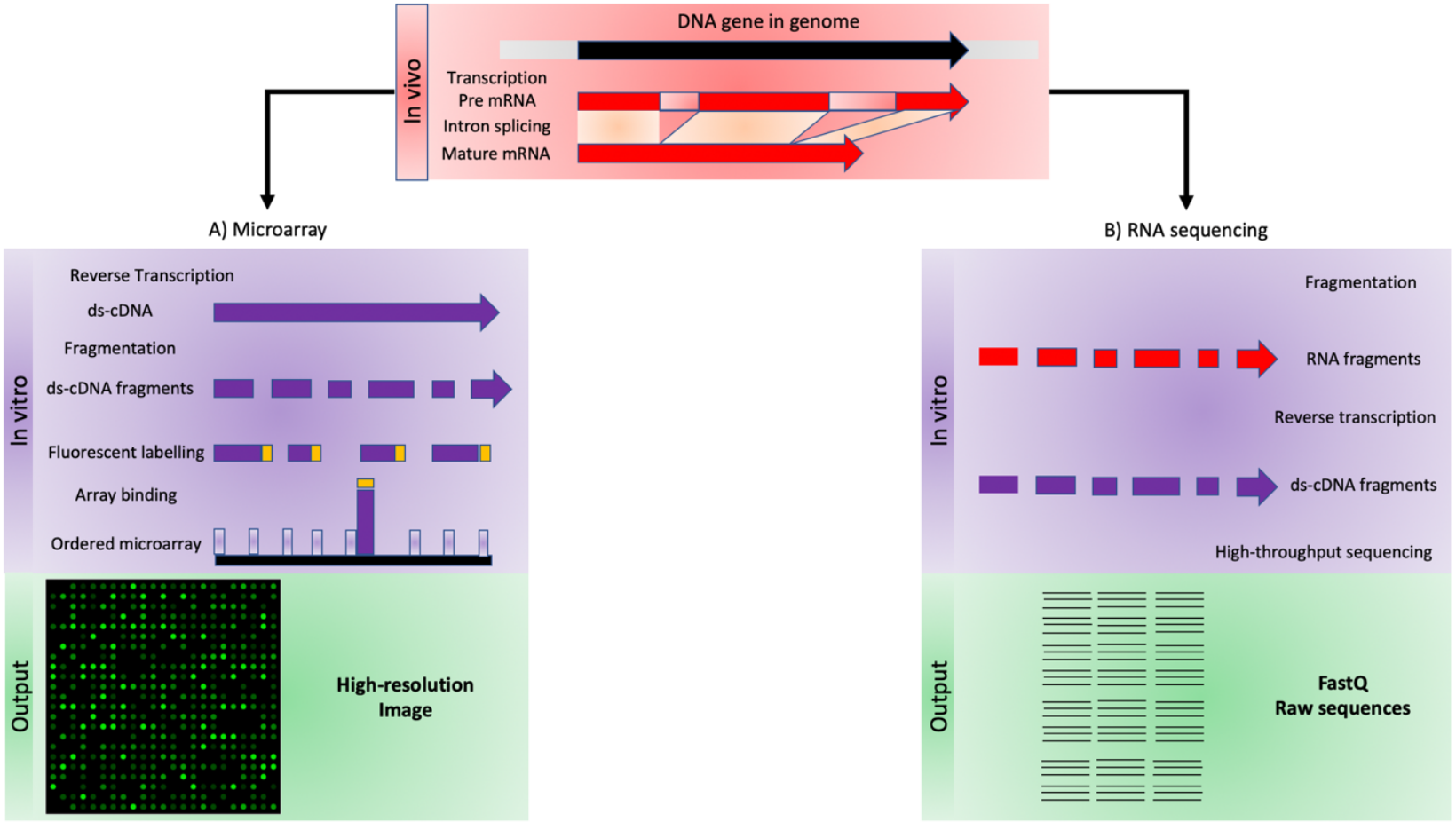
Chapter 1 - Introduction

The Metabolome is the collection of all the small molecules and their interaction within the system under study. The peculiarity of the Metabolome is its position as the final downstream product between the biological actors (such as genes and protein) and the environmental factors (such as sunlight and food). These qualities translate into a more chemically complex system than the other "omes" but it also represents the closest omics technology to the phenotype under study.

1.2. Transcriptomics

The main goal of transcriptomics analysis is the study of all the RNA transcripts from an organism (Lowe et al. 2017). The genetic information of an organism is organized inside the genome and expressed through the transcription process. The term "transcriptome" was first introduced in the late 1990s (Piétu et al. 1999; Velculescu et al. 1997). The technology has improved greatly from the earliest sequencing-based methods such as serial/cap analysis of gene expression (SAGE/CAGE), which used the Sanger sequencing platform to produce concatenated fragments (Velculescu et al. 1995). These transcripts were then quantified by matching the fragments to known genes. Nowadays, the usage of high-throughput sequencing platforms has overtaken transcriptomics platforms that rely on the Sanger approach (Lowe et al. 2017). There are two main contemporary actors in the field: Microarray (Figure 1.1A) and RNA-Seq (Figure 1.1B).

Figure 1.1 (A) Microarray. The mature mRNA molecule is subjected to a reverse transcription that results in the formation of double-strand complementary DNA (ds-cDNA) molecules. The ds-cDNA is fragmented and labelled with fluorescent molecules. The process is followed by the binding event between the fluorescent ds-cDNA and the attached oligos on the microarray surface. The presence and abundance of the transcripts are obtained through a high-resolution image obtained from the intensity of the fluorescent molecules. (B) RNA sequencing. The mature mRNA molecules are fragmented into RNA fragments. A reverse transcription step is applied to obtain ds-cDNA fragments. The fragments are then sequenced to produce FastQ files.



1.3. Microarray platform

Microarray technology quantifies a predetermined set of gene sequences from the genome (Bumgarner 2013). The technology involves the use of short nucleotide oligomers, known as "probes", which are fixed onto a solid substrate (e.g. glass). The generation of probes requires prior knowledge of the organism under study. Such knowledge can be acquired in the form of annotated genome sequence or the use of expressed sequence tags (ESTs) library to generate probes that are specific for one specific gene. The quantification process is achieved by the hybridization of fluorescently labelled transcripts (from the biological samples) to the arrayed probes. The fluorescent intensity at each probe position on the array specifies the abundance for that specific probe sequence.

Microarray platforms can be classified into two main groups: low-density spotted arrays or high-density short probe arrays (Heller 2003). Meanwhile, the presence of the transcripts can be registered with single- or dual-channel detection of fluorescent tags. Spotted low-density arrays are known to use picolitre drops of purified complementary DNA (cDNA), used as probes, and arrayed onto the surface of a glass slide (Auburn et al. 2005). The probes are typically longer than those used for high-density but they lack the high resolution of high-density arrays. Spotted arrays attach different fluorophores onto the transcripts from the test and control samples, and the ratio of fluorescent of the gene in the two conditions is used to determine a relative measure of abundance (Shalon, Smith, and Brown 1996). On the other hand, high-density arrays are based on the single-channel detection approach, and each sample is hybridized and detected individually (Lockhart et al. 1996). The Affymetrix GeneChip array (Santa Clara, CA) is based on the high-density approach, in which each transcript is quantified through the use of short 25-mer probes that collectively represents one gene (Irizarry, Bolstad, et al. 2003).

The final output of microarray data is recorded as high-resolution images that require feature detection and spectral analysis. The raw image files have a size of around 750 MB, whereas the processed intensities are about 60 MB in size. The image processing must correctly identify the features present in the image related to the regular grid and independently assign a quantification value to the fluorescence intensity for each feature. Furthermore, any image artefacts need to be found and removed from the overall analysis (Petrov and Shams 2004). Overall, the microarray technology directly correlates fluorescence intensities to the abundance of each sequence present on the array.

1.4. RNA sequencing (RNA-Seq) platform

RNA-Seq platform uses high-throughput sequencing technology to study the entirety of sequences, present in a biological sample, without prior knowledge of the organism (Stark, Grzelak, and Hadfield 2019). Furthermore, the use of computational methods permits the capture and quantify each present transcript in the sample. The length of sequenced nucleotides can range from 30 bp to 10,000 bp, based on the applied sequencing method. Nonetheless, most studies generate sequences around 100 bp in length (Stark, Grzelak, and Hadfield 2019). The deep sampling of the transcriptome during an RNA-Seq sequencing generates several short fragments that permit an in silico reconstruction of the original RNA transcript through the use of read alignment onto a reference genome or to each other (de novo assembly). The absence of prior knowledge of the genome of the organism under study presents additional uses of the platform outside the direct quantification of reads such as the identification of unknown isoforms and genes; and, the sequencing of specific reads (e.g. ChIP-seq analysis) (Park 2009). Since the advent of RNA-Seq in 2006 and 2008 (Bainbridge et al. 2006; Nagalakshmi et al. 2008), the capability of the platform has steadily improved regarding its throughput, accuracy, and read length; and, replacing microarrays as the main platform around the year 2015 (Su et al. 2014). The major advantages of RNA-Seq over microarray technologies are the deep sampling capability and the lack of prior knowledge constraint. The deep sampling approach permits the production of a dynamic range of 5 orders of magnitude of sequenced reads over microarray transcriptomes. Moreover, the amount of input RNA is lower for RNA-Seq (nanogram quantity) compared to microarrays (microgram quantity). The low quantity input of RNA-Seq consents to the study of subcellular structures and single-cell analysis when coupled with linear amplification of cDNA. Theoretically, the RNA-Seq technology does not present any upper limit of quantification, and the background signal for the typical 100 bp reads is quite low in nonrepetitive regions (Ozsolak and Milos 2011).

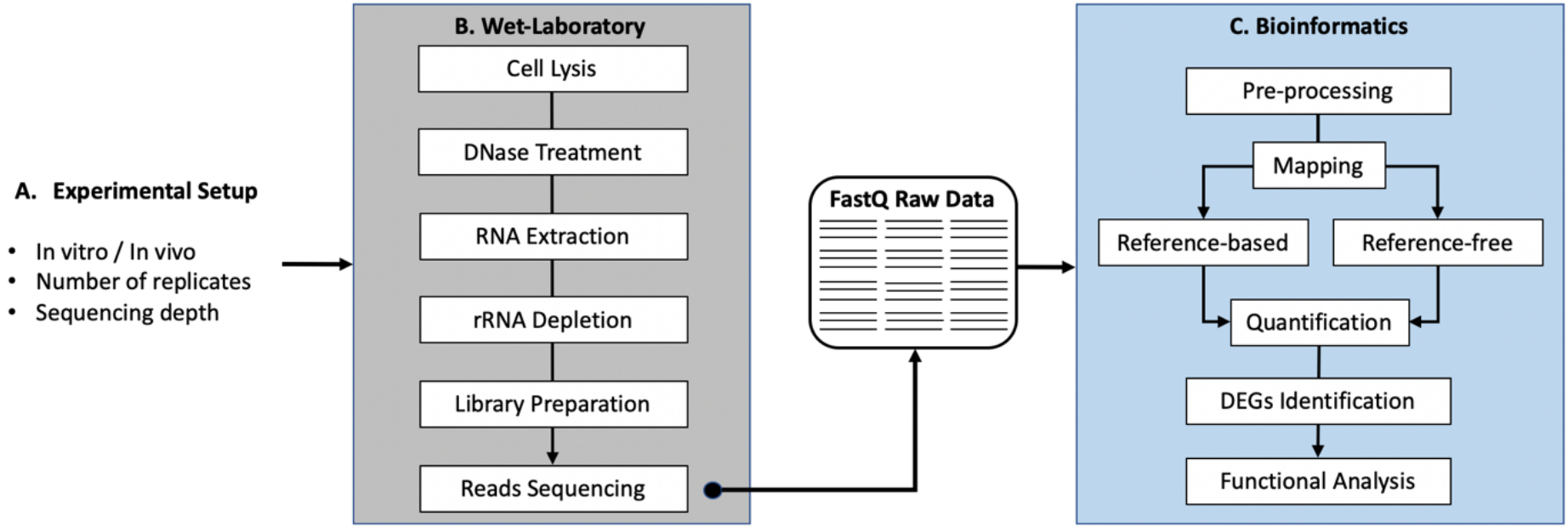
1.4.1. Data analysis

The final output of an RNA-sequencing process is the production of a FastQ file. The file has a text-based format and it displays both the biological sequence (mostly nucleotide sequence) and its corresponding quality scores. Each sequence is described by a four line-separated fields. The first line begins with the "@" symbol followed by a sequence identifier. The second line is the ASCII character of the raw sequence letters. The third line begins with the "+" symbol and is optionally followed by the same sequence identifier from the first line. The fourth line shows the quality score for the sequence in line 2, and it must contain the same number of

symbols as letters in the sequence (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847217/).

The in silico analysis of the FastQ raw reads is accomplished through several steps (Figure 1.2) with the aid of specific tools for each phase (Table 1.1 and Table 1.2). The RNA-seq pipeline is a methodology made of multiple steps that introduce human and technical errors that have an impact on the overall result (Figure 1.2). It is important to conceive a well-planned experimental design that will decrease biases and increase the overall likelihood of fulfilling the aim of the study. The first step of the analysis consists of a comprehensive literature search for the organism/s under study. The choice of carrying out the study in vivo or in vitro presents both advantages and disadvantages. In vitro protocol gives the ability to work in a more controlled environment but it also presents a simplified version of what happens in nature. In vivo protocol has the advantage of observing the biological process occurring in the right biological environment but the presence of different cell types creates a complex system where it is difficult to discern the effect of each type during the interaction. Such limitations can be overcome with a single-cell RNA-Seq protocol. It is vital to obtain an adequate quantity of mRNA molecules to capture the actual representation of the transcriptome. Furthermore, it is important to choose the right amount of sequencing depth and the number of replicates. As discussed by Liu and colleagues, a higher number of biological replications is a better option when compared to the increase of the sequencing reads (Y. Liu, Zhou, and White 2014).

Figure 1.2 Typical RNA sequencing pipeline. A) Experimental Setup. The first step of any omics study consists of proper planning of the study by considering the environment of the analysis (in vivo or in vitro), the number of replicates for each group, and the sequencing depth necessary to capture the biological diversity. B) Wet-Laboratory. The pipeline shows the main steps during the laboratory preparation that results in the sequencing of the mRNA molecules into FastQ data. C) Bioinformatics. The pipeline shows the main steps applied for the in silico analysis of the FastQ data.



1.4.2. Pre-processing

The initial step is the pre-processing phase which comprises the complete set of modifications of the FastQ raw data (Table 1.1). The FASTQ file format assigns a quality value, PHRED score, for each nucleotide present in the sequence (Cock et al. 2009). Throughout the years, different PHRED scores have been used; thus, it is important to identify the PHRED version and convert it when necessary. The conversion can be achieved by different tools such as fastq_quality_converter from the FASTX-toolkit (Hannon Lab 2009) and the FASTQ Groomer (Blankenberg et al. 2010). The quality of the reads is described by different parameters such as the quality score of the nucleotides throughout the sequence, sequence length variation, sequence duplication, contamination, GC content, etc. All of these checkpoints are vital to understanding the overall quality of the sequenced reads and finding any kind of error or bias which can influence downstream analysis. The purpose of the pre-processing is to discard adapter sequences and bases with low sequencing quality reads, to help mapping tools to achieve a better read mapping result (Conesa et al. 2016). However, the effectiveness of trimmed reads on the accuracy of downstream analysis is still unclear. Del Fabbro et al. identified a reduced number of reads mapped to annotated genes when read trimming was performed (Del Fabbro et al. 2013). Whereas, Didion et al. found that read trimming resulted in more reads mapping to annotated genes (Didion, Martin, and Collins 2017). Liao and Shi proved that the trimming process does not improve a mapping process produced by the tool Subread and its subsequent quantification (Liao and Shi 2020). Williams et al. discovered that the trimming step results in a reduced correlation of RNA-seq data to the microarray data (Williams et al. 2016). These contradicting studies showcase that trimming tools do have an impact on the mapping rate, accuracy, and speed. However, the effect is not universal but specific to the mapping tool itself and the minimum PHRED score used.

1.4.3. Alignment and Quantification

The following phase is the mapping step where the reads are aligned against a reference genome or transcriptome (Table 1.1). The tools used during the mapping phase can be divided into two main groups: alignment-based and alignment-free tools.

The alignment-based tools map the reads to a reference genome, such that relative gene expression levels can be inferred by the alignments at annotated gene loci (Trapnell et al. 2012); and, it produces a BAM file as a result of the mapping process. In the last decade, several alignment-based algorithms have been developed to replace traditional aligners such as BLAST (Camacho et al. 2009) and BLAT (Kent 2002). The advantage of the new generation aligners is the ability to analyse a high amount of sequencing data in a relatively short amount of time. Most modern aligners consider intronic regions and allow split-read during the alignment. Tools like BWA (H. Li and Durbin 2009), Bowtie2 (Langmead and Salzberg 2012), TopHat2 (D. Kim et al. 2013), and HISAT2 (D. Kim et al. 2019) are based on Burrows-Wheeler transformation methods and use the seed-extend based mapping strategy. Other aligners like STAR make use of suffix arrays as the index of the reference (Dobin et al. 2013). Similarly, Segemehl utilizes a multi-split-read aligner approach based on an enhanced suffix array. The tool is capable of mapping short and it provides a suitable tool when studying circular RNAs (Hoffmann et al. 2014). The quality control of the alignment can be checked by the following tools: RSeQC (Liguo Wang, Wang, and Li 2012), IGV (J. T. Robinson et al. 2017), Picard toolkit (Broad Institute 2019), Qualimap (Okonechnikov, Conesa, and García-Alcalde 2016), and SAMtools (H. Li et al. 2009). The tools calculate metrics such as genome coverage and the overall mapping quality achieved. The quantification tools (Table 1.1) uses the alignment-based output files (BAM/SAM) to quantify genes and/or transcript.

The alignment-free tools can be defined as any method of quantifying RNA sequence without the use or production of alignment files (BAM/SAM files) at any point during the algorithm application (Zielezinski et al. 2017). Alignment-free tools require a reference transcriptome which can be obtained in different ways. First, by a public database such as Ensembl (Yates et al. 2020) or the NCBI (Bethesda (MD): National Library of Medicine (US) 1988). Second, tools such as RSEM can be used to create a transcriptome by using the genome and its annotation (B. Li and Dewey 2011). In case neither genome nor transcriptome is available, the pre-processed reads can be used to create a de novo transcriptome by tools such as Trinity (Grabherr et al. 2013). Alignment-free tools do not depend on any knowledge of the evolutionary history of sequence changes. The speed advantage of alignment-free methods (Everaert et al. 2017) and their accuracy level comparable to alignment-based tools (Jin, Wan,

and Liu 2017) can be observed in several other fields such as expression profiling (Bray et al. 2016; Rob Patro et al. 2017); genetic variant calling (Pajuste et al. 2017; Rudewicz et al. 2016; Shajii et al. 2016; You Li et al. 2017); de novo genome assembly by long-read sequencing platforms (H. Li 2016; Berlin et al. 2015; Warren et al. 2015); phylogenetic reconstruction (Gardner, Slezak, and Hall 2015; Ren et al. 2016; Fan et al. 2015); and taxonomic classification in metagenomics studies (Wood and Salzberg 2014; Ounit and Lonardi 2016; Ames et al. 2013; A. Gupta, Jordan, and Rishishwar 2017; Roosaare et al. 2017). On the other hand, Wu et al highlight a clear deficiency of alignment-free for accurate quantification of small and lowly-expressed RNA quantification (D. C. Wu et al. 2018; Nottingham et al. 2016). Furthermore, Wu et al. demonstrated that both alignment strategies give similar results for differential expressed genes (DEGs) detection, regardless of the difference during the quantification phase (Everaert et al. 2017; Sahraeian et al. 2017).

1.4.4. Differential Expressed Genes Identification

The identification of DEGs between two sample groups is one of the most important results after an RNA-seq pipeline (Table 1.1). The initial step of the analysis is the normalization which removes biases and artefacts caused by the different sequencing depths of the libraries, length variations expressed by each isoform of the gene, by-products produced during the sequencing, and the tools used during the bioinformatics pipeline (Abrams et al. 2019). The normalization can be achieved with different measures such as RPKM (Reads Per Kilobase per Million mapped reads), FPKM (Fragment Per Kilobase per Million mapped reads), and TPM (Transcripts per Kilobase Million). Several studies have highlighted TPM as the best parameter over other normalization methods in host-pathogen publications (Klassert et al. 2017; Riege et al. 2017); and, also in benchmark analyses for other application areas (Abrams et al. 2019). DEGs tool selection can be quite laborious since each one uses a different normalization and statistical approach to find the DEGs. As shown by Costa-Silva et al., the combination of 3 to 5 tools is the right number to ensure an analysis with lower false positives (Type 1 errors) and false negatives (Type 2 errors); and, the combination of DESeq2, edgeR, limma+vomm, baySeq, and NOISeq was found to give the best result among all the other tools (Costa-Silva, Domingues, and Lopes 2017).

Table 1.1 List of tools used during the pre-processing, the mapping, quantification, and differential expressed genes during the RNA-seq analysis.

Type	Tool	Reference
Pre-processing	CutAdapt	(M. Martin 2011)
	FASTQ Groomer	(Blankenberg et al. 2010)
	FastQ Trimmer	(Blankenberg et al. 2010)
	FASTX-Toolkit	(Hannon Lab 2009)
	PRINSEQ-lite	(Schmieder, Edwards, and Bateman 2011)
	Scythe	(Buffalo 2011)
	Sickle	(Najoshi and Fass 2011)
	Trim Galore!	(Krueger 2016)
	Trimmomatic	(Bolger, Lohse, and Usadel 2014)
Pre-processing Quality Control	FastQC	(Andrews S. 2010)
	MultiQC	(Ewels, Lundin, and Max 2016)
	FastX-toolkit	(Hannon Lab 2009)
	SeqPrep	(John 2010)
	PEAR	(J. Zhang et al. 2014)
Mapping	Bowtie2	(Langmead et al. 2019)
	TopHat2	(D. Kim et al. 2013)
	BWA	(H. Li and Durbin 2009)
	ELAND	(Cox 2007)
	GLINT	(Rahmani et al. 2017)
	GSNAP	(T. D. Wu and Nacu 2010)
	Hisat2	(D. Kim et al. 2019)
	Kallisto	(Bray et al. 2016)
	Megablast	(Morgulis et al. 2008)
	NextGenMap	(Sedlazeck, Rescheneder, and Haeseler 2013)
	SOAP2	(R. Li et al. 2009)
	STAR	(Dobin et al. 2013)
	Subread	(Liao, Smyth, and Shi 2013)
	RSEM	(B. Li and Dewey 2011)
	Trinity de novo	(Grabherr et al. 2013)
	BBMap	(Bushnell et al. 2014)
	BLAST+	(Camacho et al. 2009)
	CD-HIT-EST	(W. Li and Godzik 2006)
	Oases	(Schulz et al. 2012)
	Salmon	(Robert Patro 2020)
	Velvet	(Zerbino 2010)
	Cufflinks	(Trapnell et al. 2010)
	READemption	(Förstner, Vogel, and Sharma 2014)
Mapping Quality Control	IGV	(J. T. Robinson et al. 2017)
	SAMtools	(H. Li et al. 2009)
	RSeQC	(Liguo Wang, Wang, and Li 2012)
	Picard	(Broad Institute 2019)
	Qualimap	(Okonechnikov, Conesa, and García-Alcalde 2016)
	RNA-SEQC	(Deluca et al. 2012)
Quantification	featureCount	(Liao, Smyth, and Shi 2014)
	GLINT	(Rahmani et al. 2017)
	HTSeq-count	(Anders, Pyl, and Huber 2015)
	Lox	(Z. Zhang, López-Giráldez, and Townsend 2010)
	READemption	(Förstner, Vogel, and Sharma 2014)
	Kallisto	(Bray et al. 2016)
	RSEM	(B. Li and Dewey 2011)
	Salmon	(Robert Patro 2020)

Quantification Quality Control	Scotty	(Busby et al. 2013a)
Differential Expressed Genes	baySeq	(Hardcastle 2016)
	Cuffdiff	(Trapnell et al. 2012)
	DEBrowser	(Kucukural et al. 2019)
	DESeq2	(Love, Huber, and Anders 2014)
	EBseq	(Leng et al. 2013)
	edgeR	(M. D. Robinson, Mccarthy, and Smyth 2010)
	NOISeq	(Tarazona et al. 2015)
	Limma+voom	(Law et al. 2014)
	SAMseq	(J. Li and Tibshirani 2013)
	SARTools	(Varet et al. 2016)
	Sleuth	(Bray et al. 2016)
	TCC	(Sun et al. 2013)

1.4.5. Functional analysis

The goal of the functional analysis is to extract significant biological information from the list of identified DEGs (Table 1.2).

One of the main analysis is the identification of enriched pathways from the DEGs. Several tools have been created to identify biological pathways and other tools to summarise and visualize the result (Ackermann and Strimmer 2009). Furthermore, several studies incorporate different types of protein analysis (e.g. sub-cellular location), transcription factor binding site search, sequence alignment, splicing analysis, hormone discovery, pathogen identification, taxonomic analysis, and network creation.

Table 1.2 List of tools used during the functional analysis at the end of an RNA-seq analysis.

Functional Analysis	**Tool**	**Reference**
Enrichment Analysis	amiGO	(Carbon et al. 2009)
	BayGO	(Vêncio et al. 2006)
	Blast2GO	(Conesa et al. 2005)
	CateGOrizer	(Hu et al. 2008)
	DAVID	(Dennis et al. 2003)
	Enrichr	(Kuleshov et al. 2016)
	FungiFun	(Priebe et al. 2011)
	g:Profiler2	(Kolberg et al. 2020)
	GOEAST	(Zheng and Wang 2008)
	Goseq	(Young et al. 2010)
	GSEA	(Subramanian et al. 2005a)
	KOBAS	(Xie et al. 2011)
	ROntoTools	(C. Voichita, S. Ansari 2020)
	SeqEnrich	(Becker et al. 2017)
Enrichment Analysis & Visualisation	REVIGO	(Supek et al. 2011)
	BiNGO	(Maere, Heymans, and Kuiper 2005)
	GOATOOLS	(Klopfenstein et al. 2018)
	clusterProfiler	(G. Yu et al. 2012)
	GOrilla	(Eden et al. 2009)
	Cytoscape	(Shannon et al. 2003)
	STRING	(Szklarczyk et al. 2019)

Protein Analysis	ARGOT2	(Falda et al. 2012)
	BLASTP	(Camacho et al. 2009)
	BLASTX	(Camacho et al. 2009)
	ConFunc	(Wass and Sternberg 2008)
	I-TASSER	(Y. Zhang 2008)
	Phyre	(Kelley et al. 2015)
	SignalP	(Almagro Armenteros et al. 2019)
	TargetP	(Emanuelsson et al. 2000)
	TMHMM	(Krogh et al. 2001)
	ToppGene	(J. Chen et al. 2009)
	WoLF PSORT	(Horton et al. 2007)
Gene Network Creation	BioSankey	(Platzer et al. 2018)
	GeneMania	(Franz et al. 2018)
	KAAS	(Moriya et al. 2007)
	MapMan	(Schwacke et al. 2019)
	NetGenerator	(Weber et al. 2013)
	WGCNA	(Langfelder and Horvath 2008)
Transcript Factor Binding analysis	MEME	(Bailey et al. 2009)
Alignment analysis	TOMTOM	(S. Gupta et al. 2007)
	CLUSTALW	(Thompson, Higgins, and Gibson 1994)
	MUSCLE	(R. C. Edgar 2004)
	SPADA	(P. Zhou et al. 2013)
	DIAMOND	(Buchfink, Xie, and Huson 2014)
	RNAstar	(Widmann et al. 2012)
Pathogen Identification	Pandora	(Colquhoun et al. 2021)
	Pathoscope	(Hong et al. 2014)
	RNA CoMPASS	(Xu et al. 2014)
Taxonomic Analysis	MEGA5	(Tamura et al. 2011)
	MEGAN4	(Huson et al. 2011)
Statistical Analysis	MaAsLin2	(Mallick et al. 2021)
Splicing Analysis	MISO	(Katz et al. 2010)
Hormone discovery	HORMONOMETER	(Volodarsky et al. 2009)
Package	GenePattern2	(Reich et al. 2006)
	Metascape	(Y. Zhou et al. 2019)
	MicroScope	(Vallenet et al. 2020)
	Useq	(Nix, Courdy, and Boucher 2008)
	GATK	(McKenna et al. 2010)
	SAMtools	(H. Li et al. 2009)
	BLAST+	(Camacho et al. 2009)

1.5. Transcriptomics applications discussed in the present thesis

The last decade has seen an explosion of transcriptomics data and its usage in different research fields over traditional methods due to its hypothesis-generating property. These research fields include the identification of transcriptional start sites and splicing alterations (Costa et al. 2013); the cellular response to abiotic changes; the annotation of gene function; the study of non-model organisms with non-existing or poorly available genomic resources; and, the study of non-coding RNA biological functions (e.g. impact in protein translation, RNA splicing, and transcriptional regulation (Noller 2003; Kishore and Stamm 2006; Hüttenhofer, Schattner, and Polacek 2005).

The present dissertation focuses on three transcriptomics areas of research, each discussed in their chapter and briefly introduced in this sub-chapter.

First, the use of the microarray platform to study the transcriptome signature from fibroblast cell lines affected by three different senescence conditions: disease-associated single nucleotide polymorphisms (SNP) mutation in the LMNA gene that results in Hutchinson-Gilford Progeria Syndrome (HGPS); telomere elongation on HGPS cells; and, cells treated with the UV-B light (Chapter 3 and Sub-Chapter 1.5.1.).

Second, drug repurposing analysis is a vast research field with great importance for pharmaceutical research. The improvement of transcriptomics technologies coupled with previous technology and knowledge has brought the development of several tools capable of linking genes and their expression data to drugs that are capable of reversing or mimicking the input (Chapter 4 and Sub-Chapter 1.5.2.).

Third, the study of host-pathogen interaction with the RNA-Seq platform is a novel approach termed “dual RNA-Seq”, which provides a simultaneous snapshot at the transcriptomics level for both organisms. This new methodology was applied to study the interaction between human epithelial cells and the Influenza A virus and strains from the Streptococcus pyogenes bacteria (Chapter 5 and Sub-Chapter 1.5.3.).

1.5.1. Diagnostic and disease profiling

One of the main areas of transcriptomics study is its application to disease-associated caused by genetic mutations such as SNP, allele-specific expression, and gene fusions, contributing to the understanding of disease causal variants. The transcriptomic data provides the first biological instance of the effect of genetic mutations in the genome and generates hypotheses that can provide insight into the biology of the disease. One prime example is the Progeria

syndrome (discussed in the present dissertation) caused by a single point mutation in the lamin A (LMNA) gene that results in a progressive disease that causes children to age rapidly.

1.5.2. Drug-induced gene expression database

Drug repurposing analysis is based on the identification of new uses for commercial and/or studied drugs. In the last decades, this strategy has been applied with the help of transcriptomics data with a new methodology called transcriptomics matching. The methodology consists of comparing multiple transcriptomics signatures among each other to identify the ones that have similar or reverse the behavior. The matching process relies on signatures present in public databases such as the Connectivity Map (cMap) which was established in 2006 by the Broad Institute. The project consists of an ensemble of gene expression profiles obtained from testing more than 1300 drugs against different human cell lines (J. Lamb et al. 2006). The drug-induced gene expression database can be viewed as a simplified proxy for phenotypic screening for a large number of compounds and it has been a successful instrument for drug repurposing for different disease profiles. The updated installment of the cMap data repository (cMap 3.0) is accessible at the US National Institutes of Health Library of Integrated Network-based Cellular Signatures (LINCS, https://lincsproject.org/). The new depository includes transcriptional signatures produced by tens of thousands of drugs dosed upon hundreds of human cell lines. Another advantage of cMap and LINCS is the ability to use their resource alongside other public repositories of transcriptomics data, such as the Gene Expression Omnibus (GEO) and Array Express, which contain raw gene expression data from hundreds of diseases afflicting humanity and tens of animal models used in the research process. Due to these advantages, manual curation and dedicated software (Zichen Wang et al. 2016) have been created to associate raw signatures from public repositories like GEO to drug-induced databases to find novel drug-disease connections and potential drug treatments (as discussed in Chapter 4).

1.5.3. Host-Pathogen interaction

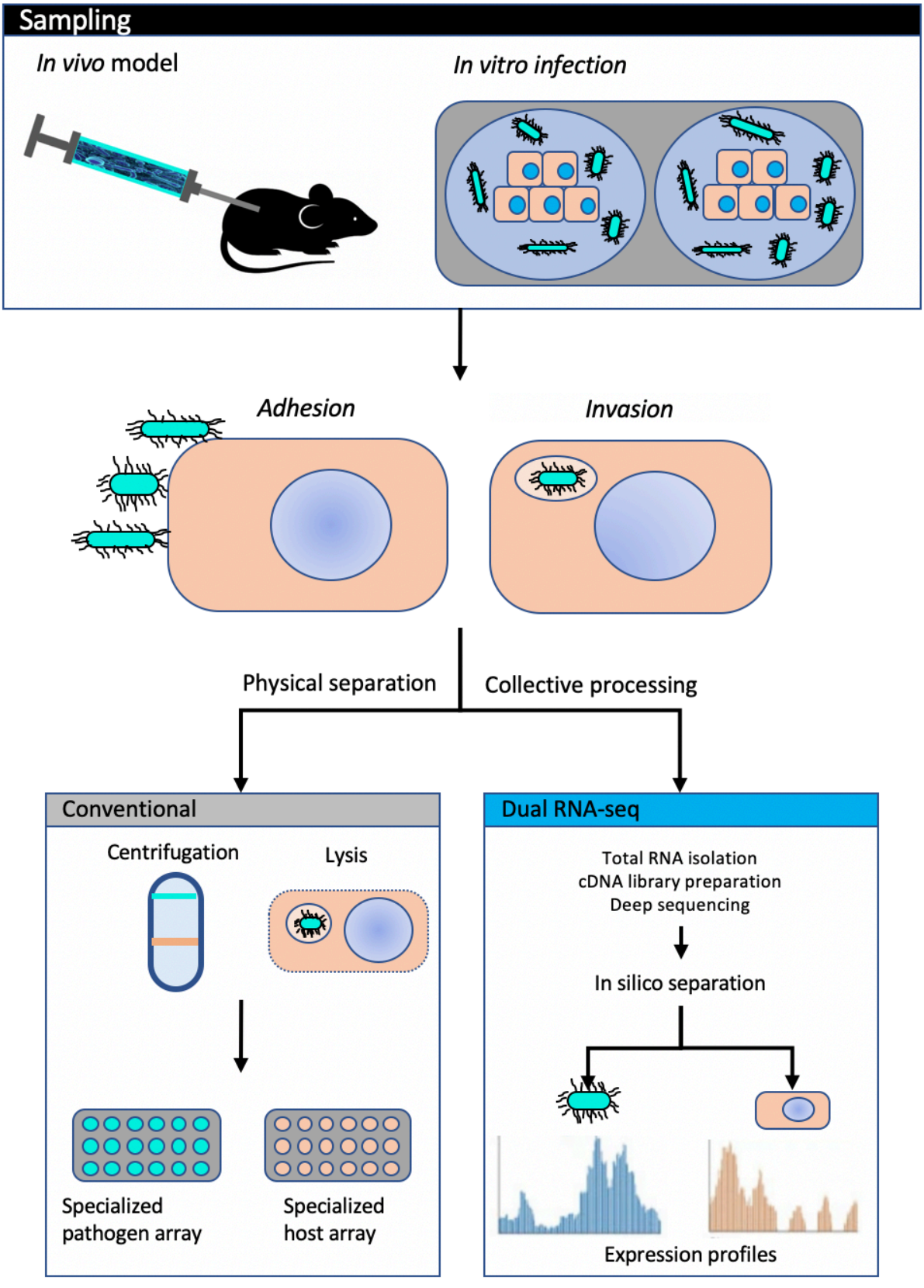
The rise of omics technology has resulted in the advancement of established research field capabilities, such as genome sequencing, and the development of new research domains like RNA-seq platforms. Advancement in genome sequencing is shown in Metagenomics studies where genetic material from environmental samples is used to identify the organisms present in such samples. The environmental samples can range from soil samples to the human gut (Daniel 2005; Qin et al. 2010; Gilbert and Dupont 2010). However, species identification does not provide information concerning the interaction between the organisms. Such a goal can

only be achieved by studying dynamic processes such as the transcriptome which offers a snapshot of the expressed genes from each present organism.

The RNA-Seq platform has become the standard for transcriptomics study over traditional technologies such as the microarray (Lowe et al. 2017). The microarray methodology is hindered by a long list of problems like high background noise, availability for only sequenced genomes, high cross-hybridization, and an expensive protocol (Shendure 2008; Bumgarner 2013). All these issues are largely solved by RNA-Seq, with its independence from prior knowledge of the genome sequence, minimal background noise during the sequencing step, and deeper coverage of the reads (Zhong Wang, Gerstein, and Snyder 2009; Marioni et al. 2008; Fu et al. 2009). Furthermore, the RNA-Seq platform presents additional protocols that can be used to tailor the platform based on the goals of the study. The strand-specific protocol provides the ability to accurately quantify gene expression from the organism that presents high gene density; overlapped genes; non-coding and antisense transcript; and the presence of operons in the genome (Hrdlickova, Toloue, and Tian 2016; Zhao et al. 2015). Additionally, the paired-end option presents an ideal system for transcriptome assembly (Corley et al. 2017). In conclusion, the high sensitivity of the RNA-Seq platform increases the likelihood of finding novel transcripts, alternative splicing events, and transcript borders compared to traditional transcript sequencing platforms (Costa-Silva, Domingues, and Lopes 2017).

Due to RNA-Seq positive traits, the platform has become the prevailing methodology used to study the interaction between different organisms. The term "dual RNA-seq" term describes studies, where the two interacting organisms are simultaneously subjected to the same RNA-Seq wet-lab protocol, followed by an in silico separation of the RNA, reads to their specific species (Westermann, Gorski, and Vogel 2012) (Figure 1.3). Dual transcriptome studies permit a better understanding of the interaction for both species compared to a canonical RNA-seq where the library preparation is focused on only one of the species (Frönicke et al. 2018).

Figure 1.3 Conventional RNA-Seq versus dual RNA-Seq pipeline. The two pipelines share the same sampling methodologies but the dual RNA-Seq analysis does not separate the organisms (collective processing) thus creating a library that consists of mRNA molecules from both organisms. The separation of the mRNA molecules to their specific species is achieved at the in silico level.



Chapter 2 - Aims and Objectives

The transcriptomics revolution has been an important key for biological research in the last decade due to its vast applications and simplicity over traditional methods. The present dissertation aims to illustrate the usage of transcriptomics data related to the understanding of human healthcare in regard to human genetic disease and human infection processes caused by viruses and bacteria; and, the discovery of new applications for approved or investigational drugs that are outside of the scope of its original medical purpose. The following chapters are centred around the three research fields I worked on for the thesis:

Chapter 3 focuses on the analysis of microarray data obtained from patients affected by the Hutchinson-Gilford Progeria syndrome (HGPS) (Sinha, Ghosh, and Raghunath 2014). HGPS is an extremely rare and progressive genetic disease that causes children to age rapidly, starting in their first and second years of life. The disease is caused by a single mutation in the LMNA gene, which encodes for the nuclear matrix lamin A. The abnormal version of the protein results in an unstable mechano-state of the nucleus and the cell itself. This unsteady state appears to be the main culprit that promotes the aging symptoms in progeria patients (Sinha, Ghosh, and Raghunath 2014). The goal of the study is to identify the transcriptomics imprint of HGPS patients compared to healthy normal people; the differences between HGPS senescence state and the senescence state caused by UV-B in healthy cells; and, the potential health benefits of telomere elongation in progeria cells.

Chapter 4 provides an overview of the current state of drug repurposing tools that rely on the gene, gene list, and expression data as their input. The chapter divides the software based on the following parameters: input and output type; sourced database used for the analysis; the platform the tool can be used (e.g. webpage); and, a specialized category based on specific characteristics of the tool (e.g. statistical methodology).

Chapter 5 illustrate the use of the RNA-Seq platform to study host-pathogen interaction between epithelial human cells and influenza A virus (IAV) and two serotypes of group A Streptococcus pyogenes (GAS). The study was conducted to understand the role of IAV infection in the increased rate of secondary GAS infection (Herrera, Huber, and Chaussee 2016). Furthermore, drug repurposing analysis was conducted with the infected transcriptomics profile, from the human host, to identify potential drugs capable of aiding the host defence against both IAV and GAS pathogens.

Chapter 3 - Hutchinson-Gilford Progeria Syndrome

3.1. Introduction

The Hutchinson-Gilford progeria syndrome (HGPS) is an autosomal dominant, fatal pediatric premature aging disease, without gender or ethnic propensity and with complete penetrance (Sinha, Ghosh, and Raghunath 2014). The first portrayal of the disease happened in 1886, by Jonathan Hutchinson (Hutchinson 1886), and a second time by his colleague Hastings Gilford who coined the condition progeria (premature aged) in 1904 (Gilford and Hutchinson 1897). The syndrome has onset in early childhood (around the first year of life) with an estimated incidence of 1 in 4 million births, and an average lifespan of 13.5 (Sinha, Ghosh, and Raghunath 2014).

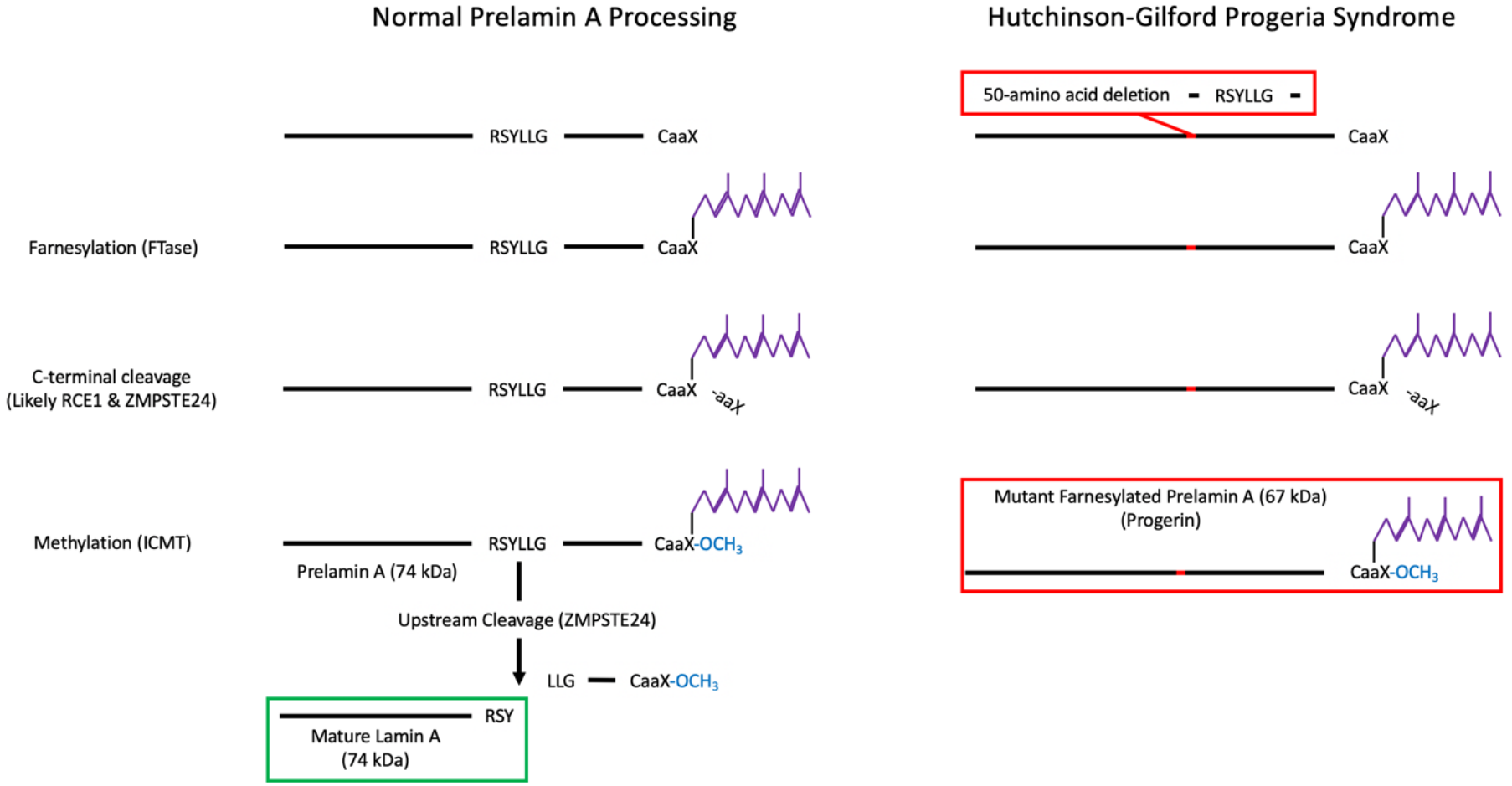
HGPS presents a wide range of symptoms that can vary based on the age of onset and the degree of severity; nonetheless, clinical features are consistent and the patients appear very in phenotype (Merideth et al. 2008). The patients present birth weight and early postnatal development are normal when compared to the healthy population. The disease presents many, but not all, clinical features of aging/senescence. Normal aging phenotypes exhibited by HGPS patients include alopecia, joint contractures, low bone density, lipoatrophy with limb wasting, and global atherosclerosis. On the other hand, symptoms that differ from aging include growth failure, skeletal dysplasia, and lack of pubertal development. In absence of strokes, HGPS patients show motor and intellectual development is typically normal. Dementia, osteoarthritis, and cancer are absent. The efficiency of the immune system is considered normal, as is wound healing. The integrity of organs such as the liver, kidney, and gastrointestinal systems remains intact. The major cause of morbidity and mortality among HGPS patients are cardiovascular disease and cerebrovascular disease (Gordon et al. 2014).

The syndrome is caused by a *de novo* mutation in the LMNA gene either within exon 11 (classic form) or at the exon 11 intronic border (atypical form) that results in the expression of a 50 amino-acid deletion version of the LMNA protein (Figure 3.1) (Ullrich and Gordon 2015). The shortened mutant version of the LMNA protein (progerin) does not impede its localization to the nucleus nor its interaction with the normal LMNA protein, because the necessary components for these functions are still present (Sinensky et al. 1994). Nonetheless, the deletion does remove the recognition site involved in the dissociation and reassociation of the nuclear membrane at each cell division (Sinensky et al. 1994; Kilic et al. 1997).

The expression of the LMNA gene is developmentally regulated and displays cell and tissue specificity, mostly in differentiated cells including fibroblast, vascular smooth muscle cells, and vascular endothelial cells (Machiels et al. 1996; Tilli et al. 2003; McClintock, Gordon, and Djabali 2006). The lamin proteins are the main constituents of the nuclear lamina, which functions as an interface between the inner membrane of the nuclear envelope and the chromatin (Goldman et al. 2004). The structural integrity of the lamina complex has relevant importance for the cell in several biological processes such as mitosis, creating and maintaining the structural stability of the nuclear scaffold, DNA replication, RNA transcription, organization of the nucleus, nuclear pore assembly, chromatin function, cell cycling, and apoptosis. Furthermore, HGPS patients show a shortened length of the telomere when compared to healthy age-matched people (Decker et al. 2009).

Chapter 3 – Hutchinson-Gilford Progeria Syndrome

Figure 3.1 Biogenesis of lamin A in normal cells and in Hutchinson-Gilford progeria cells. Normal prelamin A process produces a 74 kDa protein obtained from the upstream cleavage of the farnesyl-prelamin A. In HGPS cells, a 50-aa deletion in the prelamin A protein (aa 607-658) removes the site of recognition for the second endoproteolytic cleavage. Therefore, no mature lamin A is produced, and a farnesylated mutant prelamin A (progerin) accumulates in cells.



3.2. Aims

The understanding of HGPS is mired by different shortcomings such as the use of progeroid mouse models that do not produce progerin (Ullrich and Gordon 2015), and the complex role of the LMNA protein and the lamina complex in human cell homeostasis. The microarray platform was chosen to study the gene expression of human fibroblast cells from HGPS patients. The goal of the present study is to have a clearer understanding of the effect of progerin; the transcriptomics differences between the senescence process caused by HGPS patients and the senescence process caused by UV-B light treatment; and, to analyse the effect of telomere elongation in HGPS fibroblast cells.

3.3. Methods and Materials

3.3.1. Materials: Human fibroblast cell groups

The study analyses 24 CEL files obtained from four fibroblast cell line groups, each with six biological replicates. The "Control" group was obtained from the healthy population. The "HGPS" group was acquired from HGPS patients. The "UV-B" group describes healthy cells that have been treated with UV-B light. The last group is the "HGPS-TERT" which describes HGPS samples that have undergone telomere elongation. The laboratory experimental work was carried by our collaborative partners of the Institute of Clinical Chemistry and Laboratory Medicine at the University of Rostock under Professor Michael Walter.

3.3.2. Methods: Microarray CEL files analysis

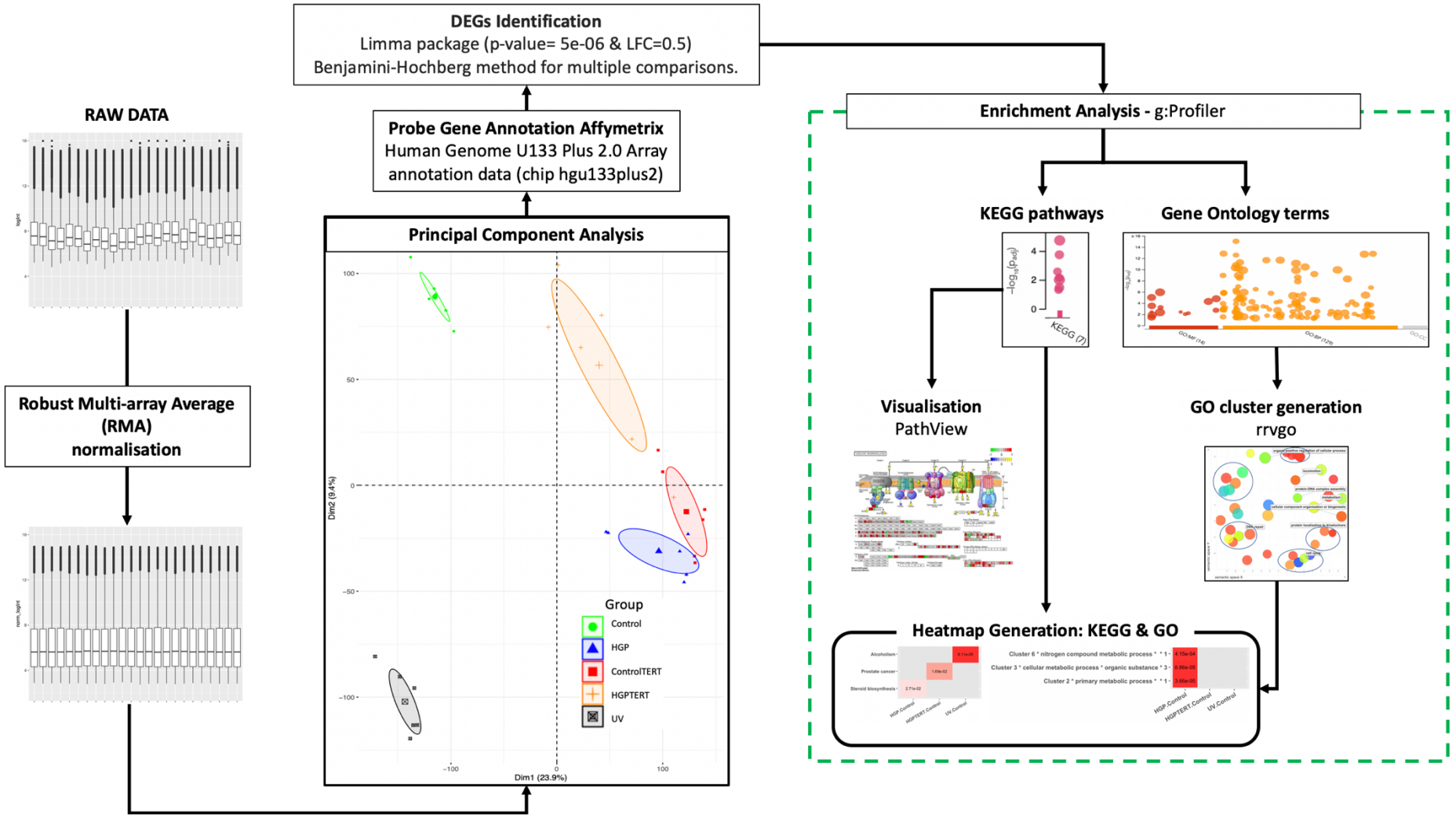
The microarray analysis of the CEL files was carried out in R (Table 3.1). The raw CEL files were normalized using the Robust Multi-array Average (RMA) method (Irizarry, Hobbs, et al. 2003). The normalization process is applied to ensure meaningful statistical analyses and inferences from all the samples. The RMA method consists of four steps: background correction for each array; quantile normalisation across the array to make all distributions the same; probe level intensity calculation through log transformation of the background corrected and normalised intensity value; and, probe set summarisation where the intensities of the probe values are combined into one.

Table 3.1 List of R package used during the CEL microarray analysis.

R Package	Description	Reference
Oligo	Pre-processing tools for oligonucleotide arrays	(Carvalho and Irizarry 2010)
Affycoretools	Functions for repetitive analyses with Affymetrix GeneChips	(MacDonald 2022)
Tidyverse	Collection of R packages designed for data science	(Wickham et al. 2019)
Factoextra	Multivariate Data Analyses and Elegant Visualization	(Alboukadel Kassambara and Fabian Mundt 2020)
hgu133plus2.db	Affymetrix Human Genome U133 Plus 2.0 Array annotation data	(Carlson 2016)
gProfiler2	Gene list functional enrichment analysis	(Raudvere et al. 2019)
rrvgo	Reduce + Visualize GO	(Sayols 2020)
Tidytext	Text mining tasks and plot generation.	(Silge and Robinson 2016)
gridExtra	Arrange multiple grid-based plots on a page	(Auguie 2017)
Scales	Graphical scales map data to aesthetics	(Hadley Wickham and Dana Seidel 2020)
Pathview	KEGG pathway based data integration and visualization	(Luo and Brouwer 2013)
ggvenn	Venn diagram creation	(Yan 2021)

After the normalisation step, the principal component analysis (PCA) was conducted to increase the readability of the data and to understand the similarity between the samples. The PCA plot does show samples from the same group to cluster closer to each other compared to other sample groups (Figure 3.2). The gene annotation of the probes was carried out by the Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2) from the Bioconductor package (Carlson 2016). Differentially expressed probes were identified by the Limma package (Ritchie et al. 2015); and, Benjamini-Hochberg for multiple comparisons: p-value < 0.0002 and Log2 Fold Change (LFC) > |0.5| (S. Y. Chen, Feng, and Yi 2017). The chosen adjusted p-value threshold was calculated as the lowest value which resulted in a number of False Positives that is less than 1 (adjusted p-value * number of DEG < 1). The DEGs were identified by the list of annotated differential expressed probes. DEGs with probes at opposing fold change values were removed from the DEG list. The g:Profiler tool was used to identify the enrichment terms (Raudvere et al. 2019) and the rrvgo R package was used to reduce redundant enriched terms (Sayols 2020). Furthermore, the Pathview package (Luo and Brouwer 2013) was used to show the expression of the genes in the enriched KEGG pathway. The R script used for the CEL microarray analysis can be found at https://github.com/SalemSueto-BioInfo/MicroArray_Affymetrix_Analysis.

Figure 3.2. Microarray CEL files Bioinformatics pipeline. The CEL raw data is RMA normalised. Followed by the PCA to check the distances of the samples. The annotation of the probes is achieved by the Human Genome U133 Plus 2.0 Array annotation data. Then, the DEGs are found by using the limma package (p-value = 5e-06 and LFC = |0.5|). The enrichment analysis is achieved by using the g:Profiler tool. The enriched KEGG pathways are visualised with PathView R tool. Meanwhile the GO clusters are generated with the rrvgo and summarised with heatmap plots.



3.4. Results

3.4.1. Bioinformatics CEL analysis

The normalisation of the raw data by the RMA method was successful resulting in the creation of normalised data where the intensities of the probes from different arrays can be compared (Figure 3.3A – 3.3B). The PCA of the normalised data shows four distinct clusters, one cluster in each quadrant of the PCA area (PCA1 = 24.1% and PCA2 = 11.4%). The identified clusters represent replicates from the same fibroblast cell group (Figure 3.3C). The DEGs identification was carried out between the healthy control group and the three conditions: HGPS, HGPS-TERT, and UV-B. The closest group comparison results are Control vs. HGPS and Control vs. HGPS-TERT, followed by the Control vs. UV-B one (Figure 3.3D).

The DEGs analysis results in the identification of 1500 specific DEGs for HGPS; 624 specific DEGs for HGPS-TERT disorder treatment; 1413 specific DEGs for the UV-B treatment. Furthermore, the three conditions share several DEGs: HGPS and HGPS-TERT have 1226 DEGs in common; HGPS and UV-B share 341 DEGs; HGPS-TERT and UV-B have in common 166 DEGs; and, the three comparisons share 417 DEGs (Figure 3.4).

Figure 3.3 CEL files analysis result summary. (A) Raw Data Log intensities. (B) RMA Data Log intensities. (C) PCA plot. (D) Cluster dendrogram between the list of DEGs between the 3 group comparisons.

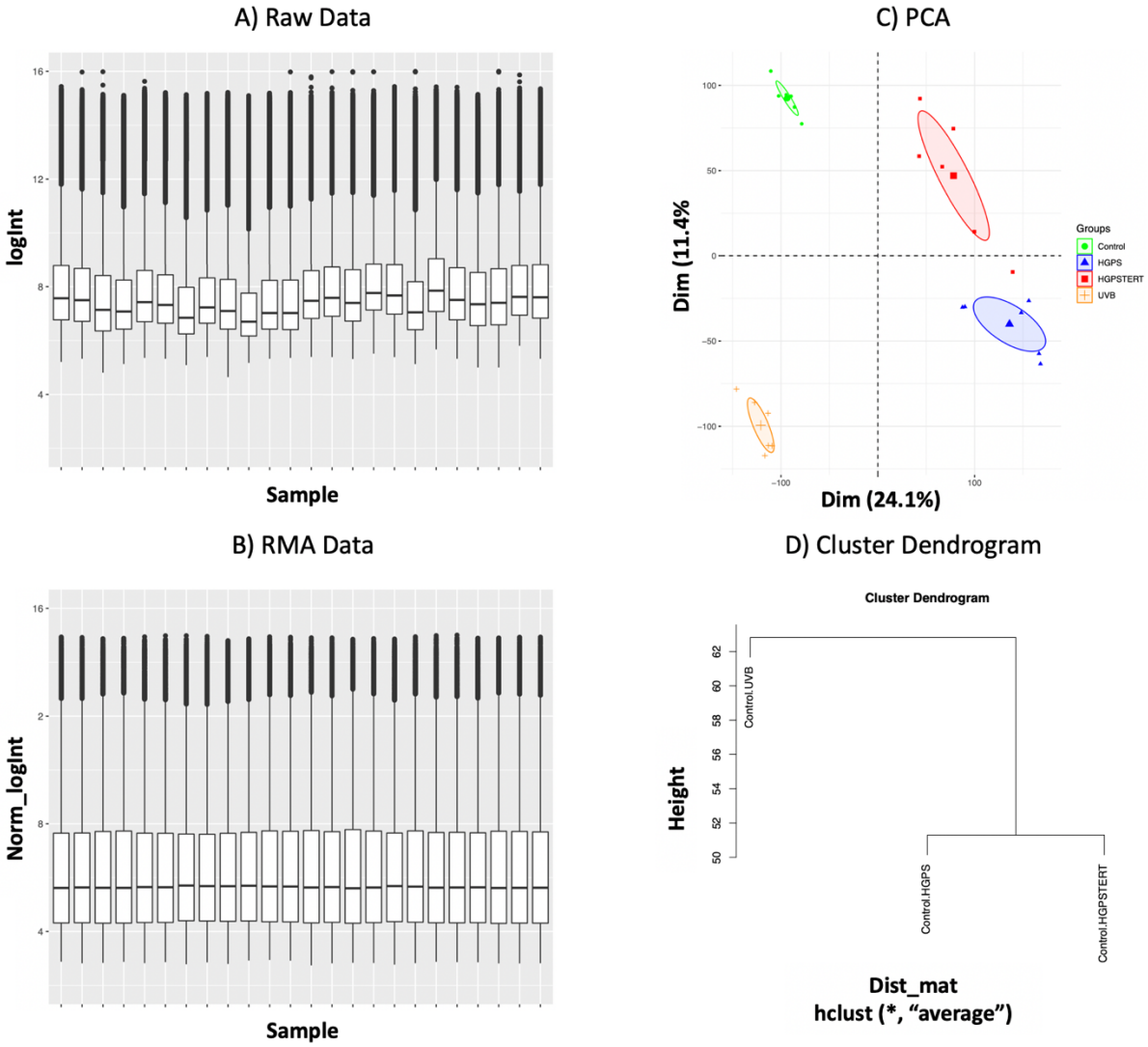
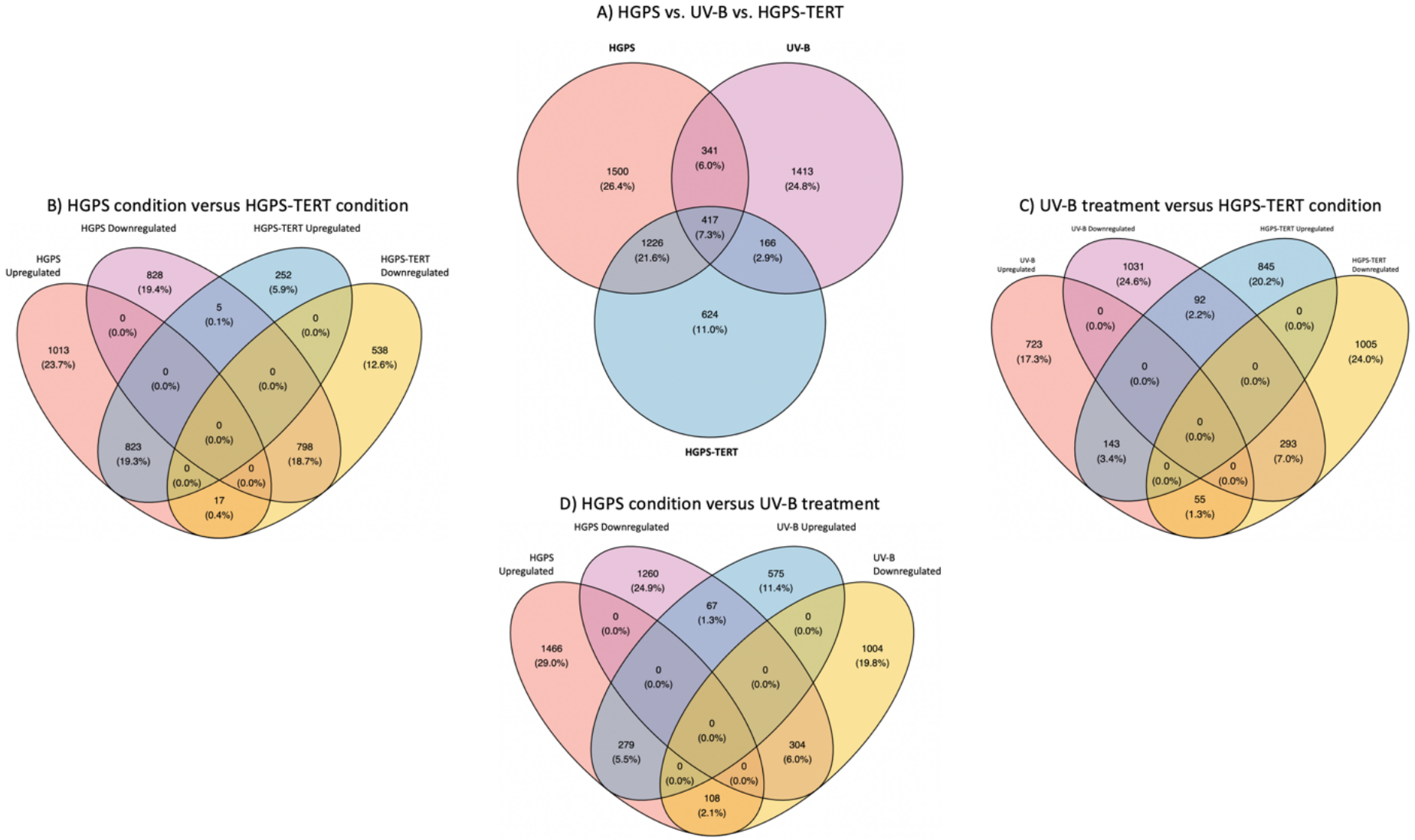


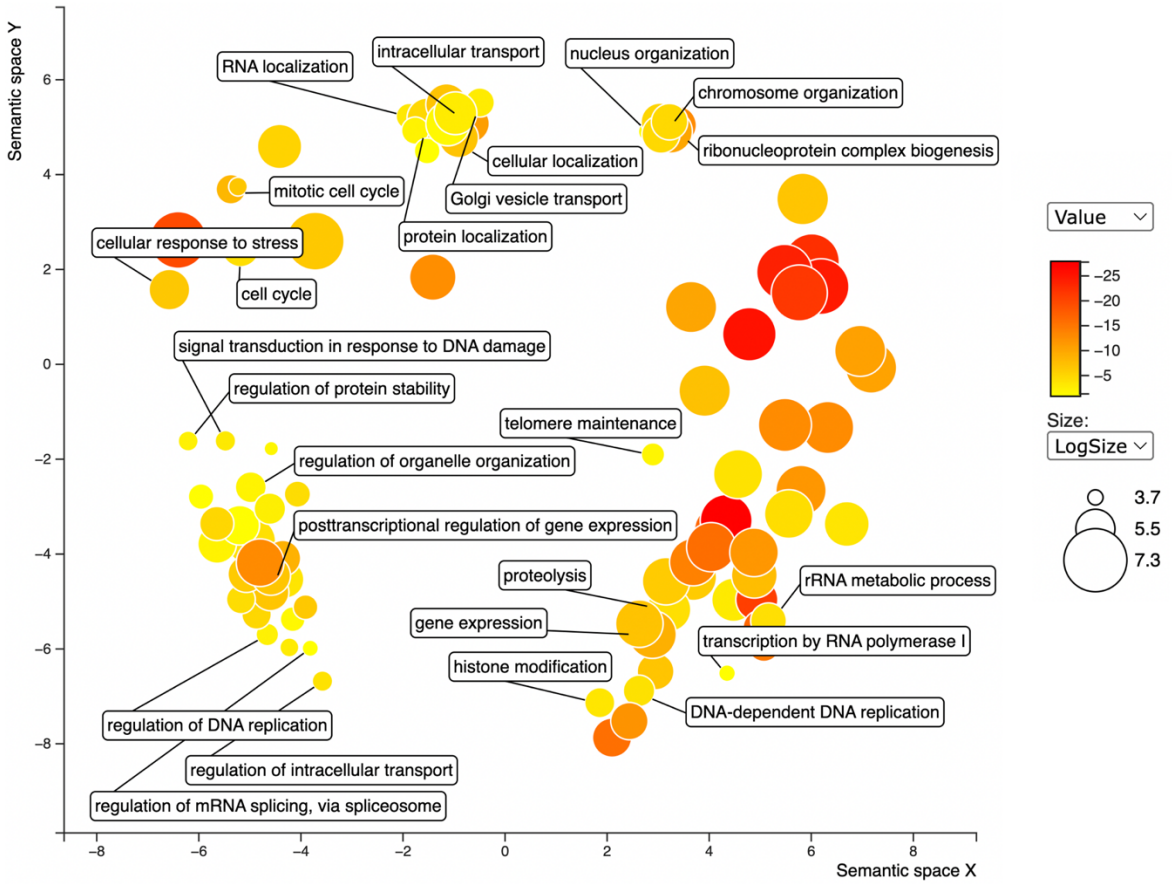
Figure 3.4 DEGs identification between the three group comparisons. (A) HGPS condition versus UV-B treatment. (B) HGPS condition versus HGPS-TERT condition. (C) UV-B treatment versus HGPS-TERT condition. (D) HGPS condition versus UV-B treatment.



3.4.2. HGPS transcriptomics signature on human fibroblast cells

The enrichment analysis of the identified DEGs between the Control and the HGPS group samples resulted in the identification of 208 enriched terms for the Gene Ontology: Biological Process (GO:BP) class. The terms were grouped based on their function to gain a better understanding of the effect of the progerin on human cells (Figure 3.5). The cell-cycle process is one of the main functions that is shown in the analysis. The identified terms are related to the regulation of G1/S and G2/M cell cycle phase transitions; the cellular response to stress caused by DNA damage and its signalling; and DNA replication processes. Furthermore, the progerin condition affects different stages of the transcription-translation phases. The analysis identified the RNA polymerase I which is the enzyme responsible for the transcription of ribosomal RNA (except for 5S rRNA, which is synthesized by RNA polymerase III). Moreover, the results show several terms related to rRNA processing and ribosomal small subunit biogenesis; and, regulation of the transcription for mRNA and ncRNA classes. The results showed several terms related to RNA post-transcriptional regulation and RNA stability. The RNA splicing event was identified with different mechanisms such as via spliceosome, via transesterification reactions, and via transesterification reactions with bulged adenosine as nucleophile. The protein-level regulation is directed towards protein stability and its proteolysis through the proteasome-mediated ubiquitin-dependent protein catabolic process. The progerin also shows regulatory functions toward organelle organization, cellular component assembly, and nucleus organization. The main target of the regulation of organelle is directed towards the cellular localisation of the organelles; the intracellular transport (for both RNA and protein) at the nucleus, cytoplasmic, and nucleocytoplasmic levels. In particular, the endoplasmic reticulum (ER) and Golgi vesicle transport were identified. The nucleus organisation is highlighted with terms like chromosome organisation, telomere maintenance, and histone modification

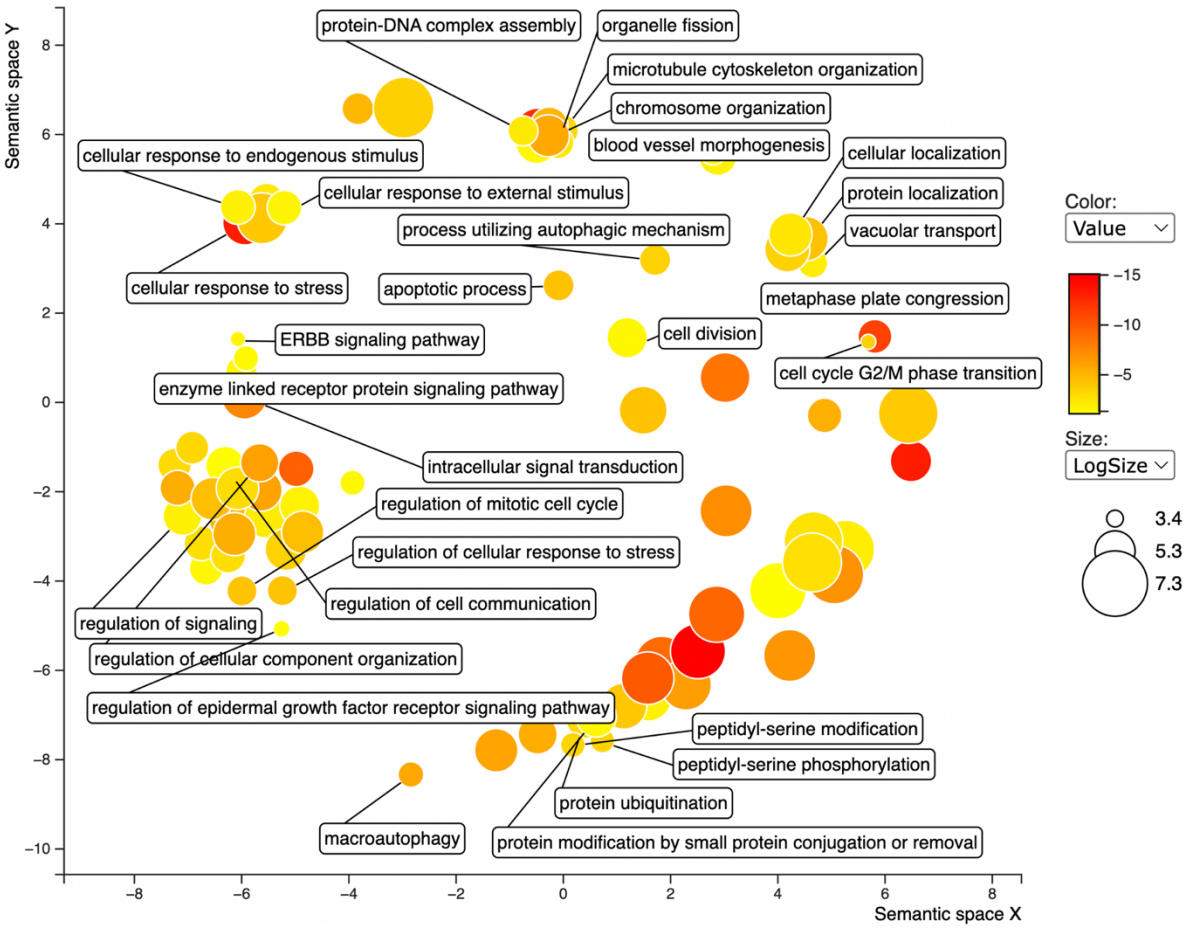
Figure 3.5 GO:BP enrichment analysis summary for the group comparison between the Healthy Control and HGPS. The plot was created with the website version of REVIGO. The semantic space axes in the plot have no intrinsic meaning. The Revigo tool uses the Multidimensional Scaling (MDS) methodology to reduce the dimensionality of a matrix of the GO terms pairwise sematic similarities. GO terms that are semantically similar should cluster closer in the plot. The highlighted GO terms were selected as the most significant ones in the clusters or the one with the most specific biological information. The bubble color shows the user-provided p-value and the size indicates the frequency of the GO term in the GOA database (the bubbles of more general terms are larger) (Barrell et al. 2009).



3.4.3. UV-B light treatment impact on healthy human fibroblast cells

The enrichment analysis of the identified DEGs between the Control group and the UV-B treatment found 169 enriched terms from the GO:BP class. The terms were grouped together based on their biological function to get a better insight into the effect of UV-B light treatment on healthy human cells (Figure 3.6). The nuclear and cell division process is one of the focal functions that is shown in the analysis. The identified terms are related to the regulation of G1/S and G2/M cell cycle phase transition; metaphase plate congression, mitotic spindle organization, microtubule cytoskeleton organization involved in mitosis, and mitotic sister chromatid segregation. Furthermore, the analysis shows the cellular response to DNA damage and nucleosome assembly. The light treatment shows cell response from both external and endogenous stimulus. The identified signalling terms are related to intracellular signal transduction, enzyme-linked receptor protein signaling pathways such as the ERBB signaling pathway, and regulation of the epidermal growth factor receptor signaling pathway. The identified translational regulations are directed at protein localization, protein ubiquitination, protein modification by small protein conjugation or removal, and peptidyl-serine phosphorylation. The UV-B condition affects the regulation of several cellular components organization like microtubule cytoskeleton, chromosome, and organelle fission; and, the cellular localization for vacuoles during their transport. The chromosome organization is highlighted by terms such as protein-DNA complex assembly and nucleosome assembly. Furthermore, the analysis shows several terms related to cell death like apoptotic process, intrinsic apoptotic signaling pathway, and programmed cell death; and, miscellaneous processes like autophagy and macroautophagy; regulation of cell communication; and, blood vessel morphogenesis.

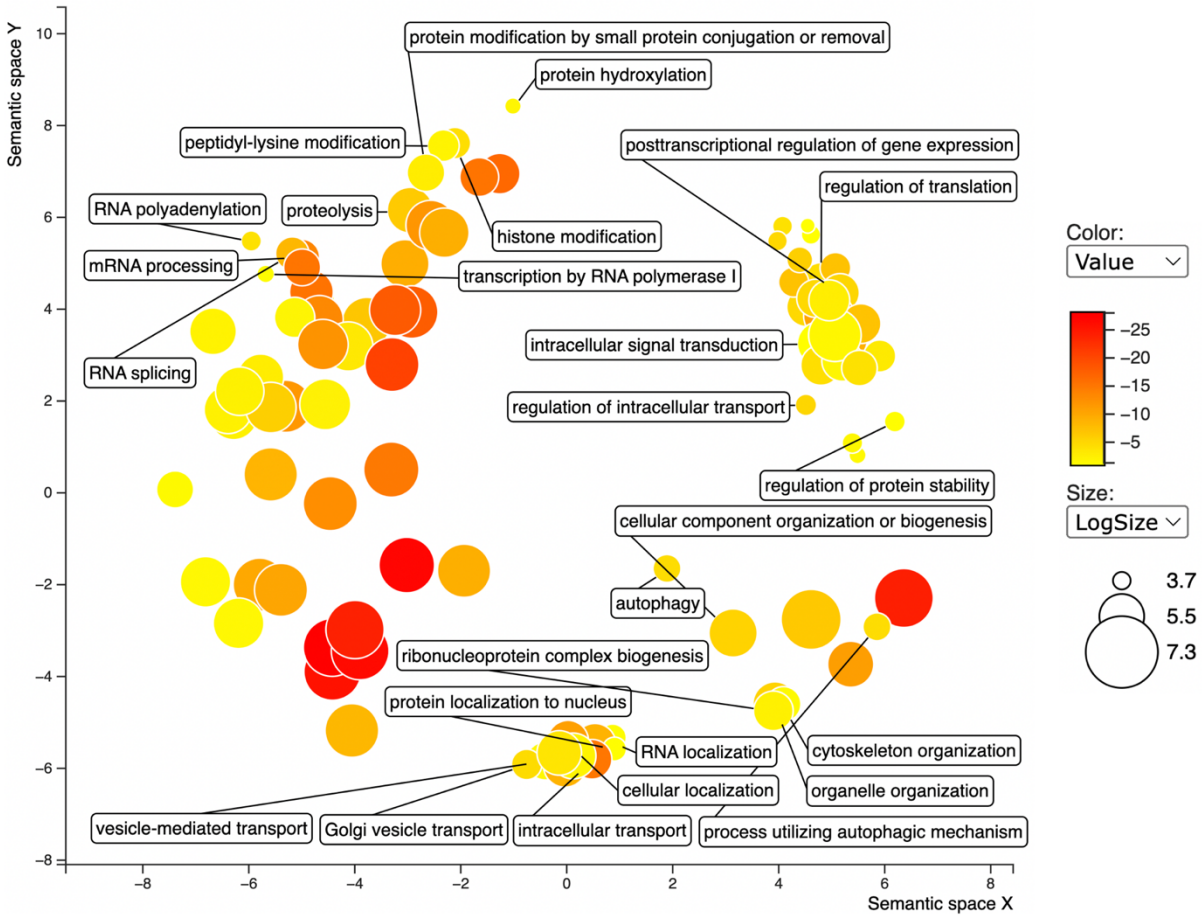
Figure 3.6 Gene Ontology: Biological Process summary for the Healthy Control versus UV-B treated cells. The plot was created with the website version of REVIGO. The semantic space axes in the plot have no intrinsic meaning. The Revigo tool uses the Multidimensional Scaling (MDS) methodology to reduce the dimensionality of a matrix of the GO terms pairwise sematic similarities. GO terms that are semantically similar should cluster closer in the plot. The highlighted GO terms were selected as the most significant ones in the clusters or the one with the most specific biological information. The bubble color shows the user-provided p-value and the size indicates the frequency of the GO term in the GOA database (the bubbles of more general terms are larger) (Barrell et al. 2009).



3.4.4. Effect of telomere elongation on HGPS fibroblast cells

The telomere elongation treatment on the HGPS cell line resulted in the identification of 187 enriched terms from the GO:BP class (Figure 3.7). The HGPS-TERT condition affects different stages of the transcription-translation steps. The transcription regulation is affected at the mRNA and rRNA macromolecules synthesis level. In particular, the analysis identified the main rRNA gene transcription polymerase, RNA polymerase I. Several posttranscriptional regulations of gene expression involve RNA polyadenylation and mRNA 3'-end processing. The splicing event was identified by three different methods: via spliceosome, via transesterification reactions, and via transesterification reactions with bulged adenosine as nucleophile. On the other hand, regulation at the translational level is regulated at protein hydroxylation; protein modification by small protein conjugation or removal; peptidyl-lysine modification; and, proteolysis. Furthermore, the telomere elongation condition affects the overall organization of the cellular organelles and organelle biogenesis. Core affected cellular organizations are the cytoskeleton; the transport mechanisms related to ER-Golgi vesicle-mediated transport, nucleocytoplasmic transport for both RNA and protein, and nuclear transport; and, the cellular localization of protein and RNA related to organelle and the nucleus. Moreover, the analysis also found biological functions related to the process utilizing autophagic mechanism, intracellular signal transduction, ribonucleoprotein complex biogenesis, and histone modification.

Figure 3.7 Gene Ontology: Biological Process summary for the Healthy Control versus HGPS fibroblast cells treated for telomere elongation. The plot was created with the website version of REVIGO. The semantic space axes in the plot have no intrinsic meaning. The Revigo tool uses the Multidimensional Scaling (MDS) methodology to reduce the dimensionality of a matrix of the GO terms pairwise sematic similarities. GO terms that are semantically similar should cluster closer in the plot. The highlighted GO terms were selected as the most significant ones in the clusters or the one with the most specific biological information. The bubble color shows the user-provided p-value and the size indicates the frequency of the GO term in the GOA database (the bubbles of more general terms are larger) (Barrell et al. 2009).



3.5. Discussions

3.5.1. Hallmark of HGPS

HGPS is known to cause cellular senescence in the affected cells and the expression of the progerin protein affects a wide range of cellular processes. The present study displays known affected cellular mechanisms and unreported ones (Figure 3.8).

Genome stability is a complex cellular phenomenon that involves a wide range of biological processes. The present study highlights several biological processes such as chromosome organization, telomere maintenance, histone modification, and cellular response to DNA damage stimulus. As shown by Musich and Zou, genomic instability and DNA damage are one of the hallmarks of HGPS (Musich and Zou 2009). The telomere maintenance in HGPS patients is shown to be shorter than their age-matched in the healthy population (Decker et al. 2009). Furthermore, specific epigenetic modifications are found to be enriched in HGPS patients in the lamina-associated domains when compared to the healthy population (Köhler et al. 2020).

Another trademark of HGPS is cell-cycle alteration. In eukaryotic cells, the cell cycle consists of four stages. The G1 stage depicts the period when the cell is metabolically active to prepare all the macromolecules necessary for DNA replication and it continuously grows. The second stage is called the S phase, during which the DNA replication starts. Followed by the G2 phase, during which the cell continues to grow and the cell synthesize several macromolecules needed for division. The final stage is called the M (mitosis) phase during which the duplicated chromosomes separate into two daughter nuclei and the cytoplasm is divided into two daughter cells, each with a full copy of DNA. Our analysis pinpoint the G1/S and the G2/M transitions as the phase that is impacted by HGPS. Such behaviour is highlighted by Dechat et al. (Dechat et al. 2007).

Phan et al. showed an elevated protein production in HGPS cells correlated with an increased number of cell cycles when compared to the normal cells (Phan, Khalid, and Iben 2019). In our study, HGPS shows an alteration of the transcription exhibited by the RNA polymerase I which is responsible for the expression of the ribosome RNA genes. The HGPS effect on the transcription phase is further expanded during the RNA splicing event with three different methods: via spliceosome, via transesterification reactions, and via transesterification reactions with bulged adenosine as nucleophile. Moreover, the effect of the syndrome is also shown at the protein stability level through the regulation of the translation process and proteasome-mediated proteolysis.

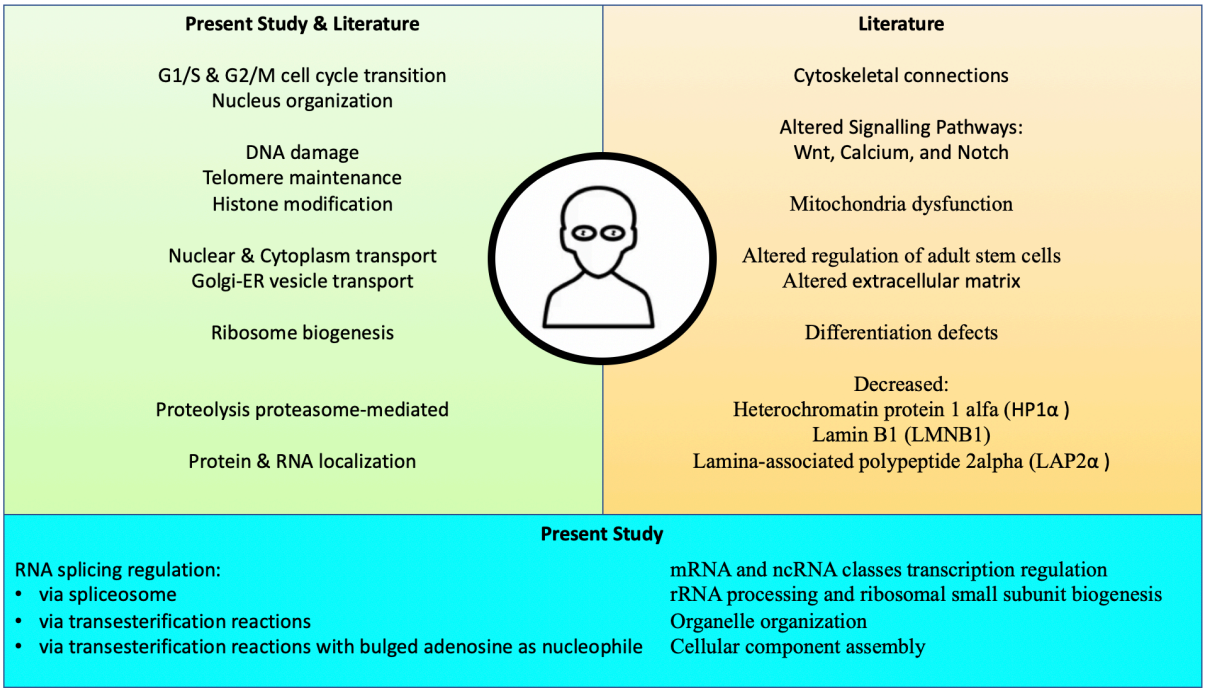
The effect of progerin expression is also detected at the organelle level such as the ER-Golgi. As proved by Chen et al., HGPS cells show the Golgi cisternae to be dispersed as opposed to the normal ones where the organelle is compact and confined to one side of the cell (Z. J. Chen et al. 2014). Our analysis does shed light on vesicle transport as a possible mechanism that is altered in HGPS cells.

The nuclear lamina is known to be an important factor in the transport of macromolecules between the nucleus and the cytoplasm. Ferri et al. demonstrated the decreased ability of the progerin-constituted lamina complex for macromolecule transport towards and from the nucleus; and, a compromised binding ability of transport protein mediators (Ferri, Storti, and Bizzarri 2017). The present analysis shows different enriched terms that show the protein and the RNA intracellular transport as compromised, as well as their intracellular localization.

A literature review of the HGPS effect on human cells shows biological processes that were not replicated in the present study. The disorder is known to disrupt the nucleo-cytoskeleton connections resulting in an impairment of nuclear movement and centrosome orientation in fibroblast polarization for migration (Chang et al. 2019). Additionally, multiple signalling pathways result in alterations in HGPS cells. Fafián-Labora et al. observed a statistically higher concentration of intracellular calcium in HGPS cell lines compared to healthy ones. The authors demonstrated the relationship between the calcium signalling pathways and mitochondria-associated membranes, apoptosis, and mitochondrial ROS production (Fafián-labora et al. 2021). The mitochondria disruption was found to be time- and dose-dependent caused by a downregulation of mitochondrial oxidative phosphorylation proteins (Rivera-Torres et al. 2013). Moreover, Hernandez et al. demonstrated the inability of the HGPS cells in murine models to produce a viable ECM which is associated with an altered Wnt signalling pathway (Hernandez et al. 2010). The Notch signalling pathway is also affected by HGPS and it regulates stem cell differentiation during osteogenesis and adipogeneses (Scaffidi and Misteli 2008). In addition, HGPS cell lines show a downregulated expression of lamin B1 (LMNB1), heterochromatin protein 1 α (HP1α), and LAP2α, and loss of nucleoplasmic lamins (Vidak and Foisner 2016).

The present study highlights biological processes that are not present in the literature such as the regulation of RNA splicing classes; the regulation of mRNA and ncRNA transcription molecules; the processing and biogenesis of rRNA molecules; the organization of organelles; and, the assembly of cellular component (Figure 3.8).

Figure 3.8 Hallmark of HGPS. The left green section shows the pathways that were identified in the present study and already found in the literature. The right orange section shows the biological pathways that are known from the literature to affect HGPS but were not identified in the present study.



3.5.2. Senescence analysis

Cellular senescence is expressed as the state of the cell-cycle arrest prompted by aged or damaged cells caused by oncogenic signalling, DNA damage, and telomere loss (Muñoz-Espín et al. 2013). Moreover, senescence is employed by the cells in normal development and it is necessary for tissue homeostasis (McHugh and Gil 2018). The three conditions in the present study do force the cells to enter senescence. The results show the HGPS and the HGPS-TERT as the closer groups and followed by the UV-B treatment (Figure 3.3D). Such behaviour is also proved by the shared number of DEGs between the groups (Figure 3.4).

3.5.2.1. Senescence comparison between HGPS and HGPS-TERT

Telomere shortening is both a hallmark of cellular ageing and HGPS. The hypothesis that telomere elongation in HGPS cells can restore some of the ageing symptoms was checked in the present study. The literature review shows ambivalent results following telomere elongation in human HGPS cells. Bikkul et al. show that telomere elongation results in chromosome mislocalization for both normal and HGPS cells (Bikkul et al. 2019). Whereas, transient expression of human telomerase in combination with farnesyltransferase inhibitor (FTI) lonafarnib could represent an improved novel therapeutic approach for HGPS (Yanhui Li et al. 2019).

The present study shows that telomere elongation results in reversing the gene expression of 5 DEGs that are downregulated in HGPS and upregulated in HGPS-TERT; and, 17 DEGs that are upregulated in HGPS and downregulated in HGPS-TERT. The remaining 1621 shared DEGs do not change expression patterns following telomere elongation (Figure 3.4B). Furthermore, 790 DEGs were detected for the HGPS-TERT condition that are normally expressed in HGPS. The telomere elongation treatment shows exclusive enriched terms such as apoptotic intrinsic pathway signalling; microtubule cytoskeleton involved during the mitosis spindle; posttranslational modification like peptidyl-serine and peptidyl-lysine phosphorylation; and, autophagy. Whereas, the HGPS presents the following terms that are absent in HGPS-TERT such as signal transduction in response to DNA damage; chromosome organization for telomere maintenance; assembly of protein-DNA nucleosome; cell cycle transition between G1/S and G2/M. The results suggest that telomere elongation resolves cellular deficiencies related to the stability of the nucleus and of the chromosome. On the other hand, it introduces new abnormalities concerning microtubule-cytoskeleton organization during mitosis, apoptotic pathways, and cellular organization of autophagy. The cell-cycle process continues to be affected in both conditions but in different phases. HGPS is affected during the G1/S transition potentially due to the signalling of DNA damage; and, G2/M transition due to abnormal chromosome organization. On the other hand, HGPS-TERT negatively affects the cell cycle due to the microtubule-cytoskeleton interaction involved during the mitosis spindle.

3.5.2.2. Senescence comparison between HGPS and UV-B

The UV-B treatment triggers cell senescence with some shared biological processes with HGPS and HGPS-TERT but it also shows a distinct biological imprint that is exclusive to the light treatment (Figure 3.9). The UV-B group shows a distinct gene expression as shown by the PCA (Figure 3.3C); and, shares 924 DEGs (with both HGPS and HGPS-TERT) which represents 39.5% of its total number of DEGs. The UV-B condition expresses 16 specific GO:BP clusters related to chromosome segregation; cell death; metaphase plate congression; cellular response to a stimulus; negative regulation of intracellular signal transduction; fission organelle; vacuolar transport; vasculature development; blood morphogenesis vessel; cell division; and, ERBB signalling pathway. The present study shows a different pattern of cell senescence caused by the two progeria conditions and the UV-B light treatment proving that cellular senescence can be triggered by different mechanisms at different cellular components.

Figure 3.9 Senescence comparison analysis between HGPS, HGPS-TERT, and UV-B treatment.

3.6. Conclusion

HGPS is a complex human genetic disease involving different cellular mechanisms ranging from telomere maintenance to intracellular transport. The scientific hypothesis that telomere elongation treatment can resolve some of HGPS symptoms was checked and the analysis resulted in the emergence of new dysregulated pathways that are absent in the original condition. Furthermore, the UV-B treatment highlights a cellular senescence state that differs from the HGPS process.

Chapter 4 - Drug repurposing from gene and expression data: A survey of bioinformatics tools and databases.

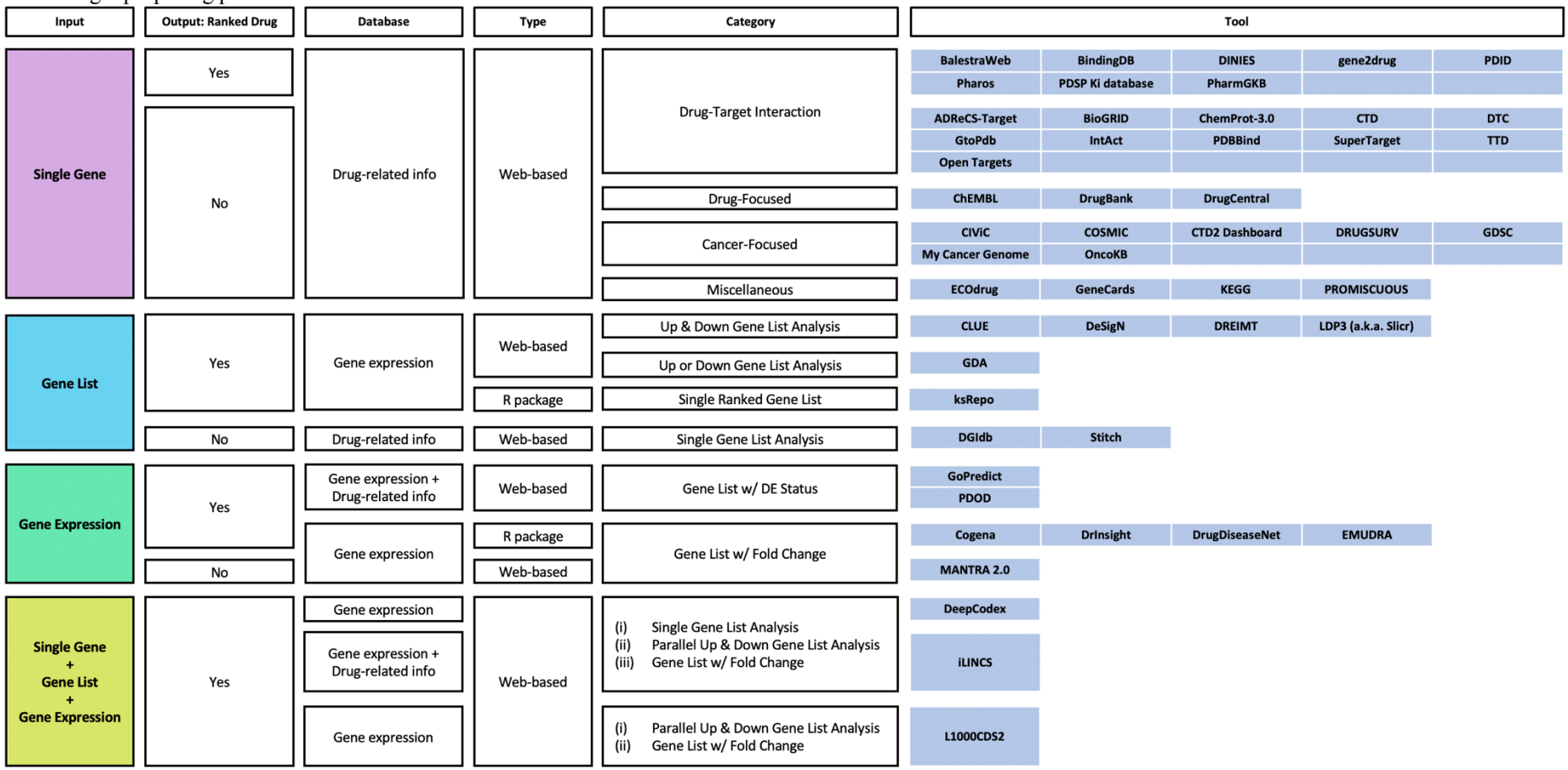
4.1. Introduction

Drug repurposing (DR) entails methods to identify new therapeutic uses for existing drugs for the treatment of diseases different than their original purposes (Sam and Athri 2019). It is also known by different names such as drug repositioning, drug redirection, drug recycling, drug re-tasking, drug reprofiling, drug rescuing, and therapeutic switching (Pushpakom et al. 2018; Jarada, Rokne, and Alhajj 2020). DR methodologies are complementary to traditional approaches for drug discovery due to their ability to select drugs with therapeutic potential in a reasonable amount of time and cost and with potentially lower risk of side effects (Sam and Athri 2019; Serçinoğlu and Sarica 2019). Choosing a particular computational drug repurposing workflow usually depends on the type of input data available, which includes genomic data (a single gene or a set of genes), gene expression information (fold changes or p-values), the chemical structure of a candidate drug molecule, the phenotype of disease in question, or a combination of these data (Jarada, Rokne, and Alhajj 2020). The present work focus on the identification of tools (standalone, web-based, and R package) capable of using human transcriptomic data to identify potential drug treatment for a diverse range of biological conditions.

4.2. Drug repurposing tools

The identification of drug repurposing tools was achieved through a literature review (Sam and Athri 2019; Serçinoğlu and Sarica 2019; Jarada, Rokne, and Alhajj 2020; Pushpakom et al. 2018; Lotfi Shahreza et al. 2018; Musa et al. 2018) (total of) and PubMed searches as of March 2021 using the following query terms: “drug repurposing” (2450 results), “drug repositioning” (4101 results), “drug redirection” (4 results), “drug recycling” (9 results), “drug re-tasking” (quoted phrase not found in phrase index), “drug reprofiling” (20 results), “drug rescuing” (quoted phrase not found in phrase index), and “therapeutic switching” (33 results). The tools were then filtered based on the type of input data they require: (i) a single gene; (ii) a list of genes; (iii) gene expression data; or a combination of any of these types of data (Figure 4.1).

Figure 4.1 Drug repurposing tool functional list as identified in the literature. The figure shows all the tools found in the literature review that are functional. The tools are divided into the following groups. The "Input" shows all the input options available to the user: the "Single Gene" expresses the usage of a single gene; the "Gene List" expresses the ability to use a regular gene list; the "Gene Expression" express the ability to use genes with their expression data such as Fold Change; the "Gene List + Gene Expression" means the user can use both gene expression and gene list. The "Output: Ranked Drug" shows whether the output drugs are associated with any statistical or measured value. The "Type" shows the system that permits the use of the tool. The "Category" is used to further divide the tools when necessary. The "Tool" shows the name of the drug repurposing platform.



4.2.1. Tools with single genes as input

The platforms in this group accept as input a single gene. The output is a drug list that is known to bind to the encoding protein, according to the sourced databases, or predicted drugs through different methodologies. The tools are all web-based, and they can be further divided based on the main topic elaborated by their sourced database.

Drug-focused resources

The group has the following members: ChEMBL (Mendez et al. 2019), DrugBank (Wishart et al. 2018), and DrugCentral (Ursu et al. 2017). The platforms rely on databases that host drug-related information such as the drug structure, protein target, drug synonyms, drug effects, mechanism of action (MOA), clinical trial, and FDA drug label. A typical gene-input outputs a list of known drug-gene interactions.

Cancer-focused resources

The group has seven members: CIViC (Griffith et al. 2017), COSMIC (Tate et al. 2019), CTD^2 Dashboard (Aksoy et al. 2017), DRUGSURV (Amelio et al. 2014), GDSC (Yang et al. 2012), My Cancer Genome (Jain et al. 2020), and OncoKB (Chakravarty et al. 2017). All the members use databases built primarily on cancer cell line data studies. The goal of these platforms is to collect and make cancer data accessible to researchers for cancer treatment and further help drug repurposing. It is possible to query the platforms with different inputs. A gene-input results in an output of a drug list with additional information based on the database source used by the direct platform.

Drug-target interactions

The platforms are built upon the binding data between small molecules and their protein target; and, known interaction between compounds and diseases. The following members output a drug list based on known drug-target interaction from their database source: ADReCS-Target (Huang et al. 2018), BindingDB (Gilson et al. 2015), BioGrid (Oughtred et al. 2019), CTD (Davis et al. 2021), DTC (Tanoli et al. 2018), GtoPdb (IUPHAR/BPS Guide to PHARMACOLOGY) (Armstrong et al. 2020), IntAct (Orchard et al. 2014), Open Targets (Ochoa et al. 2021), PDBBind (Z. Liu et al. 2017), PDSP Ki (Roth et al. 2000), PharmGKB (Whirl-Carrillo et al. 2012), Pharos (Nguyen et al. 2016), SuperTarget (Hecker et al. 2011), TTD (Y. Wang et al. 2020), and is IntAct (Orchard et al. 2014). Meanwhile, the following tools predict the potential interactions between a compound and the target based on known interaction: BalestraWeb (Cobanoglu et al. 2015), ChemProt (Kringelum et al. 2016),

gene2drug (Napolitano et al. 2018), and PDID (C. Wang et al. 2016). The DINIES tool can output both known interactions and predicted ones (Yamanishi et al. 2014).

Miscellaneous

The members from this category are built upon databases created from different sources. The ECOdrug webpage shows the potential interactions between drugs and their targets across different organisms (Verbruggen et al. 2018). The database covers 640 eukaryotic species, and it integrates data from Ensembl (Yates et al. 2020), EggNOG (Huerta-Cepas et al. 2016), and InParanoid (Sonnhammer and Östlund 2015). Moreover, the tool can predict drug interactions with ortholog genes from different species within an ecosystem. The GeneCards database integrates data from around 150 sources including gene and genome variants; protein-related information; pathways; cells lines; diseases; omics data (genomic, transcriptomic); clinical information; drugs-related information; gene and protein expression; orthologs and paralogs information; and, gene ontologies (Safran et al. 2010). The output from a gene input consists of several tabs, one of them being the drug tab, which gives a list of drugs with their status, mechanism of action, clinical trials, role, and associated diseases. The KEGG is a well-known manually curated resource for information about genomes, pathways, diseases, and drugs (Kanehisa et al. 2021). KEGG provides information about gene-drug interactions through the KEGG DRUG. The last member is the PROMISCUOUS database which includes three types of data entities: drug-target binding interactions, drug-centric databases, and drug-side effect relations (Gallo et al. 2021). It is possible to query the database by the drug and by the gene. The output format for a gene input is a list of drugs, with their pharmacological action.

4.2.2. Tools with a list of genes as input

The group relies on two types of sources: gene expressions profile databases such as the CMap (J. Lamb et al. 2006), the GEO (R. Edgar, Domrachev, and Lash 2002), and the LINCS project (Subramanian et al. 2017); and drug-related information databases, that contain data such as protein targets, side effects, interaction type, sources, and references. The members of this group are further divided into four categories, depending on the type of input: gene list, ranked gene list, lists of up- or down-regulated genes, and simultaneous sets of up- and down-regulated genes. On the other hand, the output data depends on the source database. In the case of gene expression profile sources, the drug list output is obtained through similarities or concordantly expressed genes between the input data and profiles in the database (J. Lamb et al. 2006; Campillos et al. 2008; Pilarczyk et al. 2019), while tools that rely on drug-related information

databases without expression data provide a list of drugs, based on known drug-target interaction information.

Gene lists as inputs

The sub-group has two members: DGIdb and STITCH. The DGIdb platform organizes data from 30 different sources and offers the possibility to access detailed information concerning drugs, genes, drug-gene interactions, and their references in PubMed (Cotto et al. 2018). The platform outputs a list of drugs with additional information from the sourced databases. Meanwhile, STITCH provides a comprehensive map of drug-protein network interactions, together with a diverse range of filters and visualization options (Szklarczyk et al. 2016). STITCH also integrates multiple source databases and provides a confidence score for each reported interaction. The tool accepts as input a chemical name or a structure, a gene name, or a protein sequence, and it outputs a drug-protein network involving the query data. The network also displays all the known binding affinity constant values using the edge width of the drug-protein interaction.

Ranked gene lists as inputs

The sole member of this category is ksRepo (Brown et al. 2016). It compares the ranked gene list input data to a database of signatures or compound-gene interaction lists; and, it identifies drugs with similar expressions to the input.

Up- or down-regulated genes as inputs

The sub-group has one sole member: GDA (Caroli et al. 2018). The tool integrates human cancer cell lines (from the NCI60 panel and the cancer cell line encyclopedia), drug responses, and gene mutation data. It offers four possible usages based on the input: from gene to drug; from drug to gene; from gene signature to drug; and, from drug to gene signature. GDA requires the user to choose between an up- or down-regulated expression. The output displays a table that includes a list of drugs with diverse data such as drug family, MOA, drug score in cancer cell lines, p-value, and links to the related literature.

Up- and down-regulated genes as inputs

The category includes four members that differ primarily by their data sources. The CLUE platform uses the CMap L1000 dataset as its sourced material and the methodology explained by Subramanian et al. to identify the drugs that elicit similar patterns of up- and down-regulation as the input data (Subramanian et al. 2017). DeSigN connects the input data with gene profiles associated with cancer cell line drug response data (B. K. B. Lee et al. 2017). The source databases are from the Genomics of Drug Sensitivity in Cancer Project (Yang et al. 2012) and the Cancer Cell Line Encyclopedia (Ghandi et al. 2019). The output is a list of drugs

with their gene targets associated with a connectivity score and a p-value. The output drugs are identified by the use of the non-parametric Kolmogorov-Smirnov (KS) statistic for a rank-based pattern-matching approach between the query signatures to the reference database. The connectivity score is computed according to Lamb et al. (J. Lamb et al. 2006). The DREIMT tool focuses on the identification of drugs capable of regulating the human immune system from 70 immune cell subtypes, 4,960 drug profiles, and ~2,600 immune gene expression signatures (Troulé et al. 2020). The output is a list of associated drugs with additional information concerning their pharmacological status, MOA, association scores, and drug approval status. The last platform of the sub-group is LDP3 (a.k.a. Slicr) (Zichen Wang et al. 2018). It uses the LINCS database as its source and the Characteristic Direction (CD) method to connect up- and down-regulated genes to drugs (Subramanian et al. 2017). The CD is a multivariate method used to identify differentially expressed genes; and, it tends to give more weight to genes that show a low value of fold change but 'move' together with a larger group of other genes (Duan et al. 2016a; Clark et al. 2014). The output of the platform is a list of drugs and gene knockdowns that mimics or reverse the input.

4.2.3. Tools with gene expression data as input

The cluster includes tools that require as input a gene list associated with their corresponding expression fold-changes, or a numeric value depicting the differentially expressed status, e.g. 1 for upregulated, -1 for downregulated. This category relies upon large-scale public expression databases such as the GEO, the CMap (J. Lamb et al. 2006), and the LINCS (Subramanian et al. 2017). These databases provide gene expression patterns induced by drugs, diseases, or other environmental factors. Drug repurposing tools exploit this knowledge to find similarities between the input data provided by the user, and the gene expression profiles in databases (J. Lamb et al. 2006; Campillos et al. 2008; Pilarczyk et al. 2019). In this manner, it is possible to identify drugs that can mimic (positive correlation) or reverse (negative correlation) the expression patterns from the input data. For the sake of simplicity, we classified these tools into two categories: (i) data including the status of differential expression; or (ii) a table of genes with their respective expression fold changes.

<u>Gene list with differentially expressed status as input</u>

The two tools in this sub-category accept as input a two-column table, with the gene names on the first column, and the corresponding differential expressed status (1 for upregulated, -1 downregulated, and 0 for non-differentially expressed) in the second. The GoPredict tool integrates transcriptomic and epigenomic data primarily from breast and ovarian cancer, and

signaling pathway information (Louhimo et al. 2016). The output consists of a table with a ranked list of drugs, entailing drug names, inhibition score, and a pro-cancer effect score (penalty term). The PDOD tool identifies drugs with the potential to reverse the expression of the input data (H. Yu et al. 2016).

Gene list with fold change data as input

This category has six members divided into two web platforms: CDA and MANTRA; and, four R packages: Cogena, DrInsight, DrugDiseaseNet, and EMUDRA.

The CDA web tool performs expression pattern matching between signaling pathways components and drug-induced expression patterns (J. H. Lee et al. 2012). The CDA first identifies the signaling pathways through gene set enrichment analysis on the input gene expression data and then finds drugs affecting these pathways. The output is a table of single and combinatorial drugs, that are ranked according to the number of affected pathways. The MANTRA web tool exploits the CMap database to identify similar gene expression profiles between multiple drugs through gene set enrichment analysis (Carrella et al. 2014) (Subramanian et al. 2005b). MANTRA also allows the user to upload their own drug-induced gene expression profiles from microarray analyses. The output is a network where the nodes represent the drugs and the edges stand for the distances between the compounds, calculated through GSEA-based similarity between drugs.

The Cogena R package uses the CMap dataset to identify potential therapeutic drugs (Jia et al. 2016). First, Cogena finds co-expressed gene sets in the input data through various clustering methods and then carries out pathway enrichment analysis for each co-expressed gene cluster using the hypergeometric test. The drug repurposing analysis is achieved by selecting the top 100 from both up- and down-regulated genes for each drug-induced gene expression data from the CMap; and, identifying the gene sets they affect. A hypergeometric test is conducted to find a possible relationship between the co-expressed gene clusters and the CMap gene sets. Finally, the identified drugs are ordered based on the overlap between the respective gene sets. DrInsight uses the CMap database as the source for drug perturbations (Chan et al. 2019). DrInsight measures the concordance between the input data and the CMap drug profile data for each gene. For this, DrInsight searches for gene sets with similar expression patterns to the input data (concordantly expressed genes), and this circumvents the necessity for a subjective selection of the query signatures, such as using fold changes or p-values. The output consists of a list of drugs, ranked by their statistical significance (p-values and false discovery rates). The DrugDiseaseNet tool uses several publicly available drug-perturbation databases (CMap, GEO, LINCS) together with databases linking genes to diseases from the Lung Genomics

Research Consortium to create drug-disease networks (Peyvandipour et al. 2018). The system provides a list of ranked drugs including their therapeutic repurposing scores. These scores are calculated from the correlation coefficients between their gene perturbation signature, and range from –1 (opposite perturbation signatures) to 1 (similar perturbations, i.e. the drug may cause the same effects). The last tool is called EMUDRA and it integrates four different methodologies (X. Zhou et al. 2018). The expression-weighted cosine method reduces the influence of the uninformative expression changes caused by lowly expressed genes (X. Zhou et al. 2018). The nonparametric KS statistic identifies relationships between gene sets and drug-induced gene expression profiles (J. Lamb et al. 2006). The weighted signed statistic, used in the sscMap tool, separates up- and down-regulated genes in disease signature and drug-induced signature and then calculates the normalized score for rankings of disease and drug signatures (S.-D. Zhang and Gant 2009). Last, the eXtreme method consists of four statistics that measure the correlation scores between disease signature and the top- and down-regulated genes in drug-induced treatment by using the sum, cosine similarity, Pearson correlation, and Spearman correlation measures (Cheng and Yang 2013). EMUDRA uses the CMap and the LINCS databases as sources, and the output consists of a list of drugs ranked by a score calculated through the integration of the four methodologies.

4.2.4. Tools with single gene, gene list, or gene expression as input

The group has three platforms: DeepCodex, iLINCS, and L1000CDS2. The tools rely on both gene expression profile databases and drug-related databases. The output is a list of drugs where a positive correlation means that the drugs are potentially able to mimic the input gene signature; and, a negative one indicates that the drugs can revert the expression patterns from the input data.

The DeepCodex website uses the drug-induced gene signature profiles from the LINCS data set. The tool exploits the advantages of deep neural networks to create an embedding that substantially denoises the expression data, making replicates of the same compound more similar and accurately predicting pharmacological similarities between drugs (Donner, Kazmierczak, and Fortney 2018). The iLINCS web platform integrates drug-induced gene expression datasets (e.g. LINCS) with different omics data resources. It performs pathway analysis, functional network analysis, drug repurposing analysis, drug-induced gene expression, and upstream regulatory network analysis (Pilarczyk et al. 2022). In conclusion, the L1000CDS2 website focuses on the integration of the CMap database with L1000FWD, a web application that provides interactive visualization of over 16,000 drug and small-molecule-

induced gene expression signatures (Duan et al. 2016a). It provides either the fifty most positively (mimic) or negatively (reverse) correlated drugs to the input data, according to the input configuration choices by the user.

4.3. Conclusion

The present literature review summarizes the vast diversity of tools used for drug repurposing analysis that use transcriptomics data as input. The heterogeneity of these platforms reveals the profound changes in drug development approach: from single gene targets to network-based approaches. Nevertheless, improvement can still be achieved by merging all the different public databases with drug-related information; increasing the coverage of the CMap (J. Lamb et al. 2006) and the LINCS project (Subramanian et al. 2017) by adding new drug perturbations and new human cell lines; and, a platform that integrates all the different statistical approaches would likely increase the likelihood of identifying relevant therapeutic drugs.

Chapter 5 - Human epithelial single-infection with Influenza A virus and Streptococcus pyogenes

5.1. Introduction

Influenza A virus (IAV) and Streptococcus pyogenes (group A streptococci, GAS), are major human healthcare issues worldwide (R. A. Lamb 2008; Reglinski and Sriskandan 2015). IAV is the sole member of the genus Alphainfluenzavirus of the viral Orthomyxoviridae family. The members of the family present a segmented, negative-sense, and single-strand RNA segments. IAV strains are known to cause influenza in birds and some mammals such as pigs, horses, and humans. The virus causes the flu, a contagious respiratory disease, that infects the nose, throat, and lungs. Normal clinical symptoms of the flu are a rapid rise in temperature, limp ache, tiredness, general faintness, headache, and dry cough. The high incidence of the virus is characterized by seasonal epidemics in both hemispheres each year and the occurrence (recurrence) of IAV subtype pandemics throughout the years. Seasonal epidemics result in approximately 500,000 human death per year worldwide (Fauci 2006). However, influenza pandemics are more severe as shown by the Spanish Flu of 1918 (subtype H1N1) with 40 million deaths; the Asian influenza of 1957 (H2N2) with 1-2 million deaths; and, the Hong Kong Flu of 1968 (H2N3) with 0.75-1 million deaths (R. A. Lamb 2008). Meanwhile, GAS is a species of Gram-positive, aerotolerant, and mostly extracellular bacterial pathogen from the genus Streptococcus (Reglinski and Sriskandan 2015). GAS is categorized according to the Lancefield classification system which is a serotype grouping based on the presence of the emm protein (also known as M proteins), on the bacterial cell wall. The emm proteins coat the GAS species and are essential components for bacterial virulence, necessary for the bacterial evasion of the antiphagocytic functions of the host (Reglinski and Sriskandan 2015). GAS species are known members of the commensal human microbiota present in the skin and upper respiratory tract. Nonetheless, the species is an important global human pathogen due to its ability to cause a wide spectrum of clinical infections (pharyngitis, acute rheumatic fever, scarlet fever); and, the diverse strategies used by the bacterial to elude the host defence mechanisms (Reglinski and Sriskandan 2015).

The impact of the two pathogens is observed not only at the single infection level but also during IAV-GAS co-infections where the increased mortality rate is associated with a secondary GAS presence (Zakikhany et al. 2011; Okamoto et al. 2003); and, an exacerbation of the clinical symptoms for patients infected by both pathogens (Chaussee et al. 2011). To gain insight into the interaction between IAV and GAS during an IAV-GAS co-infection, we conducted a transcriptomics analysis of both host-pathogen during the human mono-infections by IAV and two GAS strains: AP1 serotype M1 (GAS M1-AP1) and serotype 591 (GAS M49-591). The human epithelial pharynx cell line (Detroit 562) was chosen as the host cells. The epithelial pharynx cells were selected due to their role as the first line of defence against infection and due to it being a tissue target for both pathogens (Günther and Seyfert 2018; R. A. Lamb 2008; Reglinski and Sriskandan 2015). The IAV subtype H1N1 was chosen due to its endemic nature towards humans and its causative nature during different human pandemics like the Spanish flu of 1918 (R. A. Lamb 2008). The GAS species was selected over other streptococcal species due to the lack of a vaccine therapy clinically available against GAS. Thus, IAV vaccination is the sole method of preventing these IAV-GAS co-infection (Steer et al. 2016). The GAS M1 serotype was chosen due to its association with almost all clinical manifestations of GAS infection (Metzgar and Zampolli 2011) and its association with severe streptococcal infections (D. R. Martin and Single 1993; Musser et al. 1993); whereas, the GAS M49 serotype roughly accounts for 50% of GAS isolates worldwide (Bessen and Lizano 2010).

5.2. Aims

In the present study, we exploited the advantages of transcriptomics data to study the infection processes between the human host and three pathogens in a single-infection approach, as elucidated in a typical dual RNA-Seq pipeline. The identification of host DEGs (control group versus infected group) in each infection will be used to achieve the following goals. First, to identify the effect each pathogen has on the human host. Second, to gain insight onto the differences between the two GAS serotypes, M1 and M49. Third, to identify the biological mechanisms used by IAV to promote secondary GAS infection. Fourth, the usage of host DEGs to find potential drugs that can aid the defence response of the human host cells against an IAV-GAS co-infection. The search for new potential drugs will be conducted through the DR analysis. The DR analysis is the process of finding new clinical uses for approved and/or investigational drugs that are outside its original research purpose (Pushpakom et al. 2018). The new strategy provides several advantages over traditional methodologies. First, the risk of

failure is lower due to its usage of already pre- or clinically tested drugs. Second, the time frame for drug development is virtually reduced to none thanks to the accumulation of knowledge regarding its preclinical testing and process of development. Third, the cost is minimal depending on the amount of knowledge and testing done on the specific drug. There are several approaches used for DR. This study uses the signature matching methodology where the signature (e.g. gene expression) of an event is compared against the collection of drug-induced signatures.

5.3. Methods and Materials

5.3.1. Materials

5.3.1.1. List of cells

The study involved the use of Detroit 562 cells as human host cells. The host cell line is isolated from the pharynx of a female, Caucasian, pharyngeal cancer patient. The chosen GAS serotypes are GAS M1 (strain AP1) and GAS M49 (strain 591). Meanwhile, the IAV strain is A/Bavaria/74/2009 (H1N1). The laboratory experimental work was carried by our collaborative partners of the Institute for Medical Microbiology, Virology, and Hygiene at the University of Rostock under Professor Bernd Kreikemeyer.

5.3.1.2. Infection of Detroit 562 cells with GAS

The infection was conducted during an overnight culture of GAS M1-AP1 or GAS M49-591. Human cells and bacteria were incubated for 2 h at 37 °C, 5 % CO_2 with a multiplicity of infection of 100. The cells were collected followed by library preparation (not shown).

5.3.1.3. Infection of Detroit 562 cells with IAV

Detroit 562 cells were seeded into 24-well plates and grown to confluence. Then, the IAV cells were added. Human cells and viruses were co-incubated for 48 h at 37 °C, 5 % CO_2. After incubation, supernatants, and pellets were collected and followed by the library preparation (not shown).

5.3.2. Methods

5.3.2.1. Transcriptomics Data Analysis

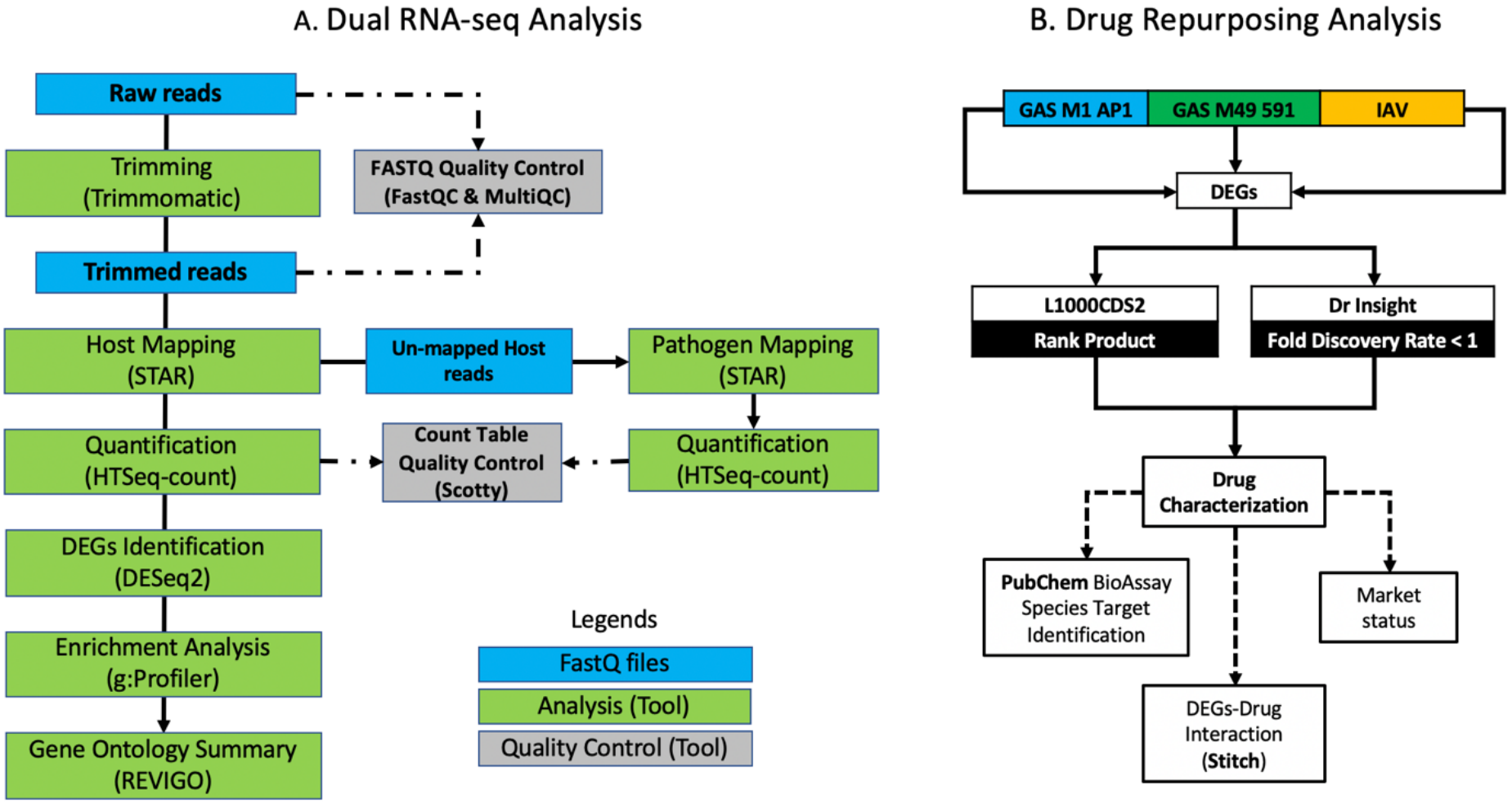
The reads were sequenced at Chronix Biomedicals on Illumina NextSeq 500 platform, as strand-specific (reverse-stranded) with single-end 75 bp reads (~30M reads per sample). The raw reads were trimmed with Trimmomatic version 0.39 (Bolger, Lohse, and Usadel 2014). Before and after trimming, the read quality of the FastQ files was checked with FastQC version 0.11.8 (Andrews S. 2010) and MultiQC version 1.2 (Ewels, Lundin, and Max 2016). Trimmed reads were mapped against the host genome with STAR version 2.7.1a (Dobin and Gingeras 2015). Then, the un-mapped host reads were aligned against the corresponding pathogen genome with STAR version 2.7.1a (Dobin and Gingeras 2015). GRCh38 and GRCh38.95 were used as the human reference genome and its annotation, respectively. On the other hand, ASM99376v1 (GenBank: GCA_000993765.1), ASM1812v1 (GenBank: GCA_000018125.1), and ViralMultiSegProj15622 (GenBank: GCA_000865085.1) reference genomes were used for GAS M1-AP1, GAS M49-591 and IAV, respectively. The gene count tables were prepared with the HTSeq-count version 0.6. tool (Anders, Pyl, and Huber 2015). The Scotty tool (Busby et al. 2013b) was used to measure the transcription depth and the coverage of the gene (p-value=0.05 and Fold Change=2) of the generated count tables. DEGs identification was achieved using DESeq2 version 1.24.0 tool (adjusted p-value < 0.0005) (Love, Huber, and Anders 2014). The adjusted p-value threshold of 0.0005 was chosen because it results in a number of false positives that is less than 1 in all three group comparisons (adjusted p-value * number of DEG < 1). Enrichment analysis was carried out with g:Profiler version e105_eg52_p16_e84549f (p-value=0.05) (Raudvere et al. 2019). The enriched terms were summarised and visualised with the use of REVIGO (Supek et al. 2011) or rrvgo R package (Sayols 2020) (Figure 5.1A). The scripts are available at https://github.com/SalemSueto-BioInfo/Dual_RNAseq_Hsapiens_Spyogenes_IAV.

5.3.2.2. Drug repurposing analysis

The following platforms were selected for DR analysis: L1000CDS2 (Duan et al. 2016b) and Dr Insight version 0.1.1 (Chan et al. 2019). The tools use different statistical approaches to identify potential anti-infection drugs from drug-perturbed gene expression profile datasets: LINCS for L1000CDS2 and CMAP for Dr Insight (Musa et al. 2018). The list of DEGs from each infection process was used as input for both platforms. The reverse mode of the

L1000CDS2 platform was used to find the top 50 drugs capable of reversing the input data. As described by Duan et al, the rank product of the drugs was calculated between the IAV infection rank and the two GAS serotypes infection (Duan et al. 2016b). The rank product was used to identify drugs with the potential of assisting the host cells against the co-infection of IAV and GAS serotypes. Thus, the analysis filter and ranks for drugs that are present in both viral and bacterial infections under study (IAV and S. pyogenes M1; and, IAV and S. pyogenes M49). The drug results from the Dr Insight platform were filtered with a p-value < 0.05. The identified drugs were then characterised for their marketed status; their species target through the use of biological assays from the PubChem website (S. Kim et al. 2021); and, a drug-gene interaction network was constructed with Stitch version 5.0 (Szklarczyk et al. 2016) between the lists of DEGs and the identified drug (Figure 5.1B). The DR scripts are available at https://github.com/SalemSueto-BioInfo/Dual_RNAseq_Hsapiens_Spyogenes_IAV.

Figure 5.1 (A) Dual RNA-seq Analysis. The figure shows the pipeline used for the analysis of the transcriptomics data. The name of the analysis step is written on top and in parenthesis the name of the tool used. (B) Drug Repurposing Analysis. The pipeline shows the step used during the DR analysis with the usage of the two platforms: L10000CDS2 and Dr Insight.



5.4. Results

5.4.1. Sample summary

The twelve sequencing runs delivered an average of 32 million raw reads per sample. After trimming, an average of 99.74% was retained from each sample. Trimmed reads were mapped to the reference genomes and an average of 95%, 0.27%, 4.02%, and 0.06% of the total raw reads were attributed to the human host, GAS M1-AP1, GAS M49-591, and IAV, respectively. The gene quantification phase for infected samples showed the host at an average of 26%; GAS M1-AP1 at 0.02%; GAS M49-591 at 0.04%; and, IAV at 0.01% of the total raw reads (Table 5.1). The analysis shows that samples of the same group cluster together (Figure 5.2). Furthermore, the Control group and the GAS M1-AP1 infected samples are the closest groups; followed by the GAS M49-591 infected samples; and, the most isolated group is the IAV-infected samples.

Figure 5.2 PCA of the gene expression of the twelve samples.

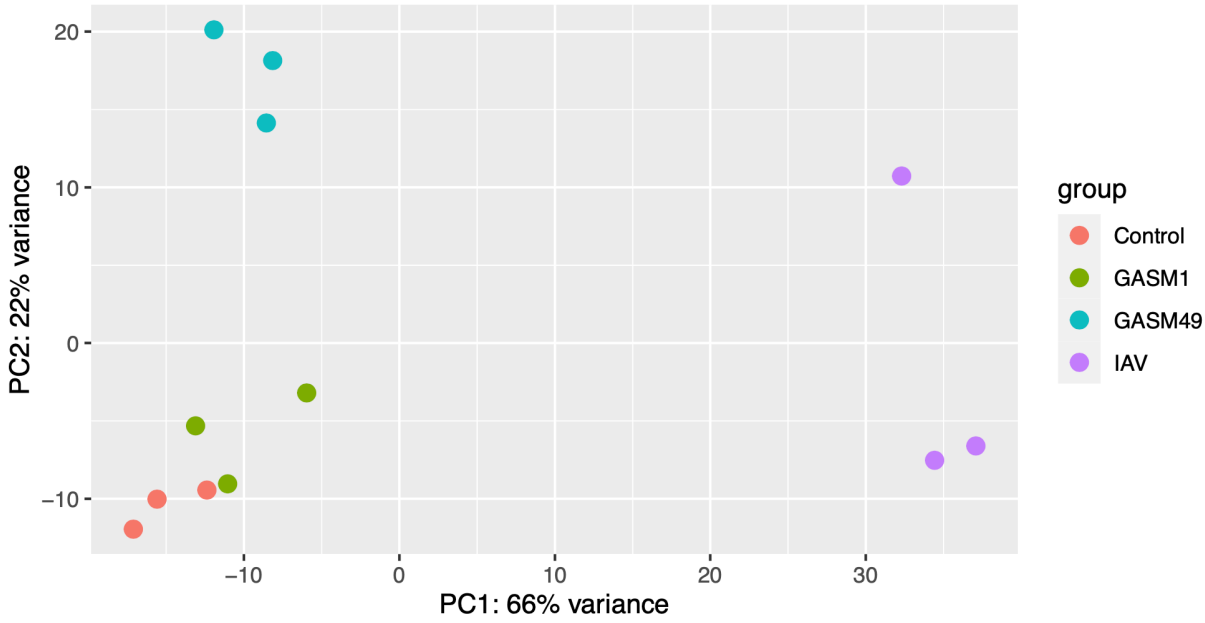


Table 5.1 Mapping statistics of the RNA sequencing libraries.

				Detroit 562 + Pathogen		Detroit 562		Pathogen	
Sample	**Species**	**Raw (Million)**	**Trimmed (Million)**	**Mapped (Million)**	**Count (Million)**	**Mapped (Million)**	**Count (Million)**	**Mapped (Million)**	**Count (Thousand)**
Sample 01	Detroit 562	29.51 (100%)	29.45 (99.81%)	28.86 (97.81%)	9.47 (32.10%)	28.86 (97.81%)	9.47 (32.10%)	0.00 (0.00%)	0 (0%)
Sample 02	Detroit 562	34.06 (100%)	34.00 (99.82%)	33.38 (98.01%)	10.79 (31.66%)	33.38 (98.01%)	10.79 (31.66%)	0.00 (0.00%)	0 (0%)
Sample 03	Detroit 562	34.86 (100%)	34.78 (99.78%)	34.23 (98.19%)	8.85 (25.38%)	34.23 (98.19%)	8.85 (25.38%)	0.00 (0.00%)	0 (0%)
Sample 04	Detroit 562 + GAS M1	32.18 (100%)	32.11 (99.79%)	31.61 (98.23%)	10.07 (31.30%)	31.53 (97.99%)	10.07 (31.29%)	0.08 (0.24%)	5.40 (0.02%)
Sample 05	Detroit 562 + GAS M1	31.43 (100%)	31.36 (99.78%)	30.76 (97.87%)	11.23 (35.74%)	30.66 (97.54%)	11.23 (35.72%)	0.11 (0.34%)	5.76 (0.02%)
Sample 06	Detroit 562 + GAS M1	33.36 (100%)	33.25 (99.67%)	32.79 (98.29%)	6.56 (19.66%)	32.71 (98.04%)	6.55 (19.64%)	0.08 (0.24%)	6.06 (0.02%)
Sample 07	Detroit 562 + GAS M49	34.23 (100%)	34.12 (99.67%)	33.35 (97.43%)	5.99 (17.51%)	32.62 (95.28%)	5.98 (17.48%)	0.73 (2.15%)	9.28 (0.03%)
Sample 08	Detroit 562 + GAS M49	31.87 (100%)	31.78 (99.72%)	30.67 (96.22%)	7.50 (23.55%)	29.81 (93.54%)	7.49 (23.51%)	0.86 (2.68%)	10.57 (0.03%)
Sample 09	Detroit 562 + GAS M49	33.22 (100%)	33.13 (99.75%)	31.70 (95.44%)	10.29 (30.98%)	29.30 (88.19%)	10.27 (30.93%)	2.41 (7.24%)	16.97 (0.05%)
Sample 10	Detroit 562 + IAV	30.01 (100%)	29.91 (99.66%)	28.85 (96.12%)	7.40 (24.64%)	28.83 (96.06%)	7.39 (24.63%)	0.02 (0.059%)	2.64 (0.01%)
Sample 11	Detroit 562 + IAV	28.05 (100%)	27.97 (99.71%)	26.59 (94.79%)	9.65 (34.39%)	26.58 (94.73%)	9.65 (34.39%)	0.02 (0.06%)	2.26 (0.01%)
Sample 12	Detroit 562 + IAV	35.22 (100%)	35.12 (99.70%)	34.07 (96.72%)	7.14 (20.29%)	34.05 (96.67%)	7.14 (20.28%)	0.02 (0.05%)	2.26 (0.01%)
Average	Detroit 562	32.81 (100 %)	32.74 (99.80 %)	32.16 (98 %)	9.7 (29.71 %)	32.16 (98 %)	9.47 (29.71 %)	0 (0 %)	0 (0 %)
	Detroit 562 + GAS M1	32.32 (100%)	32.24 (99.75 %)	31.72 (98.13 %)	9.29 (28.9 %)	31.63 (97.86 %)	9.28 (28.88 %)	0.09 (0.27 %)	5.7 (0.02 %)
	Detroit 562 + GAS M49	33.11 (100 %)	33.01 (99.71 %)	31.91 (96.36 %)	7.93 (24.01 %)	30.58 (92.34 %)	7.91 (23.97 %)	1.33 (4.02 %)	12.28 (0.04 %)
	Detroit 562 + IAV	31.09 (100 %)	31 (99.69 %)	29.84 (95.88 %)	8.06 (26.44 %)	29.82 (95.82 %)	8.06 (26.43 %)	0.02 (0.06 %)	2.39 (0.01 %)

5.4.2. Count table quality control

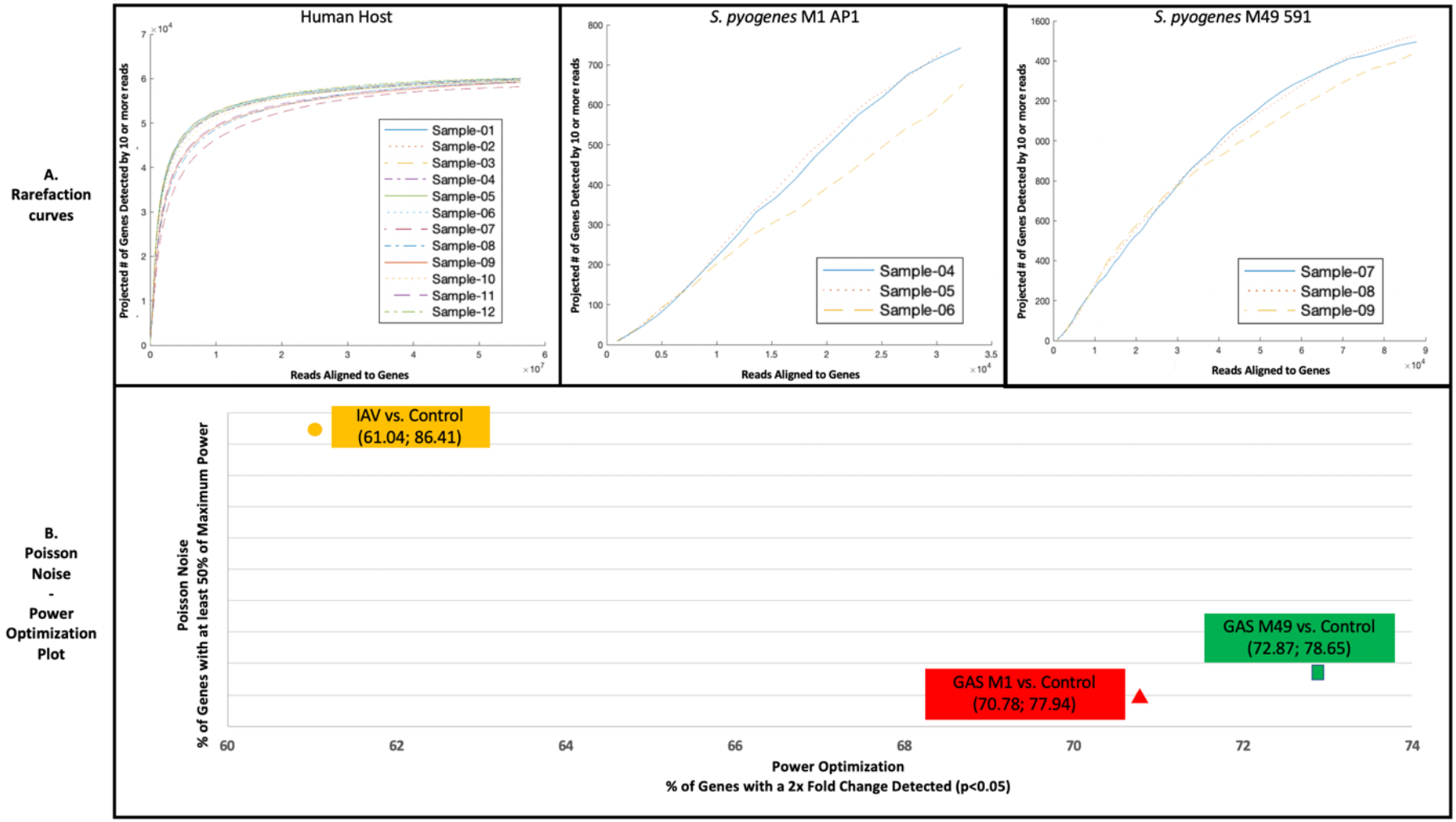
The Scotty tool was specifically created for RNA-seq data analysis (Busby et al. 2013b). It allows researchers to select the right number of biological replicates and read depth to maximise the statistical power for the identification of gene diversity in a sample and DEGs among two sample groups. The tool uses as input data the gene count table obtained for the human host and the three pathogens. The optimization is achieved by excluding configurations that necessitate a high number of replicates; high cost due to excessive read depths; and, configurations where measurements bias is present for a substantial number of genes. The present study focused on the following three plots: rarefaction curves, Power Optimization, and Poisson Noise.

The Power Optimization plot calculates the probability of the count table to identify all DEGs between two groups. The proportion of DEGs from the three infections that could be detected with a 2-fold change ($p < 0.05$) was in the range of 61.04% to 72.87%. Therefore, potentially 28%-39% of DEGs could not be detected with our approach.

The Poisson Noise plot measures the variance in RNA-seq data that occurs due to each specific RNA being selected at random among the total RNA data and counted. The variance is relative to the total count for genes with higher values being related to low counts compared to high counts. For example, the difference in expression for a gene counted with five reads versus ten reads is fundamentally less certain than a gene measured with 500 versus 1000 reads, even though both cases show a 2X fold change. The results show the values to be between 77.94% and 86.41% for the three group comparisons. Thus, our dataset has a 14%-22% of the genes with relatively high uncertainty due to its low read counts, particularly the ones below ~10 reads (Figure 5.3B).

The gene quantification was determined for both the human host and the three pathogens. Nonetheless, the pathogen reads do not provide a satisfying representation of the bacterial reads as shown by the lack of a plateau phase in the rarefaction curves (Figure 5.3A). Thus, the downstream analysis will be solely focused on the host expression data.

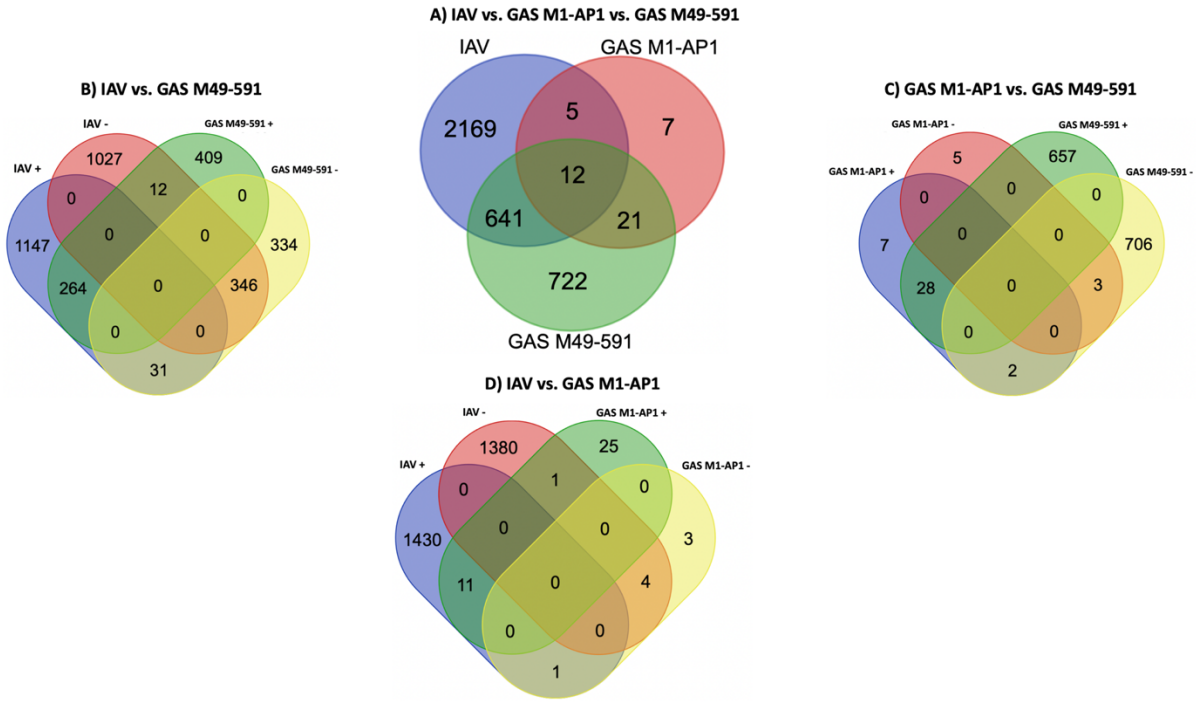
Figure 5.3 Scotty analysis. (A) Rarefaction curves. The rarefaction curves showcase the human host reads and the two GAS strains; whereas, the rarefaction curves of IAV are not present due to the inability to build the curves with a low number of genes. (B) Poisson Noise – Power Optimization plot. The Power plot and the Poisson noise plot were analysed at 2X Fold Change and $p<0.05$. The working conditions of our experimental setup are 3 replicates with an average of 8.74M of reads aligned to genes per replicate.



5.4.3. Identification of DEGs

The identification of host DEGs for the three infections shows infection-specific DEGs and shared ones (Figure 5.4). The IAV infection shows 2169 DEGs that are specific; the GAS M1-AP1 infection process has 7 specific DEGs; and, the GAS M49-591 infection presents 722 specific DEGs. The shared genes are the following: 641 DEGs are shared among IAV and GAS M49-591 infection; 5 DEGs are in common between IAV and GAS M1-AP1; 21 DEGs are shared among the two GAS strains; and, 12 DEGs are in common among all three pathogens.

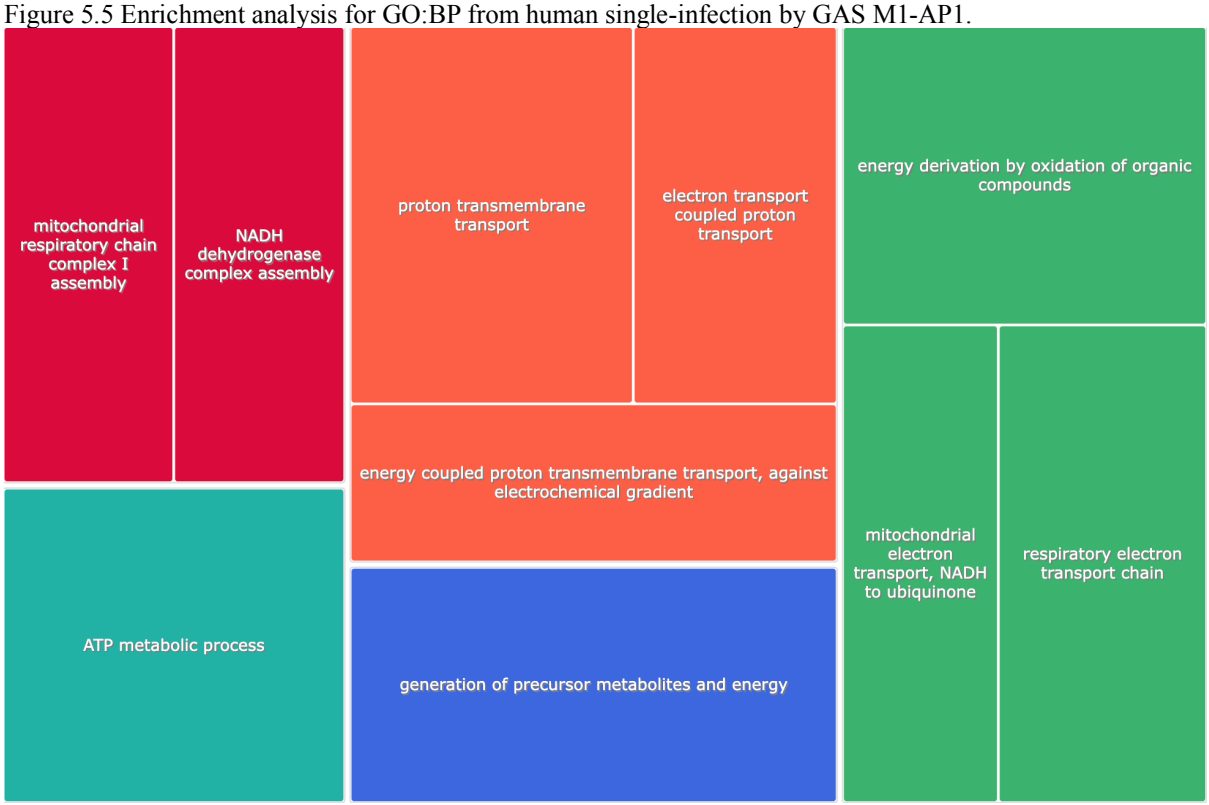
Figure 5.4 DEGs identification and comparison among the three infection process.



5.4.4. Human epithelial transcriptomics response from GAS M1-AP1 infection

The infection process caused by GAS M1-AP1 identifies 45 host DEGs. The enrichment analysis of these genes results in the identification of 18 enriched terms for the GO:BP class. The terms can be grouped into one main group: oxidative respiration (Figure 5.5). The group is characterized by terms such as “mitochondrial respiratory chain complex I assembly”, “mitochondrial electron transport, NADH to ubiquinone”, and “energy coupled proton transmembrane transport, against electrochemical gradient”.

Figure 5.5 Enrichment analysis for GO:BP from human single-infection by GAS M1-AP1.

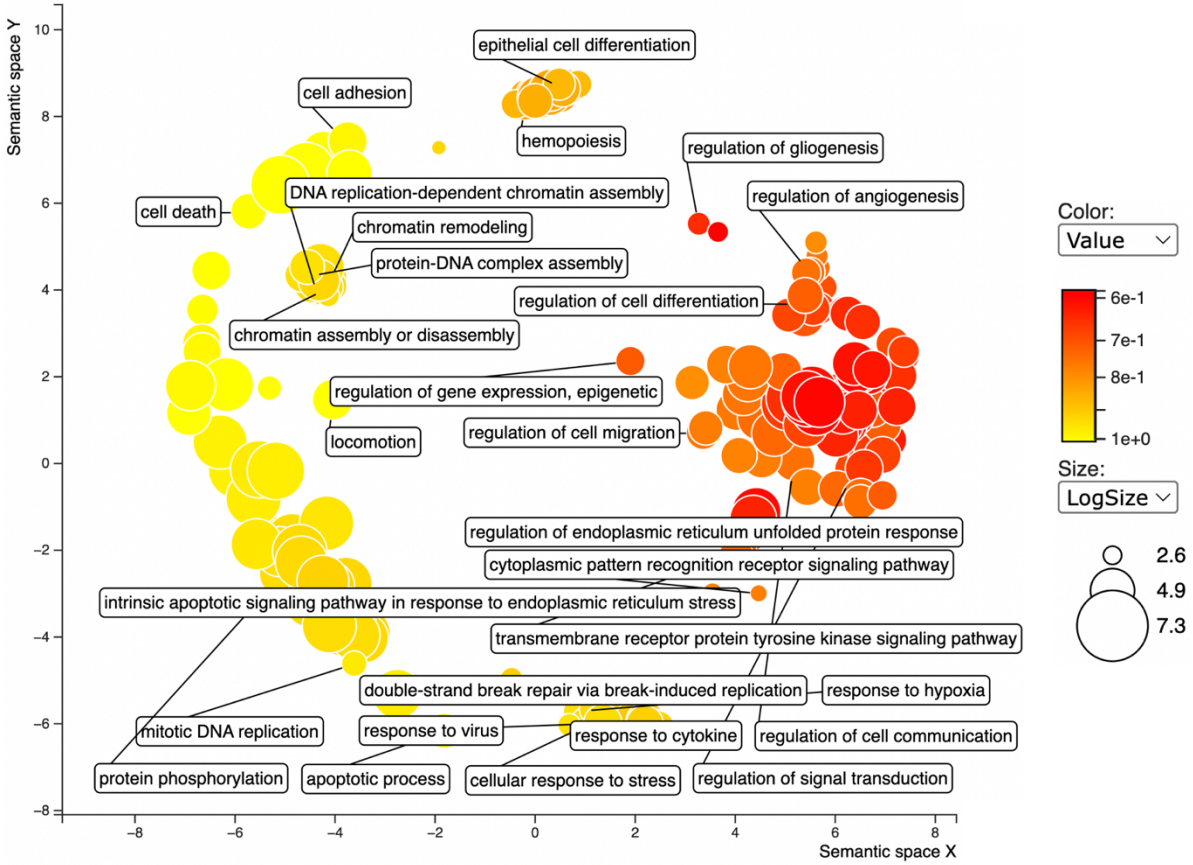


5.4.5. Human epithelial transcriptomics response from GAS M49-591 infection

The infection caused by GAS M49-591 identified 1396 human DEGs that resulted in 224 enriched terms for the GO:BP class. The terms were clustered based on their function (Figure 5.6). One of the main functions is the regulation of signal transduction. This function shows the following terms: “transmembrane receptor protein tyrosine kinase signaling pathway” and “cytoplasmic pattern recognition receptor signaling pathway”. These terms suggest a binding event between a human transmembrane protein to an extracellular ligand that results in the regulation of downstream processes such as gene transcription. The regulation of gene expression is shown to be relevant with terms like “positive regulation of transcription by RNA polymerase II” and “regulation of gene silencing”. The analysis also shows that DNA replication is affected alongside chromosome organization with terms like “chromatin assembly or disassembly”, “DNA packaging”, “telomere organization”, and “double-strand break repair via break-induced replication”. Several cell differentiation processes were found. One such process is cell differentiation for epithelial and keratinocyte cells; differentiation into haematopoietic immune system cells; and, regulation of gliogenesis. The regulation of cell stress is highlighted by terms related to the regulation of endoplasmic reticulum unfolded protein response and the regulation of apoptotic signaling pathway in response to endoplasmic reticulum stress. The host shows terms related to blood vessel development terms like “vascular endothelial growth factor production”, “angiogenesis”, and “tube morphogenesis”. Another

regulated function is cell migration where terms like "regulation of cellular component movement" and "movement of cell or subcellular component" suggests a novel organization of the organelle position. Furthermore, the analysis shows several terms related to the host response to elements such as cytokine, hypoxia, and virus. The term "cell-cell adhesion" was also identified as one of the main grouped terms.

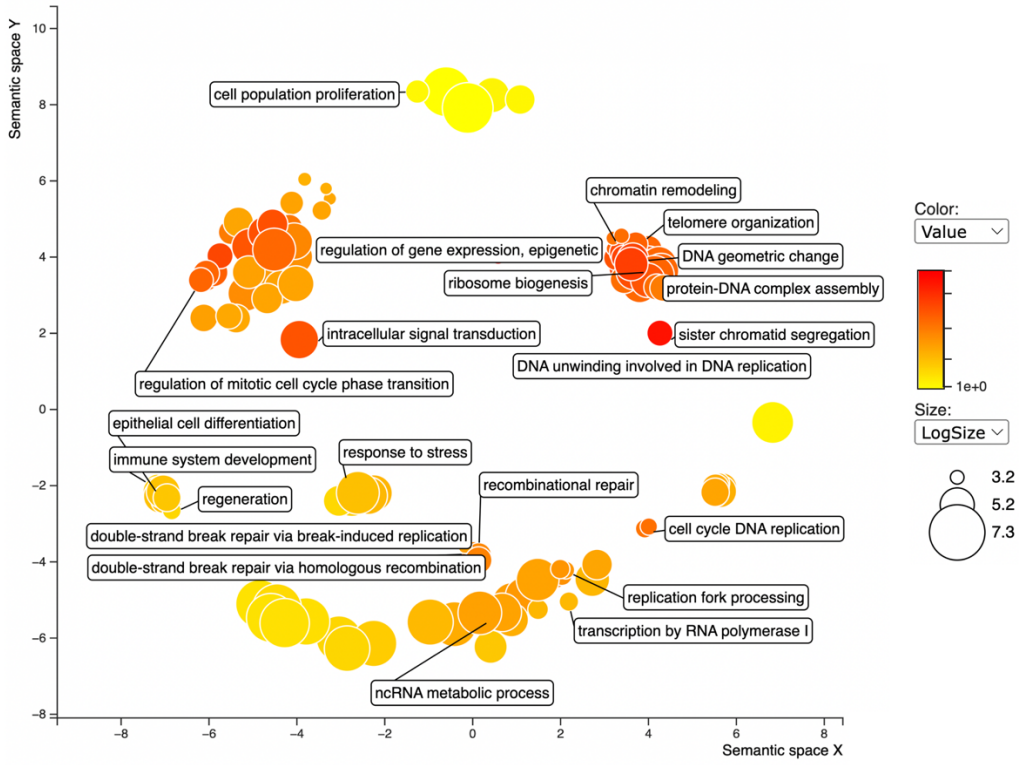
Figure 5.6 Enrichment analysis for GO:BP from human single-infection caused by GAS M49-591. The plot was created with the website version of REVIGO. The semantic space axes in the plot have no intrinsic meaning. The Revigo tool uses the Multidimensional Scaling (MDS) methodology to reduce the dimensionality of a matrix of the GO terms pairwise sematic similarities. GO terms that are semantically similar should cluster closer in the plot. The highlighted GO terms were selected as the most significant ones in the clusters or the one with the most specific biological information. The bubble color shows the user-provided p-value and the size indicates the frequency of the GO term in the GOA database (the bubbles of more general terms are larger) (Barrell et al. 2009).



5.4.6. Human epithelial transcriptomics response from IAV infection

The infection caused by IAV identified 2827 human DEGs that resulted in 133 enriched terms for the GO:BP class (Figure 5.7). One of the main identified cellular processes is related to the cell cycle and DNA replication. The cluster shows terms like "DNA replication initiation", "mitotic nuclear division", and "sister chromatid segregation". The second main group describes chromosome organization with terms like "chromatin remodelling", "telomere organization", "DNA packaging", and "DNA geometric change". Another cluster related to the chromosome organization process is the DNA repair system with terms like "double-strand break repair via homologous recombination", "double-strand break repair via break-induced replication", and "recombinational repair". Furthermore, the IAV infection shows cell differentiation for epithelial and keratinocyte cells; regulation of ribosome biogenesis; regulation of gene transcription with terms like "regulation of gene silencing" and "transcription by RNA polymerase I"; and, the term "regeneration" referred to as repair of damaged cells or tissue.

Figure 5.7 Enrichment analysis for GO:BP from human single-infection caused by IAV. The plot was created with the website version of REVIGO. The semantic space axes in the plot have no intrinsic meaning. The Revigo tool uses the Multidimensional Scaling (MDS) methodology to reduce the dimensionality of a matrix of the GO terms pairwise sematic similarities. GO terms that are semantically similar should cluster closer in the plot. The highlighted GO terms were selected as the most significant ones in the clusters or the one with the most specific biological information. The bubble color shows the user-provided p-value and the size indicates the frequency of the GO term in the GOA database (the bubbles of more general terms are larger) (Barrell et al. 2009).



5.4.7. Drug repurposing analysis

Two DR platforms (L1000CDS2 and DrInsight) were used in our analysis to identify potential drugs that can aid the host defence during the infection. The platforms use as input a list of DEGs or their expression (e.g. fold change). Thus, for each platform, we had three lists of drugs related to the three infections. Each platform identifies the drugs with a different methodology; therefore, the selection of potential IAV-GAS anti-infective drugs was tailored based on which platform the drug was first found. The L1000CDS2 web tool accepts as input a list of genes or their expression. In our analysis, we used the expression data of the DEGs from the three infections, separately. The web tool uses the CD method to identify drugs (top 50) capable of reversing the gene expression of the input data; and, the rank product methodology (product between the same drug in two distant outputs) was applied to find drugs that are common between IAV and each GAS infection (Duan et al. 2016b). Meanwhile, Dr Insight is an R package and it measures the concordance (e.g. inverse association) between the input data and the drug-perturbated data from the CMAP database (Chan et al. 2019).

L1000CDS2 platforms identified 3 drugs potentially suitable to treat S. pyogenes M1 and IAV; whereas, 18 drugs were identified as a suitable treatment for S. pyogenes M49 and IAV; and, Dr Insight found 3 potential drugs (Table 5.2).

Table 5.2 List of drugs identified through the use of DR strategy.

DR Platform	Infection	Drug	Broad molecule ID	Trade Names	PubChem BioAssay	STITCH	Rank product	p-value	FDR	Pharmaceutical Usage
L1000CDS2	IAV & GAS M1	528116.cdx	BRD-K49371609	N. A.	Present	N. A.	369	N. A.	N. A.	N. A.
		GSK-690693	BRD-A87137733	N. A.	Present	GSK-690693	748			Anti-cancer
		AZD-5438	BRD-K72414522	N. A.	Present	AZD-5438	1716			Anti-cancer
	IAV & GAS M49	BRD-U74615290	BRD-U74615290	N. A.	N. A.	N. A.	5			N. A.
		NCGC00012272-02	BRD-K07668032	N. A.	Present	AC1MM8JC	12			Pre-mRNA splicing enzymes inhibitor
		nifedipine	BRD-A30977374	Adalat - Procardia	Present	nifedipine	18			high blood pressure
		NCGC00183235-01	BRD-K30443205	N. A.	Present	AGN-PC-071I9C	22			N. A.
		PKCbeta inhibitor	BRD-K89687904	N. A.	Present	PKCbeta inhibitor	31			degenerative brain disease
		Kenpaullone	BRD-K37312348	N. A.	Present	Kenpaullone	56			Anti-cancer
		BRD-K74980345	BRD-K74980345	N. A.	Present	N. A.	63			Pre-mRNA splicing enzymes inhibitor
		BRD-K43913647	BRD-K43913647	N. A.	Present	N. A.	143			Anti-cancer
		BRD-K56024573	BRD-K56024573	N. A.	Present	N. A.	182			N. A.
		BRD-K57037351	BRD-K57037351	N. A.	Present	N. A.	200			N. A.
		BRD-K91370081	BRD-K91370081	Anisomycin - Flagecidin	Present	N. A.	270			Anti-microbial
		CHIR-99021	BRD-K16189898	N. A.	Present	CHIR99021	288			enzyme GSK-3 inhibitor
		KIN001-043	BRD-K44100512	N. A.	N. A.	N. A.	300			N. A.
		Emetine Dihydrochloride Hydrate (74)	BRD-K01976263	Emetine	Present	Emetine	399			Anti-protozoal
		AG-14361	BRD-K00615600	N. A.	Present	AG14361	714			Anti-cancer
		BRD-K78189262	BRD-K78189262	N. A.	Present	N. A.	799			N. A.
		THZ-2-98-01	BRD-U41416256	N. A.	N. A.	N. A.	1376			N. A.
		NCGC00183232-01	BRD-K80126354	N. A.	Present	AGN-PC-071I9B	2200			N. A.

Dr Insight	GAS M1	LY-294002	BRD-K27305650	N. A.	Present	LY294002 hydrochloride	N. A.	2.62E-05	0.09398182	Anti hepatitis C virus
		Sirolimus	BRD-K84937637	Rapamycin	Present	rapamycin		5.88E-05	0.10543595	organ transplant rejection
			BRD-K89626439							organ transplant rejection
	IAV	Estradiol	BRD-K18910433	Oestradiol	Present	estradiol		9.28E-05	0.33280894	hormone therapy

The PubChem BioAssay permits the identification of the list of species that have been tested with the drug in question (Figure 5.8). The heatmap shows Homo sapiens to be subjected to all identified drugs except for the PKCbeta inhibitor. Three IAV species (641809, 162536, 382835) and one S. pyogenes species (1314) were found to be tested by the following drugs: AG-14361, BRD-K43913647, Emetine, NCGC00012272-02, BRD-K57037351, NCGC00183235-01, BRD-K56024573, NCGC00183232-01, BRD-K78189262, BRD-K91370081, nifedipine, estradiol, and kenpaullone. BRD-K91370081 is the only drug that was found to be tested for both IAV (162536) and S. pyogenes (1314) species. 3 drugs do not have any BioAssay present: BRD-U74615290, KIN001-043, and THZ-2-98-01. The remaining 21 drugs have been tested on different families of both bacteria and viruses.

The Stitch web tool was used to identify the drug-gene interaction between the 24 identified drugs and the DEGs found from the three infection conditions (Figure 5.9A). The Stitch score is the integration of multiple types of evidence combined into one numerical score and it provides a level of confidence for the interaction (from 0 to 1). The analysis shows no interaction between the identified drugs and the exclusive DEGs from the GAS M1 and GAS M1 – IAV exclusive groups. The GAS M49 infection-specific DEGs group showed Rapamycin with the highest score (0.989) and Estradiol with the highest number of interactions (22). The IAV infection-exclusive DEGs showed the highest score with Rapamycin (0.992) and the highest number of interactions with Estradiol (25). For the DEGs exclusive to both GAS serotypes infection groups, Estradiol has both the highest score and the highest number of interactions at 0.987 and 4, respectively. For the DEGs shared by both GAS serotype M49 and IAV infection groups, Estradiol has the highest number of interactions (25), whereas Rapamycin has the highest score at 0.992. From the group of DEGs common to all three infections, only the MT-CO1 DEG and the Estradiol drug interaction were found.

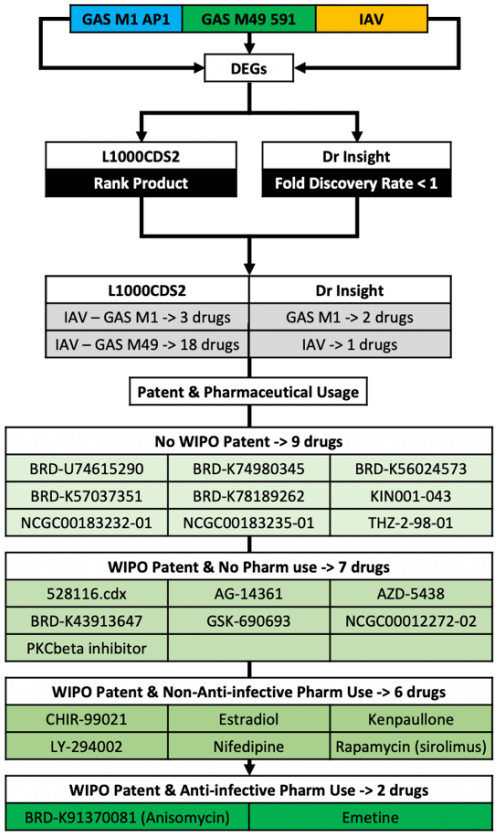
The analysis of the WIPO patent and the pharmaceutical usage of the 24 identified drugs resulted in the formation of 4 groups (Figure 5.9B). The following 9 drugs were found to have no WIPO patent: BRD-U74615290, BRD-K74980345, BRD-K56024573, BRD-K57037351, BRD-K78189262, KIN001-043, NCGC00183232-01, NCGC00183235-01, and THZ-2-98-01. The next group includes 7 drugs with known WIPO patents with no found pharmaceutical use: 528116.cdx, AG-14361, AZD-5438, BRD-K43913647, GSK-690693, NCGC00012272-02, and PKCbeta inhibitor. The 6 drug members of the third group have WIPO (World Intellectual Property Organization) patents and have pharmaceutical use but no known antibiotic properties. Nifedipine is used as a calcium channel blocker used to treat hypertension and angina pectoris (Drugbank 2022). Kenpaullone has a role as a geroprotector, a tau-protein

kinase inhibitor, a cardioprotective agent, and a cyclin-dependent kinase inhibitor (Tocris 2022c). CHIR-99021 is known as a tau-protein kinase inhibitor (Tocris 2022b). LY-294002 is an inhibitor of phosphatidylinositol 3-kinase (PI3K) and the bromodomain and extra-terminal family of proteins, with potential antineoplastic activity (Tocris 2022d). Therapeutic Estradiol is a synthetic form of estradiol, a steroid sex hormone vital to the maintenance of fertility and secondary sexual characteristics in females (PubChem 2022). Rapamycin (sirolimus) is a macrolide compound, firstly discovered in Streptomyces hygroscopicus, and used to prevent organ transplant rejection, treat a rare lung disease called lymphangioleiomyomatosis, and treat perivascular epithelioid cell tumor (Tocris 2022e). The last group has 2 drug members: BRD-K91370081 (Anisomycin) and Emetine. The drugs have WIPO patents and they have known antibiotics use against bacterial and/or virus species in the pharmaceutical market. BRD-K91370081 (also known as Anisomycin or Flagecidin) is an antibiotic, first discovered in Streptomyces griseolus, which inhibits protein biosynthesis in eukaryotic (Tocris 2022a). The Emetine drug is a pyridoisoquinoline and it has wide pharmaceutical usage such as antiprotozoal, antiviral, antimalarial, protein synthesis inhibitor, anticoronaviral, antiamoebic, and emetic. Furthermore, Emetine is known to inhibit SARS-CoV2, Zika, and Ebola virus replication (PubChem 2021; Tocris 2020).

Figure 5.8 PubChem BioAssay Heatmap – Dendrogram clustering – The figure shows the species target (Y axis) for the identified drugs (X axis), from the drug repurposing analysis, as obtained from the list of bioassays present in the PubChem. Additionally, a dendrogram clustering is highlighted for both the species and the drugs.

Figure 5.9 (A) Drug Repurposing identification pipeline. The identified DEGs were used as input data for the two DR platforms. The filtering of the drugs was carried out as follows. The rank product of the drugs from the L1000CDS2 was calculated between the IAV against the two GAS infections; and, the selection was carried out for the ones that are present in both viral and bacterial infections. The drug selection from the DrInsight R-tool was carried out for the one with an p-value < 0.05. The identified drugs were then divided into several groups based on their patented status and their possible pharmaceutical use. (B) Drug-Protein interaction network (Stitch). The drug-protein interaction network was created with the STITCH web application between the lists of DEGs and the identified drugs during the DR analysis. The Venn diagram shows the number of DEGs from the three infections. The tables show the list of drug that interacts with that specific infection group, the number of interactions between the drug and the DEGs, and the confidence score (maximum and minimum) for such interaction. The score rank from 0 to 1, with 1 as the highest value.

A. Drug Repurposing Analysis



L1000CDS2	Dr Insight
IAV – GAS M1 -> 3 drugs	GAS M1 -> 2 drugs
IAV – GAS M49 -> 18 drugs	IAV -> 1 drugs

Patent & Pharmaceutical Usage

No WIPO Patent -> 9 drugs		
BRD-U74615290	BRD-K74980345	BRD-K56024573
BRD-K57037351	BRD-K78189262	KIN001-043
NCGC00183232-01	NCGC00183235-01	THZ-2-98-01

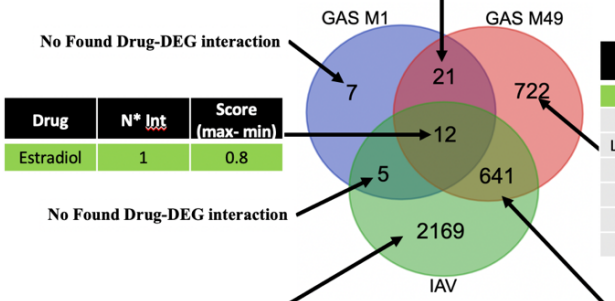
WIPO Patent & No Pharm use -> 7 drugs		
528116.cdx	AG-14361	AZD-5438
BRD-K43913647	GSK-690693	NCGC00012272-02
PKCbeta inhibitor		

WIPO Patent & Non-Anti-infective Pharm Use -> 6 drugs		
CHIR-99021	Estradiol	Kenpaullone
LY-294002	Nifedipine	Rapamycin (sirolimus)

WIPO Patent & Anti-infective Pharm Use -> 2 drugs	
BRD-K91370081 (Anisomycin)	Emetine

B. Drug-Protein interaction network (Stitch)

Drug	N* Interaction	Score (max- min)
Estradiol	4	0.987 - 0.84
Rapamycin	2	0.845 - 0.804
LY294002 hydro.	3	0.877 - 0.8



Drug	N* Int	Score (max- min)
Estradiol	1	0.8

Drug	N* Int	Score (max- min)
Rapamycin	20	0.989 - 0.411
Estradiol	22	0.963 - 0.467
LY294002 hydro.	16	0.956 - 0.45
Nifedipine	1	0.955
GSK690693	2	0.911 - 0.89
Emetine	2	0.674 - 0.548
PKCbeta inhibi.	3	0.603 - 0.403

Drug	N* Int	Score (max- min)
Rapamycin	20	0.992 - 0.401
Estradiol	25	0.968 - 0.445
LY294002 hydro.	10	0.904 - 0.471
Kenpaullone	1	0.857
AZD5438	1	0.47

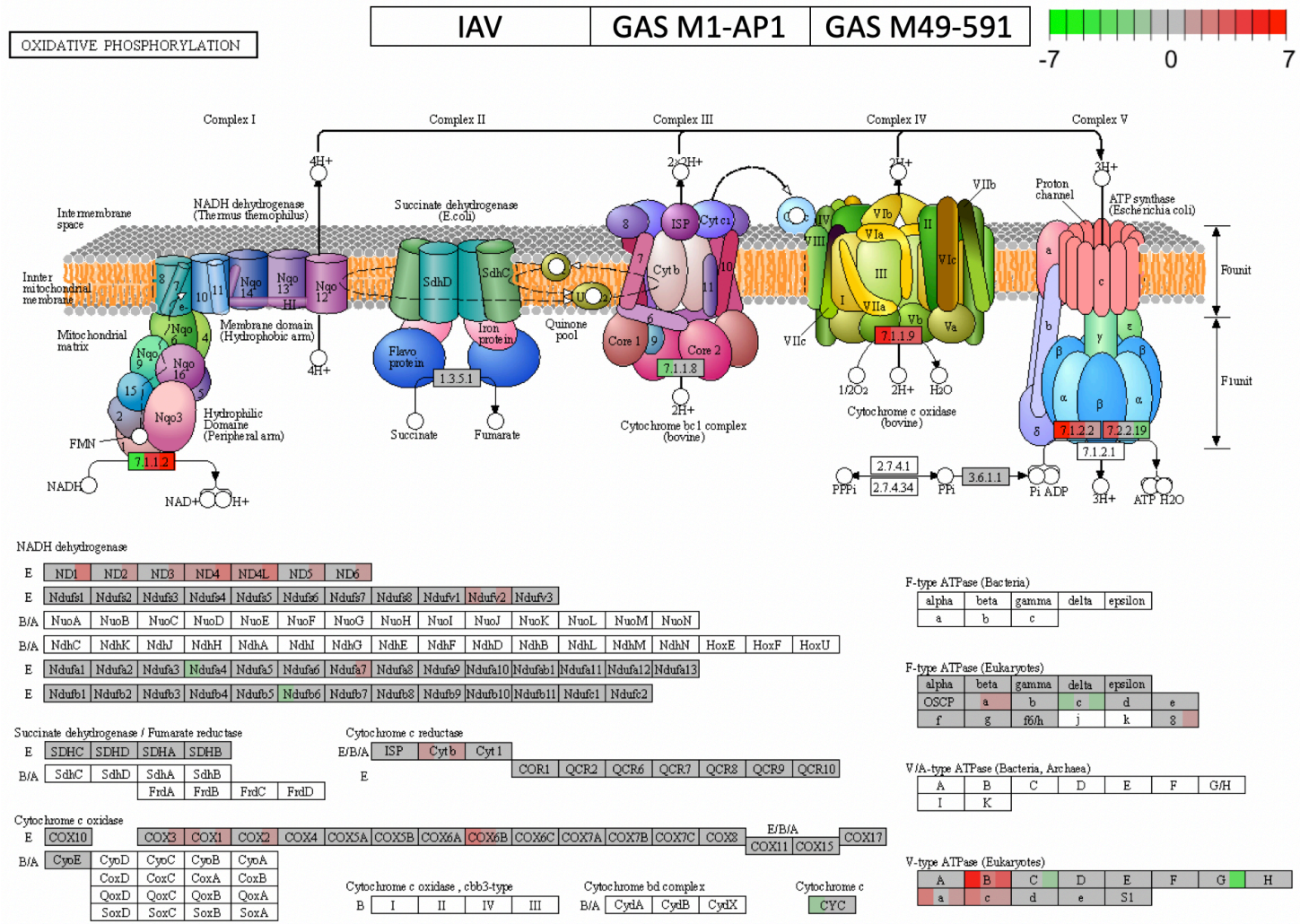
Drug	N* Int	Score (max- min)
Rapamycin	20	0.992 - 0.401
Estradiol	25	0.968 - 0.445
LY294002 hydro.	10	0.904 - 0.471
Kenpaullone	1	0.857
AZD5438	1	0.47

5.5. Discussion

5.5.1. Upregulation of oxidative respiration process among the three infections

The analysis shows 12 DEGs that are shared among all three infections (Figure 5.4). The enrichment analysis of the 12-shared-DEGs shows enriched terms related to oxidative phosphorylation and with GO:CC terms located in the mitochondrial respiratory chain complexes I and III (Figure 5.10). All identified enriched terms are based on the four mitochondrial DEGs (MT-ND4L, MT-ND4, MT-CO1, and MT-CYB). The four DEGs are upregulated under all three infections. The following DEGs are known to have a direct connection to bacterial and/or viral infections. The KLF6 gene was upregulated following all three infections and encodes a DNA-binding transcription factor. The gene is known to be activated by IAV to bind to the inducible NO synthase (iNOS) promoter, leading to iNOS transcription which is a key factor for apoptosis (Mgbemena et al. 2012). The SLC30A1 gene was downregulated under all three infection conditions. It encodes a member of the solute carrier family and regulates the transport of zinc. There is no known association between SLC30A1 and IAV infection. However, the gene is known to exert a negative effect on human cell survival during vesicular stomatitis virus infection (Moskovskich et al. 2019).

Figure 5.10 Oxidative phosphorylation KEGG pathway. The plot shows the mitochondrial complexes that regulate ATP production during oxidative phosphorylation. The Log Fold Change (-7 to +7) of the proteins in the mitochondrial complexes is shown starting from IAV, GAS M1-AP1, and GAS M49-591 infections.



5.5.2. Differential host responses to GAS and IAV infections

In this study, pharyngeal epithelial cells were infected with IAV, GAS M1-AP1, and GAS M49-591. The results of the study show the three pathogens engaged differently with the human host cells.

The 48 hours-post-infection with IAV was chosen as a time point for sampling to provide time for viral infection yet keep the host cells intact to capture the host transcriptional response to the infection. The viral infection did not result in noticeable cell death and cell death marker expression was negatively affected (data not shown). Moreover, a diminished concentration of pro-inflammatory cytokines IL-6 and IL-8 was observed in cell culture supernatants (data not shown) and a reduced expression of the IL-6 gene (ENSG00000136244) was also discovered in the RNA-Seq analysis (LFC = -2.15). Taking the abundance of count reads of viral transcripts into consideration, it can be concluded that virus production was still in process at the time point of sampling. This behavior is supported by the presence of enriched terms such as DNA damage and DNA damage response which have been previously reported as consequences following IAV infection (N. Li et al. 2015).

Infection with GAS M1-AP1 and GAS M49-591 revealed serotype-specific differences. Adherence of serotype M49 to host cells was twice as high compared to serotype M1 (data not shown). Such behavior has been shown by previous studies that streptococcal adherence to host cells is dependent on the infection model, the serotype, and the strain under investigation (Ryan and Juncosa 2016) (Berkower et al. 1999).

Infection with either GAS serotype resulted in decreased IL-8 concentration (data not shown). This observation can be ascribed to the activity of GAS virulence factor SpyCEP (or ScpC), an IL-8-degrading protease (Zinkernagel et al. 2008). Infection with serotype M49-591 caused an increase in IL-1β concentration. Also, an elevated expression of IL1B (ENSG00000125538) was observed in the RNA-seq analysis (LFC of 2.4). Previous studies attribute a critical defensive role to IL-1β during GAS infection (Hsu et al. 2011; LaRock and Nizet 2015).

Marked cell death was observed only after infection with GAS M49-591 (data not shown). The concept of pathogenic bacteria protecting their host cells from cell death has already been described for Neisseria gonorrhoeae (Binnicker, Williams, and Apicella 2003), Neisseria meningitidis (Massari et al. 2003), Chlamydia pneumoniae (Fischer et al. 2001), and Helicobacter pylori (Shirin et al. 2000). In the present study, enrichment analyses from infections with GAS M1-AP1 did not indicate cell death-related activity. However, GAS has

been reported to cause cell death in several epithelial cell lines, including Detroit 562 cells (Agarwal et al. 2012).

GAS is known to elicit unfolded protein response and ER stress in host cells (Baruch et al. 2014); and, prolonged ER stress leads to apoptotic cell death (Adams et al. 2019; M. Wang and Kaufman 2016). These biological processes were shown during GAS M49-591 infection with the identification of terms like "regulation of endoplasmic reticulum unfolded protein response" (GO:1900101), "response to endoplasmic reticulum stress" (GO:0034976), "cell death" (GO:0008219) and "apoptotic process" (GO:0006915). The present study shows GAS M49-591 infection displayed significant upregulation in gene expression of IRE1 (Adams et al. 2019), JNK1, MAPK14, TRAF2, PERK, CHOP, and GADD34. These findings suggest an apoptotic death modality involving ER stress.

5.5.3. IAV role on secondary bacterial GAS infection

As described previously, IAV is capable of enhancing a secondary infection of GAS. Though the processes underlying this link are not fully understood, there are known mechanisms that IAV promotes that end up benefitting a secondary bacterial infection.

5.5.3.1. IAV infection effect on cell-cell adhesion and extracellular matrix stability

Several studies, in vitro and in vivo, have shown the ability of IAV to disrupt the epithelial cell tight junctions and thereby favouring subsequent infections by Streptococci (Short et al. 2016; Nita-Lazar et al. 2015). The enrichment analysis of Detroit 562 cell DEGs following IAV infection revealed several enriched terms related to cell differentiation into epithelial cells and epithelium development (Figure 5.11).

The GO:BP term “biological adhesion” (GO:0022610) describes all the different types of attachment between cell-cell and cell-substrate such as the extracellular matrix. The present study identified 326 DEGs (status of differentially expressed from at least one infection process) that are part of the GO term. As classified by Zhong et al, a subset of 74 DEGs was classified as a cell adhesion molecule (CAM) and annotated into 5 groups: ag (primary roles in axonal guidance), fa (primary involvement in focal adhesions), i (information predominant CAM), m (primary involvement in interactions with cell-matrix), and tj (primary involvement in tight junctions); whereas, 252 DEGs were not classified as CAM by Zhong and colleagues (Zhong et al. 2015).

Among the CAM group, several members of the desmosome were found to be differentially upregulated only during the IAV infection: DSG2 (CAM class fa, LFC=1), DSC2 (CAM class

fa, LFC=2.6), PKP2 (CAM class fa, LFC=-1.7), DSP (CAM class tj, LFC=2.6), and JUP (CAM class tj, LFC=1.2). The desmosome, known as a macula adherent, is a specialized intercellular junction found in animal cells designed for cell-cell adhesion. The desmosome can be divided into three sections: the extracellular core region, the outer dense plaque, and the inner dense plaque (Delva, Tucker, and Kowalczyk 2009). The desmoplakin (e.g. DSP) proteins connect the keratin intermediate filaments, present in the inner dense plaque, to the plakoglobin (e.g. JUP) within the outer dense plaque. Two desmoplakin proteins are connected through their interaction with the plakophilins (e.g. PKP2). The desmogleins (e.g. DSG2) and desmocollins (e.g. DSC2) are transmembrane proteins present in the extracellular core region. The two transmembrane proteins interact with the plakoglobins. The loss of any member of the desmosome complex will result in a compromised complex and a disruption in the cell-cell physical interaction.

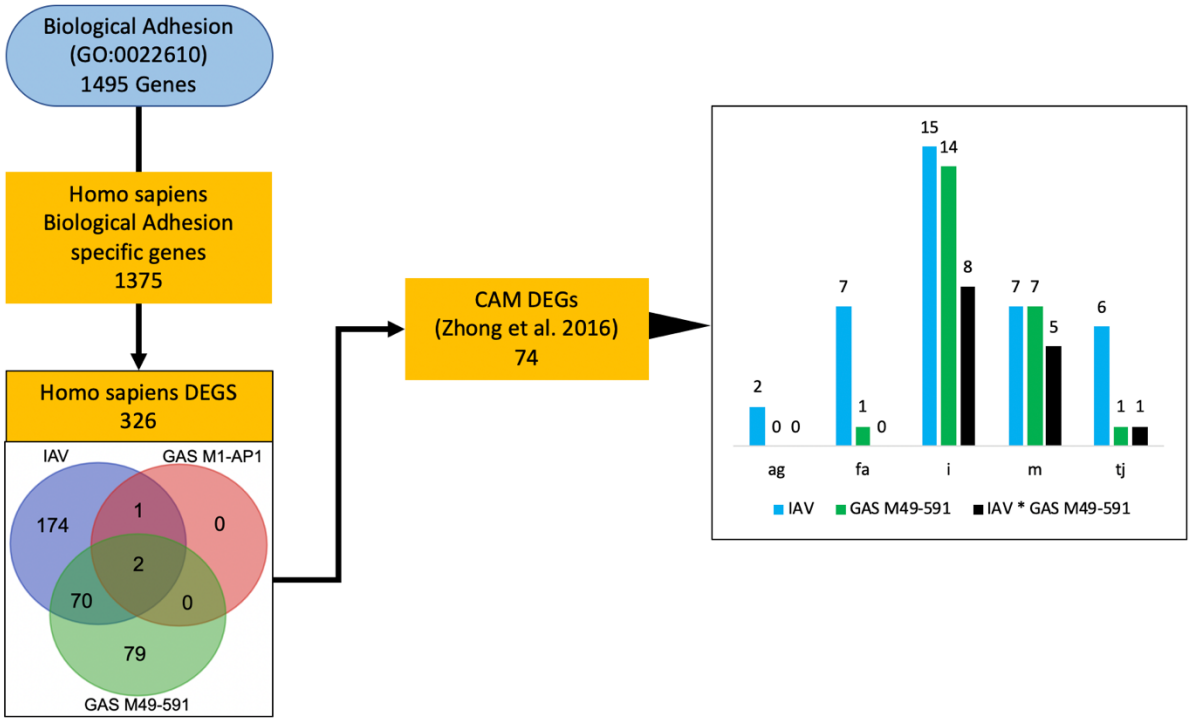
The DSG2 protein had been identified as the primary receptor used by adenovirus (H. Wang et al. 2011). Initially, the adenovirus binds to DSG2 and initiates intracellular signalling that concludes with the cleavage of the extracellular domain of the gene, thus eliminating the DSG2 homodimers between epithelial cells (H. Wang et al. 2015). Direct interaction between IAV and the two transmembrane proteins of the desmosomes (DSG2 and DSC2) has not been identified in the literature. Nonetheless, the upregulation of the desmosomes genes expression during IAV infection might be the attempt of the host to stabilize the cell-cell interaction after the disruption of the extracellular matrix. An in vitro experiment is needed to test this hypothesis.

In contrast, the upregulation of DSP and PKP2 have different outcomes. Several studies have established the importance of the interaction between the bacterial OmpA gene and DSP host desmosomal proteins for an efficient bacterial invasion (Confer and Ayalew 2013; Schweppe et al. 2015). An upregulated DSP gene would increase the efficiency of secondary GAS infection. Moreover, Wang et al. showed the restriction of IAV replication by a binding competition between the host gene PKP2 and the viral protein PB2 for PB1 binding. The competition limited the polymerase activity and reduced the rate of viral replication (Lingyan Wang et al. 2017). Thus, the downregulation of PKP2 results in a positive factor for IAV infection.

In addition, the GAS M1-AP1 and IAV infections shared the membrane-associated mucins (MAM) gene, MUC16, among their list of DEGs. MUC16 is upregulated in IAV influenza (LFC=3.24); whereas, it is downregulated during GAS M1-AP1 infection (LFC=-0.85). The

MAMs act as a physical barrier against bacterial infection and their shredding by S. pneumoniae enzymes increases the likelihood of bacterial infection (Govindarajan et al. 2012).

Figure 5.11 Biological Adhesion expression. Biological adhesion of human host genes during the three infections. The GO:0022610 term was used to identify all genes that are part of the biological adhesion between cells. Then, the differential expressed human genes were selected and analysed. 74 DEGs from all infection processes were found to be part of the cell adhesion molecules genes as discussed by Zhong et al., 2016. The identified groups are ag (primary roles in axonal guidance), fa (primary involvement in focal adhesions), I (information predominant CAM), m (primarily involved in interactions with cell-matrix), and tj (primary involvement in tight junctions). Meanwhile, 252 DEGs from all infections were not found to be CAM genes as discussed by Zhong et al., 2016.

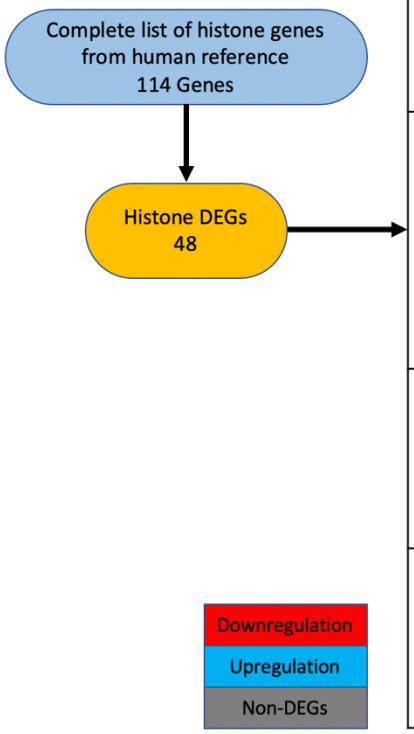


5.5.3.2. IAV infection effect on extracellular histone presence

Histones are evolutionary basic conserved proteins in all eukaryotic cells. They are known for their ability to bind double-strand DNA and to regulate their organization and gene expression. However, the histones are also present in the cell cytoplasm and extracellular fluids where they function as antimicrobial and promote inflammatory response (Hoeksema et al. 2016). GAS is susceptible to the bacteria-killing function of all histones; and, the bacteria protect themselves by the acquisition and activation of plasminogen which degrades all classes of histones (Nitzsche et al. 2016). The transcriptomics analysis from the IAV infection results in the identification of 46 differential expressed (DE) histone genes. The DE histone genes are members of each of the five classes of histone. Furthermore, 45 of the DE histone genes results to be downregulated with H1-0 gene (ENSG00000189060) being the sole member to be upregulated. The GAS M1-AP1 shows one DE histone gene (ENSG00000275126 - H4C13) to be downregulated. Whereas, GAS M49-591 infection shares 36 downregulated DE histone

genes with the IAV infection; whereas, the downregulated H2AZ2 gene (ENSG00000105968) and the upregulated H1-10 (ENSG00000184897) are only present for GAS M49-591 (Figure 5.12. The present study shows the ability of the IAV to create an environment of low concentration of histone proteins following its infection of human epithelial cells. Such an environment permits a favourable subsequent infection of GAS strains due to the diminished anti-microbial activity present in the cytoplasm and extracellular environment.

Figure 5.12 – Histone DEGs Analysis. The table shows the list of histone genes found in the complete list of genes from human reference and their gene expression in the three infections. The following histone genes were not found in the reference: H2BW4P, CENPA, and H3-7. The "Gene" and "Symbol" columns provide the gene Ensembl ID and the gene symbol, respectively. The "Histone Class" column indicates the histone family that the gene is part of. The columns under the "Log Fold Change" column umbrella define the DEG status and the LFC of the histone gene related to the infection (red cell for downregulated, blue cell for upregulated, and grey cell for non-differential-expressed.



			Log Fold Change		
Gene	Symbol	Histone Class	IAV	M49-591	M1-AP1
ENSG00000184357	H1-5	H1	-2.090371975	-0.901169365	-0.509921488
ENSG00000124575	H1-3	H1	-1.242344266	-0.159800907	-0.402898642
ENSG00000168298	H1-4	H1	-1.014826338	-0.065378507	-0.173920661
ENSG00000189060	H1-0	H1	0.88911647	0.723494438	-0.019598212
ENSG00000184897	H1-10	H1	0.457194309	1.004807184	0.166500316
ENSG00000276368	H2AC14	H2A	-2.808407534	-1.113379953	-0.654746305
ENSG00000274997	H2AC12	H2A	-2.348041459	-0.899201918	-0.744584857
ENSG00000278463	H2AC4	H2A	-2.322767108	-0.901468866	-0.79525334
ENSG00000276903	H2AC16	H2A	-2.288026505	-1.173029672	-0.924919277
ENSG00000196747	H2AC13	H2A	-2.104100863	-0.749451581	-0.579478293
ENSG00000275221	H2AC15	H2A	-1.956267134	-1.161875202	-0.994423969
ENSG00000196787	H2AC11	H2A	-1.806230278	-0.61793364	-0.545775546
ENSG00000184270	H2AC21	H2A	-1.754417588	-0.680431371	-0.319396261
ENSG00000277075	H2AC8	H2A	-1.747092574	-0.839764879	-0.686507938
ENSG00000188486	H2AX	H2A	-1.675506093	-0.825887518	-0.255184673
ENSG00000184260	H2AC20	H2A	-1.372512022	-0.33359533	-0.562445734
ENSG00000105968	H2AZ2	H2A	-0.83758445	-0.779611997	-0.187195282
ENSG00000273703	H2BC14	H2B	-3.187794673	-1.649973482	-1.183190294
ENSG00000278588	H2BC10	H2B	-2.427054509	-0.922948102	-0.586767853
ENSG00000276410	H2BC3	H2B	-2.159320922	-1.173883304	-0.575891042
ENSG00000185130	H2BC13	H2B	-1.958882426	-0.864033422	-0.39810866
ENSG00000274641	H2BC17	H2B	-1.795572786	-0.684946126	-0.513482001
ENSG00000275713	H2BC9	H2B	-1.661215159	-1.049428802	-0.678659138
ENSG00000203814	H2BC18	H2B	-1.659546218	-0.870444748	-0.4690096
ENSG00000197903	H2BC12	H2B	-1.388267235	-0.641985663	-0.462925002
ENSG00000273802	H2BC8	H2B	-1.371235215	-0.799000351	-0.646286784
ENSG00000277224	H2BC7	H2B	-1.211997402	-0.869625606	-0.574246538
ENSG00000274290	H2BC6	H2B	-2.077929075	-0.768432013	-0.555280004
ENSG00000124635	H2BC11	H2B	-1.570476782	-0.534123038	-0.567001459
ENSG00000233822	H2BC15	H2B	-1.263808795	-0.398045054	-0.229558909
ENSG00000275714	H3C1	H3	-2.673823631	-1.122059053	-0.529350058
ENSG00000197153	H3C12	H3	-2.441818619	-1.482082675	-0.977624049
ENSG00000286522	H3C2	H3	-2.324723675	-0.699362119	-0.395694111
ENSG00000287080	H3C3	H3	-2.212258434	-0.922839474	-0.577768137
ENSG00000277775	H3C7	H3	-1.942945259	-0.760509026	-0.392287475
ENSG00000273983	H3C8	H3	-1.882962709	-0.717388652	-0.485010202
ENSG00000275379	H3C11	H3	-1.818147183	-0.743179278	-0.632657447
ENSG00000183598	H3C13	H3	-2.629103169	-0.918157059	-0.688257604
ENSG00000278828	H3C10	H3	-1.599863373	-0.718144029	-0.656938746
ENSG00000275126	H4C13	H4	-1.719911069	-1.466932944	-0.86008471
ENSG00000276180	H4C9	H4	-2.618337099	-1.474108905	-0.948238151
ENSG00000278637	H4C1	H4	-2.050568407	-1.317092861	-0.598121425
ENSG00000277157	H4C4	H4	-1.914138582	-0.810572706	-0.383038286
ENSG00000197061	H4C3	H4	-1.679612636	-0.721108753	-0.399686216
ENSG00000278705	H4C2	H4	-1.59953124	-1.253656108	-0.854465337
ENSG00000158406	H4C8	H4	-1.494830722	-0.845615992	-0.394864104
ENSG00000276966	H4C5	H4	-1.184347896	-0.802895632	-0.348475944
ENSG00000197837	H4-16	H4	-1.383627989	-0.440871661	-0.496558078

5.5.4. DR potential on the identification of anti-infective drug against IAV-GAS

The following platforms were selected for DR analysis: L1000CDS2 (Duan et al. 2016b) and Dr Insight version 0.1.1 (Chan et al. 2019). The tools use different statistical approaches to identify potential anti-infection drugs from drug-perturbed gene expression profile datasets:

LINCS for L1000CDS2 and CMAP for Dr Insight (Musa et al. 2018). The list of DEGs from each infection process was used as input for both platforms. The reverse mode of the L1000CDS2 platform was used to find the top 50 drugs capable of reversing the input data. As described by Duan et al, the rank product of the drugs was calculated between the IAV infection rank and the two GAS serotypes infection (Duan et al. 2016b). The rank product was used to identify drugs with the potential of assisting the host cells against the co-infection of IAV and GAS serotypes. Thus, the analysis filter and ranks for drugs that are present in both viral and bacterial infections under study (IAV and S. pyogenes M1; and, IAV and S. pyogenes M49). The drug results from the Dr Insight platform were filtered with a p-value < 0.05. The identified drugs were then analyzed or their marketed status; their species target through the use of biological assays from the PubChem website (S. Kim et al. 2021); and, a drug-gene interaction network was constructed with Stitch version 5.0 (Szklarczyk et al. 2016) between the lists of DEGs and the identified drug (Figure 5.9B). The DR scripts are available at https://github.com/SalemSueto-BioInfo/Dual_RNAseq_Hsapiens_Spyogenes_IAV.

5.6. Conclusion

The present study focus on the infection processes of two M serotypes from GAS and IAV; and, the identification of potential drugs capable of assisting the human host against both pathogens. The three infections showed several differences from each other. These dissimilarities were observed at the transcriptomics level; the rate of the human host cell death; the level of secreted cytokine in the supernatant; and, the strain-specific adherence of GAS to the human host cells. The goal of using the dual RNA-seq technique to study both host and pathogen simultaneously during the infection was not achieved due to the low gene count observed for the pathogens, as observed by the rarefaction curves from the Scotty tool. Therefore, the infection and the DR analysis were studied from the human host perspective. The transcriptomics data provides several indications about the impact of IAV on the secondary GAS infection through its effects on the integrity of cell-cell interaction and the downregulation of histone proteins. The study concludes with the identification of 40 potential drug treatments for either or both pathogens. The identification of several drugs with known anti-influenza and anti-GAS infection shows the potential of DR platforms for the treatment of infectious disease.

Chapter 6 – Conclusion

The aim of the present dissertation was the application of transcriptomics data to study two main biological processes: HGPS and the host-pathogen interaction between human and GAS and IAV. The microarray platform was used to study the effect of the single mutation in the LMNA gene that causes the HGPS disease in human. The disease is known to produce a mutant version of the LMNA protein that has a significant effect on the stability of the entire cellular mechanical structure starting from the nuclear lamina. The study identified known biological processes such as the organization of the nucleus and cell cycle transition; and, unknown ones like the regulation of mRNA and ncRNA transcription. Furthermore, the study does show differences at the transcription level for cellular senescence process caused by HGPS, UV-B light, and telomere elongation. The application of the RNA-seq technology was used to study the host-pathogen interaction study. The analysis highlights the IAV pathogen to have an impact on secondary bacterial infection by GAS by affecting the integrity of the extracellular matrix between epithelial cells; and, the down-regulation of histone genes transcription. Moreover, the drug repurposing analysis from the human gene expression identified 24 potential drugs including drugs with known anti-infective properties such as Anisomycin, Emetine, and LY-294002.

Bibliography

Abrams, Zachary B., Travis S. Johnson, Kun Huang, Philip R.O. Payne, and Kevin Coombes. 2019. “A Protocol to Evaluate RNA Sequencing Normalization Methods.” *BMC Bioinformatics* 20 (S24): 679. https://doi.org/10.1186/s12859-019-3247-x.

Ackermann, Marit, and Korbinian Strimmer. 2009. “A General Modular Framework for Gene Set Enrichment Analysis.” *BMC Bioinformatics* 10 (1): 1–20. https://doi.org/10.1186/1471-2105-10-47/TABLES/8.

Adams, Christopher J., Megan C. Kopp, Natacha Larburu, Piotr R. Nowak, and Maruf M. U. Ali. 2019. “Structure and Molecular Mechanism of ER Stress Signaling by the Unfolded Protein Response Signal Activator IRE1.” *Frontiers in Molecular Biosciences* 6 (March): 11. https://doi.org/10.3389/fmolb.2019.00011.

Agarwal, Shivani, Shivangi Agarwal, Hong Jin, Preeti Pancholi, and Vijay Pancholi. 2012. “Serine/Threonine Phosphatase (SP-STP), Secreted from Streptococcus Pyogenes, Is a Pro-Apoptotic Protein.” *Journal of Biological Chemistry* 287 (12): 9147–67. https://doi.org/10.1074/JBC.M111.316554.

Aksoy, Bülent Arman, Vlado Dancík, Kenneth Smith, Jessica N. Mazerik, Zhou Ji, Benjamin Gross, Olga Nikolova, et al. 2017. “CTD2 Dashboard: A Searchable Web Interface to Connect Validated Results from the Cancer Target Discovery and Development Network.” *Database : The Journal of Biological Databases and Curation* 2017 (January): 54. https://doi.org/10.1093/database/bax054.

Alboukadel Kassambara and Fabian Mundt. 2020. “Factoextra : Extract and Visualize the Results of Multivariate Data Analyses. R Package Version 1.0.7.” 2020.

Almagro Armenteros, José Juan, Konstantinos D. Tsirigos, Casper Kaae Sønderby, Thomas Nordahl Petersen, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2019. “SignalP 5.0 Improves Signal Peptide Predictions Using Deep Neural Networks.” *Nature Biotechnology* 37 (4): 420–23. https://doi.org/10.1038/s41587-019-0036-z.

Amelio, I., M. Gostev, R. A. Knight, A. E. Willis, G. Melino, and A. V. Antonov. 2014. “DRUGSURV: A Resource for Repositioning of Approved and Experimental Drugs in Oncology Based on Patient Survival Information.” *Cell Death and Disease* 5 (2): e1051. https://doi.org/10.1038/cddis.2014.9.

Ames, Sasha K., David A. Hysom, Shea N. Gardner, G. Scott Lloyd, Maya B. Gokhale, and

Bibliography

Jonathan E. Allen. 2013. “Scalable Metagenomic Taxonomy Classification Using a Reference Genome Database.” *Bioinformatics* 29 (18): 2253–60. https://doi.org/10.1093/bioinformatics/btt389.

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. 2015. “HTSeq-A Python Framework to Work with High-Throughput Sequencing Data.” *Bioinformatics* 31 (2): 166–69. https://doi.org/10.1093/bioinformatics/btu638.

Andrews S. 2010. “FastQC: A Quality Control Tool for High Throughput Sequence Data.” 2010.

Armstrong, Jane F., Elena Faccenda, Simon D. Harding, Adam J. Pawson, Christopher Southan, Joanna L. Sharman, Brice Campo, et al. 2020. “The IUPHAR/BPS Guide to PHARMACOLOGY in 2020: Extending Immunopharmacology Content and Introducing the IUPHAR/MMV Guide to MALARIA PHARMACOLOGY.” *Nucleic Acids Research* 48 (D1): D1006–21. https://doi.org/10.1093/nar/gkz951.

Auburn, Richard P., David P. Kreil, Lisa A. Meadows, Bettina Fischer, Santiago Sevillano Matilla, and Steven Russell. 2005. “Robotic Spotting of CDNA and Oligonucleotide Microarrays.” *Trends in Biotechnology* 23 (7): 374–79. https://doi.org/10.1016/J.TIBTECH.2005.04.002.

Auguie, Baptiste. 2017. “GridExtra: Miscellaneous Functions for ‘Grid’ Graphics. R Package Version 2.3.” 2017.

Bailey, Timothy L., Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, and William S. Noble. 2009. “MEME Suite: Tools for Motif Discovery and Searching.” *Nucleic Acids Research* 37 (SUPPL. 2): W202–8. https://doi.org/10.1093/nar/gkp335.

Bainbridge, Matthew N., René L. Warren, Martin Hirst, Tammy Romanuik, Thomas Zeng, Anne Go, Allen Delaney, et al. 2006. “Analysis of the Prostate Cancer Cell Line LNCaP Transcriptome Using a Sequencing-by-Synthesis Approach.” *BMC Genomics* 7 (1): 1–11. https://doi.org/10.1186/1471-2164-7-246/TABLES/5.

Barrell, Daniel, Emily Dimmer, Rachael P. Huntley, David Binns, Claire O’Donovan, and Rolf Apweiler. 2009. “The GOA Database in 2009—an Integrated Gene Ontology Annotation Resource.” *Nucleic Acids Research* 37 (suppl_1): D396–403. https://doi.org/10.1093/NAR/GKN803.

Baruch, Moshe, Ilia Belotserkovsky, Baruch B Hertzog, Miriam Ravins, Eran Dov, Kevin S

Bibliography

McIver, Yoann S Le Breton, et al. 2014. “An Extracellular Bacterial Pathogen Modulates Host Metabolism to Regulate Its Own Sensing and Proliferation.” *Cell* 156 (1–2): 97–108. https://doi.org/10.1016/j.cell.2013.12.007.

Becker, Michael G., Philip L. Walker, Nadège C. Pulgar-Vidal, and Mark F. Belmonte. 2017. “SeqEnrich: A Tool to Predict Transcription Factor Networks from Co-Expressed Arabidopsis and Brassica Napus Gene Sets.” *PLOS ONE* 12 (6): e0178256. https://doi.org/10.1371/JOURNAL.PONE.0178256.

Berkower, Carol, Miriam Ravins, Allon E. Moses, and Emanuel Hanski. 1999. “Expression of Different Group A Streptococcal M Proteins in an Isogenic Background Demonstrates Diversity in Adherence to and Invasion of Eukaryotic Cells.” *Molecular Microbiology* 31 (5): 1463–75. https://doi.org/10.1046/j.1365-2958.1999.01289.x.

Berlin, Konstantin, Sergey Koren, Chen Shan Chin, James P. Drake, Jane M. Landolin, and Adam M. Phillippy. 2015. “Assembling Large Genomes with Single-Molecule Sequencing and Locality-Sensitive Hashing.” *Nature Biotechnology* 33 (6): 623–30. https://doi.org/10.1038/nbt.3238.

Bessen, Debra E, and Sergio Lizano. 2010. “Tissue Tropisms in Group A Streptococcal Infections.” *Future Microbiology* 5 (4): 623–38. https://doi.org/10.2217/fmb.10.28.

Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988]. 1988. “National Center for Biotechnology Information (NCBI).” 1988. https://www.ncbi.nlm.nih.gov/.

Bikkul, Mehmet U., Richard G.A. Faragher, Gemma Worthington, Peter Meinke, Alastair R.W. Kerr, Aakila Sammy, Kumars Riyahi, et al. 2019. “Telomere Elongation through HTERT Immortalization Leads to Chromosome Repositioning in Control Cells and Genomic Instability in Hutchinson-Gilford Progeria Syndrome Fibroblasts, Expressing a Novel SUN1 Isoform.” *Genes, Chromosomes & Cancer* 58 (6): 341. https://doi.org/10.1002/GCC.22711.

Binnicker, Matthew J., Richard D. Williams, and Michael A. Apicella. 2003. “Infection of Human Urethral Epithelium with Neisseria Gonorrhoeae Elicits an Upregulation of Host Anti-Apoptotic Factors and Protects Cells from Staurosporine-Induced Apoptosis.” *Cellular Microbiology* 5 (8): 549–60. https://doi.org/10.1046/j.1462-5822.2003.00300.x.

Blankenberg, Daniel, Assaf Gordon, Gregory Von Kuster, Nathan Coraor, James Taylor,

Bibliography

Anton Nekrutenko, and Galaxy Team. 2010. “Manipulation of FASTQ Data with Galaxy.” *BIOINFORMATICS APPLICATIONS NOTE* 26 (14): 1783–85. https://doi.org/10.1093/bioinformatics/btq281.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. “Near-Optimal Probabilistic RNA-Seq Quantification.” *Nature Biotechnology* 34 (5): 525–27. https://doi.org/10.1038/nbt.3519.

Broad Institute. 2019. “Picard Toolkit.” 2019. https://github.com/broadinstitute/picard.

Brown, Adam S., Sek Won Kong, Isaac S. Kohane, and Chirag J. Patel. 2016. “KsRepo: A Generalized Platform for Computational Drug Repositioning.” *BMC Bioinformatics* 17 (1). https://doi.org/10.1186/s12859-016-0931-y.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2014. “Fast and Sensitive Protein Alignment Using DIAMOND.” *Nature Methods 2014 12:1* 12 (1): 59–60. https://doi.org/10.1038/nmeth.3176.

Buffalo, Vince. 2011. “Scythe - A Bayesian Adapter Trimmer.” 2011.

Bumgarner, Roger. 2013. “DNA Microarrays: Types, Applications and Their Future.” *Curr Protoc Mol Biol*. https://doi.org/10.1002/0471142727.mb2201s101.

Busby, Michele A., Chip Stewart, Chase A. Miller, Krzysztof R. Grzeda, and Gabor T. Marth. 2013a. “Scotty: A Web Tool for Designing RNA-Seq Experiments to Measure Differential Gene Expression.” *Bioinformatics* 29 (5): 656. https://doi.org/10.1093/BIOINFORMATICS/BTT015.

Busby, Michele A, Chip Stewart, Chase A Miller, Krzysztof R Grzeda, and Gabor T Marth. 2013b. “Gene Expression Scotty : A Web Tool for Designing RNA-Seq Experiments to Measure Differential Gene Expression” 29 (5): 656–57. https://doi.org/10.1093/bioinformatics/btt015.

Bushnell, Brian, Rob Egan, Alex Copeland, Brian Foster, Alicia Clum, Hui Sun, Vasanth Singan, et al. 2014. “BBMap : A Fast , Accurate , Splice-Aware Aligner BBMap : A Fast , Accurate , Splice-Aware Aligner,” 1–2. https://doi.org/10.1186/1471-2105-13-238.

C. Voichita, S. Ansari, S. Draghici. 2020. “ROntoTools: R Onto-Tools Suite. R Package Version 2.16.0.” 2020.

Bibliography

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (December): 421. https://doi.org/10.1186/1471-2105-10-421.

Campillos, Monica, Michael Kuhn, Anne Claude Gavin, Lars Juhl Jensen, and Peer Bork. 2008. "Drug Target Identification Using Side-Effect Similarity." *Science* 321 (5886): 263–66. https://doi.org/10.1126/science.1158140.

Carbon, Seth, Amelia Ireland, Christopher J. Mungall, Shengqiang Shu, Brad Marshall, Suzanna Lewis, Jane Lomax, et al. 2009. "AmiGO: Online Access to Ontology and Annotation Data." *Bioinformatics* 25 (2): 288–89. https://doi.org/10.1093/bioinformatics/btn615.

Carlson, Marc. 2016. "Hgu133plus2.Db: Affymetrix Human Genome U133 Plus 2.0 Array Annotation Data (Chip Hgu133plus2). R Package Version 3.2.3." 2016.

Caroli, Jimmy, Giovanni Sorrentino, Mattia Forcato, Giannino Del Sal, and Silvio Bicciato. 2018. "GDA, a Web-Based Tool for Genomics and Drugs Integrated Analysis." *Web Server Issue Published Online* 46. https://doi.org/10.1093/nar/gky434.

Carrella, D., F. Napolitano, R. Rispoli, M. Miglietta, A. Carissimo, L. Cutillo, F. Sirci, F. Gregoretti, and D. Di Bernardo. 2014. "Mantra 2.0: An Online Collaborative Resource for Drug Mode of Action and Repurposing by Network Analysis." *Bioinformatics* 30 (12): 1787–88. https://doi.org/10.1093/bioinformatics/btu058.

Carvalho, Benilton S., and Rafael A. Irizarry. 2010. "A Framework for Oligonucleotide Microarray Preprocessing." *Bioinformatics* 26 (19): 2363–67. https://doi.org/10.1093/BIOINFORMATICS/BTQ431.

Chakravarty, Debyani, Jianjiong Gao, Sarah Phillips, Ritika Kundra, Hongxin Zhang, Jiaojiao Wang, Julia E. Rudolph, et al. 2017. "OncoKB: A Precision Oncology Knowledge Base." *JCO Precision Oncology* 1 (1): 1–16. https://doi.org/10.1200/po.17.00011.

Chan, Jinyan, Xuan Wang, Jacob A Turner, Nicole E Baldwin, and Jinghua Gu. 2019. "Breaking the Paradigm: Dr Insight Empowers Signature-Free, Enhanced Drug Repurposing." Edited by Oliver Stegle. *Bioinformatics* 35 (16): 2818–26. https://doi.org/10.1093/bioinformatics/btz006.

Chang, Wakam, Yuexia Wang, G. W. Gant Luxton, Cecilia Östlund, Howard J. Worman, and Gregg G. Gundersen. 2019. "Imbalanced Nucleocytoskeletal Connections Create Common Polarity Defects in Progeria and Physiological Aging." *Proceedings of the*

National Academy of Sciences of the United States of America 116 (9): 3578–83. https://doi.org/10.1073/PNAS.1809683116/-/DCSUPPLEMENTAL.

Chaussee, Michael S., Heather R. Sandbulte, Margaret J. Schuneman, Frank P. DePaula, Leslie A. Addengast, Evelyn H. Schlenker, and Victor C. Huber. 2011. "Inactivated and Live, Attenuated Influenza Vaccines Protect Mice against Influenza:Streptococcus Pyogenes Super-Infections." *Vaccine* 29 (21): 3773–81. https://doi.org/10.1016/J.VACCINE.2011.03.031.

Chen, Jing, Eric E. Bardes, Bruce J. Aronow, and Anil G. Jegga. 2009. "ToppGene Suite for Gene List Enrichment Analysis and Candidate Gene Prioritization." *Nucleic Acids Research* 37 (SUPPL. 2): W305. https://doi.org/10.1093/nar/gkp427.

Chen, Shi Yi, Zhe Feng, and Xiaolian Yi. 2017. "A General Introduction to Adjustment for Multiple Comparisons." *Journal of Thoracic Disease* 9 (6): 1725–29. https://doi.org/10.21037/jtd.2017.05.34.

Chen, Zi Jie, Wan Ping Wang, Yu Ching Chen, Jing Ya Wang, Wen Hsin Lin, Lin Ai Tai, Gan Guang Liou, Chung Shi Yang, and Ya Hui Chi. 2014. "Dysregulated Interactions between Lamin A and SUN1 Induce Abnormalities in the Nuclear Envelope and Endoplasmic Reticulum in Progeric Laminopathies." *Journal of Cell Science* 127 (8): 1792–1804. https://doi.org/10.1242/JCS.139683/VIDEO-5.

Cheng, Jie, and Lun Yang. 2013. "Comparing Gene Expression Similarity Metrics for Connectivity Map." In *Proceedings - 2013 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2013*, 165–70. https://doi.org/10.1109/BIBM.2013.6732481.

Clark, Neil R., Kevin S. Hu, Axel S. Feldmann, Yan Kou, Edward Y. Chen, Qiaonan Duan, and Avi Ma'ayan. 2014. "The Characteristic Direction: A Geometrical Approach to Identify Differentially Expressed Genes." *BMC Bioinformatics* 15 (1): 79. https://doi.org/10.1186/1471-2105-15-79.

Cobanoglu, Murat Can, Zoltán N. Oltvai, D. Lansing Taylor, and Ivet Bahar. 2015. "BalestraWeb: Efficient Online Evaluation of Drug-Target Interactions." *Bioinformatics* 31 (1): 131–33. https://doi.org/10.1093/bioinformatics/btu599.

Cock, Peter J.A., Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. 2009. "The Sanger FASTQ File Format for Sequences with Quality Scores, and the Solexa/Illumina FASTQ Variants." *Nucleic Acids Research* 38 (6): 1767–71.

https://doi.org/10.1093/nar/gkp1137.

Colquhoun, Rachel M., Michael B. Hall, Leandro Lima, Leah W. Roberts, Kerri M. Malone, Martin Hunt, Brice Letcher, et al. 2021. "Pandora: Nucleotide-Resolution Bacterial Pan-Genomics with Reference Graphs." *Genome Biology* 22 (1): 1–30. https://doi.org/10.1186/S13059-021-02473-1/FIGURES/9.

Conesa, Ana, Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talón, and Montserrat Robles. 2005. "Blast2GO: A Universal Tool for Annotation, Visualization and Analysis in Functional Genomics Research." *Bioinformatics* 21 (18): 3674–76. https://doi.org/10.1093/bioinformatics/bti610.

Conesa, Ana, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michal Wojciech Szcześniak, et al. 2016. "A Survey of Best Practices for RNA-Seq Data Analysis." *Genome Biology*. BioMed Central Ltd. https://doi.org/10.1186/s13059-016-0881-8.

Confer, Anthony W., and Sahlu Ayalew. 2013. "The OmpA Family of Proteins: Roles in Bacterial Pathogenesis and Immunity." *Veterinary Microbiology*. Elsevier. https://doi.org/10.1016/j.vetmic.2012.08.019.

Corley, Susan M., Karen L. MacKenzie, Annemiek Beverdam, Louise F. Roddam, and Marc R. Wilkins. 2017. "Differentially Expressed Genes from RNA-Seq and Functional Enrichment Results Are Affected by the Choice of Single-End versus Paired-End Reads and Stranded versus Non-Stranded Protocols." *BMC Genomics* 18 (1). https://doi.org/10.1186/S12864-017-3797-0.

Costa-Silva, Juliana, Douglas Domingues, and Fabricio Martins Lopes. 2017. "RNA-Seq Differential Expression Analysis: An Extended Review and a Software Tool." Edited by Zhi Wei. *PLOS ONE* 12 (12): e0190152. https://doi.org/10.1371/journal.pone.0190152.

Costa, Valerio, Marianna Aprile, Roberta Esposito, and Alfredo Ciccodicola. 2013. "RNA-Seq and Human Complex Diseases: Recent Accomplishments and Future Perspectives." *European Journal of Human Genetics* 21 (2): 134. https://doi.org/10.1038/EJHG.2012.129.

Cotto, Kelsy C, Alex H Wagner, Yang-Yang Feng, Susanna Kiwala, Adam C Coffman, Gregory Spies, Alex Wollam, Nicholas C Spies, Obi L Griffith, and Malachi Griffith. 2018. "DGIdb 3.0: A Redesign and Expansion of the Drug-Gene Interaction Database." *Nucleic Acids Research* 46 (D1): D1068–73. https://doi.org/10.1093/nar/gkx1143.

Bibliography

Cox, A. 2007. “ELAND: Efficient Local Alignment of Nucleotide Data (Unpublished).” *Unpublished*.

Dai, Xiaofeng, and Li Shen. 2022. “Advances and Trends in Omics Technology Development.” *Frontiers in Medicine* 0 (July): 1546. https://doi.org/10.3389/FMED.2022.911861.

Daniel, Rolf. 2005. “The Metagenomics of Soil.” *Nature Reviews Microbiology*. Nature Publishing Group. https://doi.org/10.1038/nrmicro1160.

Davis, Allan Peter, Cynthia J. Grondin, Robin J. Johnson, Daniela Sciaky, Jolene Wiegers, Thomas C. Wiegers, and Carolyn J. Mattingly. 2021. “Comparative Toxicogenomics Database (CTD): Update 2021.” *Nucleic Acids Research* 49 (D1): D1138–43. https://doi.org/10.1093/nar/gkaa891.

Dechat, Thomas, Takeshi Shimi, Stephen A Adam, Antonio E Rusinol, Douglas A Andres, H Peter Spielmann, Michael S Sinensky, and Robert D Goldman. 2007. “Alterations in Mitosis and Cell Cycle Progression Caused by a Mutant Lamin A Known to Accelerate Human Aging.” *PNAS* 104: 4955–60.

Decker, Michelle L., Elizabeth Chavez, Irma Vulto, and Peter M. Lansdorp. 2009. “Telomere Length in Hutchinson-Gilford Progeria Syndrome.” *Mechanisms of Ageing and Development* 130 (6): 377–83. https://doi.org/10.1016/J.MAD.2009.03.001.

Deluca, David S., Joshua Z. Levin, Andrey Sivachenko, Timothy Fennell, Marc Danie Nazaire, Chris Williams, Michael Reich, Wendy Winckler, and Gad Getz. 2012. “RNA-SeQC: RNA-Seq Metrics for Quality Control and Process Optimization.” *Bioinformatics* 28 (11): 1530. https://doi.org/10.1093/BIOINFORMATICS/BTS196.

Delva, Emmanuella, Dana K. Tucker, and Andrew P. Kowalczyk. 2009. “The Desmosome.” *Cold Spring Harbor Perspectives in Biology* 1 (2). https://doi.org/10.1101/CSHPERSPECT.A002543.

Dennis, Glynn, Brad T. Sherman, Douglas A. Hosack, Jun Yang, Wei Gao, H. Clifford Lane, and Richard A. Lempicki. 2003. “DAVID: Database for Annotation, Visualization, and Integrated Discovery.” *Genome Biology* 4 (5): R60. https://doi.org/10.1186/gb-2003-4-9-r60.

Didion, John P., Marcel Martin, and Francis S. Collins. 2017. “Atropos: Specific, Sensitive, and Speedy Trimming of Sequencing Reads.” *PeerJ* 2017 (8). https://doi.org/10.7717/peerj.3720.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha,

Bibliography

Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics* 29 (1): 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Dobin, Alexander, and Thomas R. Gingeras. 2015. “Mapping RNA-Seq Reads with STAR.” *Current Protocols in Bioinformatics* 51 (1): 11.14.1-11.14.19. https://doi.org/10.1002/0471250953.bi1114s51.

Donner, Yoni, Stéphane Kazmierczak, and Kristen Fortney. 2018. “Drug Repurposing Using Deep Embeddings of Gene Expression Profiles.” *Molecular Pharmaceutics* 15 (10): 4314–25. https://doi.org/10.1021/acs.molpharmaceut.8b00284.

Drugbank. 2022. “DrugBank: Nifedipine.” 2022. https://go.drugbank.com/drugs/DB01115.

Duan, Qiaonan, St Patrick Reid, Neil R. Clark, Zichen Wang, Nicolas F. Fernandez, Andrew D. Rouillard, Ben Readhead, et al. 2016a. “L1000CDS2: LINCS L1000 Characteristic Direction Signatures Search Engine.” *Npj Systems Biology and Applications* 2 (1): 1–12. https://doi.org/10.1038/npjsba.2016.15.

———. 2016b. “L1000CDS2: LINCS L1000 Characteristic Direction Signatures Search Engine.” *Npj Systems Biology and Applications* 2 (1): 1–12. https://doi.org/10.1038/npjsba.2016.15.

Eden, Eran, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. 2009. “GOrilla: A Tool for Discovery and Visualization of Enriched GO Terms in Ranked Gene Lists.” *BMC Bioinformatics* 10 (1): 48. https://doi.org/10.1186/1471-2105-10-48.

Edgar, Robert C. 2004. “MUSCLE: A Multiple Sequence Alignment Method with Reduced Time and Space Complexity.” *BMC Bioinformatics* 5 (1): 113. https://doi.org/10.1186/1471-2105-5-113.

Edgar, Ron, Michael Domrachev, and Alex E. Lash. 2002. “Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository.” *Nucleic Acids Research* 30 (1): 207–10. https://doi.org/10.1093/nar/30.1.207.

Emanuelsson, Olof, Henrik Nielsen, Søren Brunak, and Gunnar Von Heijne. 2000. “Predicting Subcellular Localization of Proteins Based on Their N-Terminal Amino Acid Sequence.” *Journal of Molecular Biology* 300 (4): 1005–16. https://doi.org/10.1006/jmbi.2000.3903.

Everaert, Celine, Manuel Luypaert, Jesper L V Maag, Quek Xiu Cheng, Marcel E Dinger, Jan Hellemans, and Pieter Mestdagh. 2017. “Benchmarking of RNA-Sequencing Analysis

Workflows Using Whole-Transcriptome RT-QPCR Expression Data." https://doi.org/10.1038/s41598-017-01617-3.

Ewels, Philip, Sverker Lundin, and K Max. 2016. "MultiQC : Summarize Analysis Results for Multiple Tools and Samples in a Single Report" 32 (June): 3047–48. https://doi.org/10.1093/bioinformatics/btw354.

Fabbro, Cristian Del, Simone Scalabrin, Michele Morgante, and Federico M. Giorgi. 2013. "An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis." *PLoS ONE* 8 (12): 85024. https://doi.org/10.1371/journal.pone.0085024.

Fafián-labora, Juan A., Miriam Morente-lópez, Fco Javier de Toro, and María C. Arufe. 2021. "High-Throughput Screen Detects Calcium Signaling Dysfunction in Hutchinson-Gilford Progeria Syndrome." *International Journal of Molecular Sciences* 22 (14). https://doi.org/10.3390/IJMS22147327.

Falda, Marco, Stefano Toppo, Alessandro Pescarolo, Enrico Lavezzo, Barbara Di Camillo, Andrea Facchinetti, Elisa Cilia, Riccardo Velasco, and Paolo Fontana. 2012. "Argot2: A Large Scale Function Prediction Tool Relying on Semantic Similarity of Weighted Gene Ontology Terms." *BMC Bioinformatics* 13 (SUPPL.4): S14. https://doi.org/10.1186/1471-2105-13-S4-S14.

Fan, Huan, Anthony R. Ives, Yann Surget-Groba, and Charles H. Cannon. 2015. "An Assembly and Alignment-Free Method of Phylogeny Reconstruction from next-Generation Sequencing Data." *BMC Genomics* 16 (1). https://doi.org/10.1186/s12864-015-1647-5.

Fauci, Anthony S. 2006. "Seasonal and Pandemic Influenza Preparedness: Science and Countermeasures." *The Journal of Infectious Diseases* 194 (Supplement_2): S73–76. https://doi.org/10.1086/507550.

Ferri, Gianmarco, Barbara Storti, and Ranieri Bizzarri. 2017. "Nucleocytoplasmic Transport in Cells with Progerin-Induced Defective Nuclear Lamina." *Biophysical Chemistry* 229 (October): 77–83. https://doi.org/10.1016/J.BPC.2017.06.003.

Fischer, Silke F., Claudia Schwarz, Juliane Vier, and Georg Häcker. 2001. "Characterization of Antiapoptotic Activities of *Chlamydia Pneumoniae* in Human Cells." Edited by D. L. Burns. *Infection and Immunity* 69 (11): 7121–29. https://doi.org/10.1128/IAI.69.11.7121-7129.2001.

Förstner, Konrad U., Jörg Vogel, and Cynthia M. Sharma. 2014. "READemption—a Tool for the Computational Analysis of Deep-Sequencing–Based Transcriptome Data."

Bioinformatics 30 (23): 3421–23. https://doi.org/10.1093/BIOINFORMATICS/BTU533.

Franz, Max, Harold Rodriguez, Christian Lopes, Khalid Zuberi, Jason Montojo, Gary D. Bader, and Quaid Morris. 2018. "GeneMANIA Update 2018." *Nucleic Acids Research* 46 (W1): W60–64. https://doi.org/10.1093/nar/gky311.

Frönicke, Lutz, Denise N. Bronner, Mariana X. Byndloss, Bridget McLaughlin, Andreas J. Bäumler, and Alexander J. Westermann. 2018. "Toward Cell Type-Specific In Vivo Dual RNA-Seq." In *Methods in Enzymology*, 612:505–22. Academic Press Inc. https://doi.org/10.1016/bs.mie.2018.08.013.

Fu, Xing, Ning Fu, Song Guo, Zheng Yan, Ying Xu, Hao Hu, Corinna Menzel, et al. 2009. "Estimating Accuracy of RNA-Seq and Microarrays with Proteomics." *BMC Genomics* 10 (1): 161. https://doi.org/10.1186/1471-2164-10-161.

Gallo, Kathleen, Andrean Goede, Andreas Eckert, Barbara Moahamed, Robert Preissner, and Björn Oliver Gohlke. 2021. "PROMISCUOUS 2.0: A Resource for Drug-Repositioning." *Nucleic Acids Research* 49 (D1): D1373–80. https://doi.org/10.1093/nar/gkaa1061.

Gardner, Shea N., Tom Slezak, and Barry G. Hall. 2015. "KSNP3.0: SNP Detection and Phylogenetic Analysis of Genomes without Genome Alignment or Reference Genome." *Bioinformatics* 31 (17): 2877–78. https://doi.org/10.1093/bioinformatics/btv271.

Ghandi, Mahmoud, Franklin W. Huang, Judit Jané-Valbuena, Gregory V. Kryukov, Christopher C. Lo, E. Robert McDonald, Jordi Barretina, et al. 2019. "Next-Generation Characterization of the Cancer Cell Line Encyclopedia." *Nature* 569 (7757): 503–8. https://doi.org/10.1038/s41586-019-1186-3.

Gilbert, Jack A, and Christopher L Dupont. 2010. "Microbial Metagenomics: Beyond the Genome." https://doi.org/10.1146/annurev-marine-120709-142811.

Gilford, Hastings, and Jonathan Hutchinson. 1897. "On a Condition of Mixed Premature and Immature Development." *Medico-Chirurgical Transactions* 80 (1): 17. https://doi.org/10.1177/095952879708000105.

Gilson, Michael K, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang, and Jenny Chong. 2015. "BindingDB in 2015: A Public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology." *Nucleic Acids Research* 44: 1045–53. https://doi.org/10.1093/nar/gkv1072.

Goldman, Robert D., Dale K. Shumaker, Michael R. Erdos, Maria Eriksson, Anne E. Goldman, Leslie B. Gordon, Yosef Gruenbaum, et al. 2004. "Accumulation of Mutant Lamin A

Causes Progressive Changes in Nuclear Architecture in Hutchinson–Gilford Progeria Syndrome." *Proceedings of the National Academy of Sciences* 101 (24): 8963–68. https://doi.org/10.1073/PNAS.0402943101.

Gordon, Leslie B., Joe Massaro, Ralph B. D'Agostino, Susan E. Campbell, Joan Brazier, W. Ted Brown, Monica E. Kleinman, and Mark W. Kieran. 2014. "Impact of Farnesylation Inhibitors on Survival in Hutchinson-Gilford Progeria Syndrome." *Circulation* 130 (1): 27. https://doi.org/10.1161/CIRCULATIONAHA.113.008285.

Govindarajan, Bharathi, Balaraj B. Menon, Sandra Spurr-Michaud, Komal Rastogi, Michael S. Gilmore, Pablo Argüeso, and Ilene K. Gipson. 2012. "A Metalloproteinase Secreted by Streptococcus Pneumoniae Removes Membrane Mucin MUC16 from the Epithelial Glycocalyx Barrier." *PLoS ONE* 7 (3). https://doi.org/10.1371/journal.pone.0032418.

Grabherr, Manfred G, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, et al. 2013. "Trinity: Reconstructing a Full-Length Transcriptome without a Genome from RNA-Seq Data HHS Public Access Author Manuscript." *Nat Biotechnol* 29 (7): 644–52. https://doi.org/10.1038/nbt.1883.

Griffith, Malachi, Nicholas C Spies, Kilannin Krysiak, Joshua F McMichael, Adam C Coffman, Arpad M Danos, Benjamin J Ainscough, et al. 2017. "CIViC Is a Community Knowledgebase for Expert Crowdsourcing the Clinical Interpretation of Variants in Cancer." *Nature Genetics* 49 (2): 170–74. https://doi.org/10.1038/ng.3774.

Günther, Juliane, and Hans Martin Seyfert. 2018. "The First Line of Defence: Insights into Mechanisms and Relevance of Phagocytosis in Epithelial Cells." *Seminars in Immunopathology* 40 (6): 555. https://doi.org/10.1007/S00281-018-0701-1.

Gupta, Anuj, I. King Jordan, and Lavanya Rishishwar. 2017. "StringMLST: A Fast k-Mer Based Tool for Multilocus Sequence Typing." *Bioinformatics* 33 (1): 119–21. https://doi.org/10.1093/bioinformatics/btw586.

Gupta, Shobhit, John A. Stamatoyannopoulos, Timothy L. Bailey, and William Stafford Noble. 2007. "Quantifying Similarity between Motifs." *Genome Biology* 8 (2): 1–9. https://doi.org/10.1186/GB-2007-8-2-R24/TABLES/3.

Hadley Wickham and Dana Seidel. 2020. "Scales: Scale Functions for Visualization. R Package Version 1.1.1." 2020.

Hannon Lab. 2009. "FASTX-Toolkit: FASTQ/A Short-Reads Pre-Processing Tools." 2009. http://hannonlab.cshl.edu/fastx_toolkit/index.html.

Bibliography

Hardcastle, Thomas J. 2016. “Generalized Empirical Bayesian Methods for Discovery of Differential Data in High-Throughput Biology.” *Bioinformatics* 32 (2): 195–202. https://doi.org/10.1093/bioinformatics/btv569.

Hecker, Nikolai, Jessica Ahmed, Joachim Von Eichborn, Mathias Dunkel, Karel Macha, Andreas Eckert, Michael K Gilson, Philip E Bourne, and Robert Preissner. 2011. “SuperTarget Goes Quantitative: Update on Drug-Target Interactions.” https://doi.org/10.1093/nar/gkr912.

Heller, Michael J. 2003. “DNA Microarray Technology: Devices, Systems, and Applications.” *Http://Dx.Doi.Org/10.1146/Annurev.Bioeng.4.020702.153438* 4 (November): 129–53. https://doi.org/10.1146/ANNUREV.BIOENG.4.020702.153438.

Hernandez, Lidia, Kyle J. Roux, Esther Sook Miin Wong, Leslie C. Mounkes, Rafidah Mutalif, Raju Navasankari, Bina Rai, et al. 2010. “Functional Coupling between the Extracellular Matrix and Nuclear Lamina by Wnt Signaling in Progeria.” *Developmental Cell* 19 (3): 413–25. https://doi.org/10.1016/J.DEVCEL.2010.08.013.

Herrera, Andrea L., Victor C. Huber, and Michael S. Chaussee. 2016. “The Association between Invasive Group A Streptococcal Diseases and Viral Respiratory Tract Infections.” *Frontiers in Microbiology* 7 (MAR): 342. https://doi.org/10.3389/FMICB.2016.00342.

Hoeksema, Marloes, Martin Van Eijk, Henk P. Haagsman, and Kevan L. Hartshorn. 2016. “Histones as Mediators of Host Defense, Inflammation and Thrombosis.” *Future Microbiology* 11 (3): 441. https://doi.org/10.2217/FMB.15.151.

Hoffmann, Steve, Christian Otto, Gero Doose, Andrea Tanzer, David Langenberger, Sabina Christ, Manfred Kunz, et al. 2014. “A Multi-Split Mapping Algorithm for Circular RNA, Splicing, Trans-Splicing and Fusion Detection.” *Genome Biology 2014 15:2* 15 (2): 1–11. https://doi.org/10.1186/GB-2014-15-2-R34.

Hong, Changjin, Solaiappan Manimaran, Ying Shen, Joseph F. Perez-Rogers, Allyson L. Byrd, Eduardo Castro-Nallar, Keith A. Crandall, and William E. Johnson. 2014. “PathoScope 2.0: A Complete Computational Framework for Strain Identification in Environmental or Clinical Sequencing Samples.” *Microbiome* 2 (1): 33. https://doi.org/10.1186/2049-2618-2-33.

Horton, Paul, Keun Joon Park, Takeshi Obayashi, Naoya Fujita, Hajime Harada, C. J. Adams-Collier, and Kenta Nakai. 2007. “WoLF PSORT: Protein Localization Predictor.” *Nucleic*

Acids Research 35 (SUPPL.2): W585. https://doi.org/10.1093/nar/gkm259.

Hrdlickova, Radmila, Masoud Toloue, and Bin Tian. 2016. "RNA-Seq Methods for Transcriptome Analysis." https://doi.org/10.1002/wrna.1364.

Hsu, Li Chung, Thomas Enzler, Jun Seita, Anjuli M. Timmer, Chih Yuan Lee, Ting Yu Lai, Guann Yi Yu, et al. 2011. "IL-1β-Driven Neutrophila Preserves Antibacterial Defense in the Absence of the Kinase IKKβ." *Nature Immunology* 12 (2): 144. https://doi.org/10.1038/NI.1976.

Hu, Zhiliang, Jie Bao, James M Reecy, and Zhi-Liang Hu. 2008. "CateGOrizer: A Web-Based Program to Batch Analyze Gene Ontology Classification Categories CateGOrizer: A Web-Based Program to Batch Analyze Gene On-Tology Classification Categories." *Online Journal of Bioinformatics* 9 (2): 108–12. http://www.animalgenome.org/bioinfo/tools/catego/.

Huang, Li Hong, Qiu Shun He, Ke Liu, Jiao Cheng, Min Dong Zhong, Lin Shan Chen, Li Xia Yao, and Zhi Liang Ji. 2018. "ADReCS-Target: Target Profiles for Aiding Drug Safety Research and Application." *Nucleic Acids Research* 46 (D1): D911–17. https://doi.org/10.1093/nar/gkx899.

Huerta-Cepas, Jaime, Damian Szklarczyk, Kristoffer Forslund, Helen Cook, Davide Heller, Mathias C. Walter, Thomas Rattei, et al. 2016. "EGGNOG 4.5: A Hierarchical Orthology Framework with Improved Functional Annotations for Eukaryotic, Prokaryotic and Viral Sequences." *Nucleic Acids Research* 44 (D1): D286–93. https://doi.org/10.1093/nar/gkv1248.

Huson, Daniel H., Suparna Mitra, Hans Joachim Ruscheweyh, Nico Weber, and Stephan C. Schuster. 2011. "Integrative Analysis of Environmental Sequences Using MEGAN4." *Genome Research* 21 (9): 1552. https://doi.org/10.1101/GR.120618.111.

Hutchinson, Jonathan. 1886. "Congenital Absence of Hair and Mammary Glands with Atrophic Condition of the Skin and Its Appendages, in a Boy Whose Mother Had Been Almost Wholly Bald from Alopecia Areata from the Age of Six." *Medico-Chirurgical Transactions* 69 (1): 473. https://doi.org/10.1177/095952878606900127.

Hüttenhofer, Alexander, Peter Schattner, and Norbert Polacek. 2005. "Non-Coding RNAs: Hope or Hype?" *Trends in Genetics* 21 (5): 289–97. https://doi.org/10.1016/J.TIG.2005.03.007.

Irizarry, Rafael A., Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs, and

Terence P. Speed. 2003. “Summaries of Affymetrix GeneChip Probe Level Data.” *Nucleic Acids Research* 31 (4): e15–e15. https://doi.org/10.1093/NAR/GNG015.

Irizarry, Rafael A, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. 2003. “Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data.” *Biostatistics*. Vol. 4. http://www.biostat.jhsph.edu/.

Iyer, Niranjani. 2022. “Promises and Benefits of Omics Approaches to Data-Driven Science Industries.” *Bioinformatics in Agriculture*, January, 23–36. https://doi.org/10.1016/B978-0-323-89778-5.00031-3.

Jain, Neha, Kathleen F. Mittendorf, Marilyn Holt, Michele Lenoue-Newton, Ian Maurer, Clinton Miller, Matthew Stachowiak, et al. 2020. “The My Cancer Genome Clinical Trial Data Model and Trial Curation Workflow.” *Journal of the American Medical Informatics Association : JAMIA* 27 (7): 1057–66. https://doi.org/10.1093/jamia/ocaa066.

Jarada, Tamer N., Jon G. Rokne, and Reda Alhajj. 2020. “A Review of Computational Drug Repositioning: Strategies, Approaches, Opportunities, Challenges, and Directions.” *Journal of Cheminformatics*. BioMed Central. https://doi.org/10.1186/s13321-020-00450-7.

Jia, Zhilong, Ying Liu, Naiyang Guan, Xiaochen Bo, Zhigang Luo, and Michael R Barnes. 2016. “Cogena, a Novel Tool for Co-Expressed Gene-Set Enrichment Analysis, Applied to Drug Repositioning and Drug Mode of Action Discovery.” https://doi.org/10.1186/s12864-016-2737-8.

Jin, Haijing, Ying Wooi Wan, and Zhandong Liu. 2017. “Comprehensive Evaluation of RNA-Seq Quantification Methods for Linearity.” *BMC Bioinformatics* 18 (S4): 117. https://doi.org/10.1186/s12859-017-1526-y.

John, John St. 2010. “Seqprep.” 2010.

Kanehisa, Minoru, Miho Furumichi, Yoko Sato, Mari Ishiguro-Watanabe, and Mao Tanabe. 2021. “KEGG: Integrating Viruses and Cellular Organisms.” *Nucleic Acids Research* 49 (D1): D545–51. https://doi.org/10.1093/NAR/GKAA970.

Katz, Yarden, Eric T. Wang, Edoardo M. Airoldi, and Christopher B. Burge. 2010. “Analysis and Design of RNA Sequencing Experiments for Identifying Isoform Regulation.” *Nature Methods* 7 (12): 1009–15. https://doi.org/10.1038/nmeth.1528.

Kelley, Lawrence A., Stefans Mezulis, Christopher M. Yates, Mark N. Wass, and Michael J.E.

Sternberg. 2015. “The Phyre2 Web Portal for Protein Modeling, Prediction and Analysis.” *Nature Protocols* 10 (6): 845–58. https://doi.org/10.1038/nprot.2015.053.

Kent, W. James. 2002. “BLAT—The BLAST-Like Alignment Tool.” *Genome Research* 12 (4): 656. https://doi.org/10.1101/GR.229202.

Kilic, Fusun, Marguerite B. Dalton, Sarah K. Burrell, John P. Mayer, Scott D. Patterson, and Michael Sinensky. 1997. “In Vitro Assay and Characterization of the Farnesylation-Dependent Prelamin A Endoprotease.” *Journal of Biological Chemistry* 272 (8): 5298–5304. https://doi.org/10.1074/jbc.272.8.5298.

Kim, Daehwan, Joseph M. Paggi, Chanhee Park, Christopher Bennett, and Steven L. Salzberg. 2019. “Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-Genotype.” *Nature Biotechnology* 37 (8): 907–15. https://doi.org/10.1038/s41587-019-0201-4.

Kim, Daehwan, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. 2013. “TopHat2 : Accurate Alignment of Transcriptomes in the Presence of Insertions , Deletions and Gene Fusions,” 1–13.

Kim, Sunghwan, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, et al. 2021. “PubChem in 2021: New Data Content and Improved Web Interfaces.” *Nucleic Acids Research* 49 (D1): D1388–95. https://doi.org/10.1093/NAR/GKAA971.

Kishore, Shivendra, and Stefan Stamm. 2006. “The SnoRNA HBII-52 Regulates Alternative Splicing of the Serotonin Receptor 2C.” *Science* 311 (5758): 230–32. https://doi.org/10.1126/SCIENCE.1118265/SUPPL_FILE/KISHORE.SOM.REVISED.PDF.

Klassert, Tilman E., Julia Bräuer, Martin Hölzer, Magdalena Stock, Konstantin Riege, Cristina Zubiría-Barrera, Mario M. Müller, et al. 2017. “Differential Effects of Vitamins A and D on the Transcriptional Landscape of Human Monocytes during Infection.” *Scientific Reports* 7 (January). https://doi.org/10.1038/SREP40599.

Klopfenstein, D. V., Liangsheng Zhang, Brent S. Pedersen, Fidel Ramírez, Alex Warwick Vesztrocy, Aurélien Naldi, Christopher J. Mungall, et al. 2018. “GOATOOLS: A Python Library for Gene Ontology Analyses.” *Scientific Reports* 8 (1): 1–17. https://doi.org/10.1038/s41598-018-28948-z.

Köhler, Florian, Felix Bormann, Günter Raddatz, Julian Gutekunst, Samuel Corless, Tanja Musch, Anke S. Lonsdorf, Sylvia Erhardt, Frank Lyko, and Manuel Rodríguez-Paredes. 2020. “Epigenetic Deregulation of Lamina-Associated Domains in Hutchinson-Gilford

Progeria Syndrome." *Genome Medicine* 12 (1): 1–16. https://doi.org/10.1186/S13073-020-00749-Y/FIGURES/6.

Kolberg, Liis, Uku Raudvere, Ivan Kuzmin, Jaak Vilo, and Hedi Peterson. 2020. "Gprofiler2 -- an R Package for Gene List Functional Enrichment Analysis and Namespace Conversion Toolset g:Profiler." *F1000Research* 9 (July): 709. https://doi.org/10.12688/F1000RESEARCH.24956.1.

Kringelum, Jens, Sonny Kim Kjaerulff, Søren Brunak, Ole Lund, Tudor I. Oprea, and Olivier Taboureau. 2016. "ChemProt-3.0: A Global Chemical Biology Diseases Mapping." *Database* 2016. https://doi.org/10.1093/database/bav123.

Krogh, Anders, Björn Larsson, Gunnar Von Heijne, and Erik L.L. Sonnhammer. 2001. "Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes." *Journal of Molecular Biology* 305 (3): 567–80. https://doi.org/10.1006/jmbi.2000.4315.

Krueger, Felix. 2016. "Trim Galore." 2016.

Kucukural, Alper, Onur Yukselen, Deniz M. Ozata, Melissa J. Moore, and Manuel Garber. 2019. "DEBrowser: Interactive Differential Expression Analysis and Visualization Tool for Count Data 06 Biological Sciences 0604 Genetics 08 Information and Computing Sciences 0806 Information Systems." *BMC Genomics* 20 (1): 1–12. https://doi.org/10.1186/S12864-018-5362-X/FIGURES/10.

Kuleshov, Maxim V., Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, et al. 2016. "Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update." *Nucleic Acids Research* 44 (W1): W90–97. https://doi.org/10.1093/nar/gkw377.

Lamb, Justin, Emily D. Crawford, David Peck, Joshua W. Modell, Irene C. Blat, Matthew J. Wrobel, Jim Lerner, et al. 2006. "The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease." *Science* 313 (5795): 1929–35. https://doi.org/10.1126/science.1132939.

Lamb, R. A. 2008. "Influenza." *Encyclopedia of Virology*, January, 95–104. https://doi.org/10.1016/B978-012374410-4.00654-3.

Langfelder, Peter, and Steve Horvath. 2008. "WGCNA: An R Package for Weighted Correlation Network Analysis." *BMC Bioinformatics* 9 (1): 559. https://doi.org/10.1186/1471-2105-9-559.

Bibliography

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Langmead, Ben, Christopher Wilks, Valentin Antonescu, and Rone Charles. 2019. “Scaling Read Aligners to Hundreds of Threads on General-Purpose Processors.” *Bioinformatics* 35 (3): 421–32. https://doi.org/10.1093/BIOINFORMATICS/BTY648.

LaRock, Christopher N, and Victor Nizet. 2015. “Inflammasome/IL-1β Responses to Streptococcal Pathogens.” *Frontiers in Immunology* 6: 518. https://doi.org/10.3389/fimmu.2015.00518.

Law, Charity W, Yunshun Chen, Wei Shi, and Gordon K Smyth. 2014. “Voom : Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts,” 1–17.

Lee, Bernard Kok Bang, Kai Hung Tiong, Jit Kang Chang, Chee Sun Liew, Zainal Ariff Abdul Rahman, Aik Choon Tan, Tsung Fei Khang, and Sok Ching Cheong. 2017. “DeSigN: Connecting Gene Expression with Therapeutics for Drug Repurposing and Development.” *BMC Genomics* 18 (S1): 934. https://doi.org/10.1186/s12864-016-3260-7.

Lee, Ji Hyun, Dae Gyu Kim, Tae Jeong Bae, Kyoohyoung Rho, Ji Tae Kim, Jong Jun Lee, Yeongjun Jang, Byung Cheol Kim, Kyoung Mii Park, and Sunghoon Kim. 2012. “CDA: Combinatorial Drug Discovery Using Transcriptional Response Modules.” *PLoS ONE* 7 (8). https://doi.org/10.1371/journal.pone.0042573.

Leng, Ning, John A Dawson, James A Thomson, Victor Ruotti, Anna I Rissman, Bart M G Smits, Jill D Haag, Michael N Gould, Ron M Stewart, and Christina Kendziorski. 2013. “Gene Expression EBSeq : An Empirical Bayes Hierarchical Model for Inference in RNA-Seq Experiments” 29 (8): 1035–43. https://doi.org/10.1093/bioinformatics/btt087.

Li, Bo, and Colin N Dewey. 2011. “RSEM : Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome.” https://doi.org/10.1186/1471-2105-12-323.

Li, Heng. 2016. “Minimap and Miniasm: Fast Mapping and de Novo Assembly for Noisy Long Sequences.” *Bioinformatics* 32 (14): 2103–10. https://doi.org/10.1093/bioinformatics/btw152.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.

Bibliography

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Li, Jun, and Robert Tibshirani. 2013. “Finding Consistent Patterns: A Nonparametric Approach for Identifying Differential Expression in RNA-Seq Data.” *Statistical Methods in Medical Research* 22 (5): 519–36. https://doi.org/10.1177/0962280211428386.

Li, Na, Marcus Parrish, Tze Khee Chan, Lu Yin, Prashant Rai, Yamada Yoshiyuki, Nona Abolhassani, et al. 2015. “Influenza Infection Induces Host DNA Damage and Dynamic DNA Damage Responses during Tissue Regeneration.” *Cellular and Molecular Life Sciences : CMLS* 72 (15): 2973. https://doi.org/10.1007/S00018-015-1879-1.

Li, Ruiqiang, Chang Yu, Yingrui Li, Tak-wah Lam, Siu-ming Yiu, Karsten Kristiansen, and Jun Wang. 2009. “SOAP2 : An Improved Ultrafast Tool for Short Read Alignment” 25 (15): 1966–67. https://doi.org/10.1093/bioinformatics/btp336.

Li, Weizhong, and Adam Godzik. 2006. “Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences.” *Bioinformatics* 22 (13): 1658–59. https://doi.org/10.1093/bioinformatics/btl158.

Li, Yanhui, Gang Zhou, Ivone G. Bruno, Ning Zhang, Sei Sho, Enzo Tedone, Tsung Po Lai, John P. Cooke, and Jerry W. Shay. 2019. “Transient Introduction of Human Telomerase MRNA Improves Hallmarks of Progeria Cells.” *Aging Cell* 18 (4). https://doi.org/10.1111/ACEL.12979.

Li, You, Tayla B. Heavican, Neetha N. Vellichirammal, Javeed Iqbal, and Chittibabu Guda. 2017. “ChimeRScope: A Novel Alignment-Free Algorithm for Fusion Transcript Prediction Using Paired-End RNA-Seq Data.” *Nucleic Acids Research* 45 (13). https://doi.org/10.1093/nar/gkx315.

Liao, Yang, and Wei Shi. 2020. “Read Trimming Is Not Required for Mapping and Quantification of RNA-Seq Reads at the Gene Level.” *NAR Genomics and Bioinformatics* 2 (3). https://doi.org/10.1093/nargab/lqaa068.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. “FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics* 30 (7): 923–30. https://doi.org/10.1093/bioinformatics/btt656.

Liao, Yang, Gordon K Smyth, and Wei Shi. 2013. “The Subread Aligner : Fast , Accurate and

Scalable Read Mapping by Seed-and-Vote" 41 (10). https://doi.org/10.1093/nar/gkt214.

Liu, Yuwen, Jie Zhou, and Kevin P. White. 2014. "RNA-Seq Differential Expression Studies: More Sequence or More Replication?" *Bioinformatics* 30 (3): 301. https://doi.org/10.1093/BIOINFORMATICS/BTT688.

Liu, Zhihai, Minyi Su, Li Han, Jie Liu, Qifan Yang, Yan Li, and Renxiao Wang. 2017. "Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions." *Accounts of Chemical Research* 50 (2): 302–9. https://doi.org/10.1021/acs.accounts.6b00491.

Lockhart, David J., Helin Dong, Michael C. Byrne, Maximillian T. Follettie, Michael V. Gallo, Mark S. Chee, Michael Mittmann, et al. 1996. "Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays." *Nature Biotechnology 1996 14:13* 14 (13): 1675–80. https://doi.org/10.1038/nbt1296-1675.

Lotfi Shahreza, Maryam, Nasser Ghadiri, Sayed Rasoul Mousavi, Jaleh Varshosaz, and James R Green. 2018. "A Review of Network-Based Approaches to Drug Repositioning." *Briefings in Bioinformatics* 19 (5): 878–92. https://doi.org/10.1093/bib/bbx017.

Louhimo, Riku, Marko Laakso, Denis Belitskin, Juha Klefström, Rainer Lehtonen, and Sampsa Hautaniemi. 2016. "Data Integration to Prioritize Drugs Using Genomics and Curated Data." *BioData Mining* 9 (1). https://doi.org/10.1186/s13040-016-0097-1.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12). https://doi.org/10.1186/s13059-014-0550-8.

Lowe, Rohan, Neil Shirley, Mark Bleackley, Stephen Dolan, and Thomas Shafee. 2017. "Transcriptomics Technologies." *PLoS Computational Biology* 13 (5). https://doi.org/10.1371/journal.pcbi.1005457.

Luo, Weijun, and Cory Brouwer. 2013. "Pathview: An R/Bioconductor Package for Pathway-Based Data Integration and Visualization." *Bioinformatics* 29 (14): 1830–31. https://doi.org/10.1093/BIOINFORMATICS/BTT285.

MacDonald, J. W. 2022. "Affycoretools: Functions Useful for Those Doing Repetitive Analyses with Affymetrix GeneChips." 2022.

Machiels, Barbie M., Antoine H.G. Zorenc, Jorike M. Endert, Helma J.H. Kuijpers, Guillaume J.J.M. Van Eys, Frans C.S. Ramaekers, and Jos L.V. Broers. 1996. "An Alternative Splicing Product of the Lamin A/C Gene Lacks Exon 10." *Journal of Biological Chemistry* 271

(16): 9249–53. https://doi.org/10.1074/jbc.271.16.9249.

Maere, Steven, Karel Heymans, and Martin Kuiper. 2005. "BiNGO: A Cytoscape Plugin to Assess Overrepresentation of Gene Ontology Categories in Biological Networks." *Bioinformatics* 21 (16): 3448–49. https://doi.org/10.1093/bioinformatics/bti551.

Mallick, Himel, Ali Rahnavard, Lauren J. McIver, Siyuan Ma, Yancong Zhang, Long H. Nguyen, Timothy L. Tickle, et al. 2021. "Multivariable Association Discovery in Population-Scale Meta-Omics Studies." *PLOS Computational Biology* 17 (11): e1009442. https://doi.org/10.1371/JOURNAL.PCBI.1009442.

Marioni, John C., Christopher E. Mason, Shrikant M. Mane, Matthew Stephens, and Yoav Gilad. 2008. "RNA-Seq: An Assessment of Technical Reproducibility and Comparison with Gene Expression Arrays." *Genome Research* 18 (9): 1509–17. https://doi.org/10.1101/gr.079558.108.

Martin, Diana R., and Leigh A. Single. 1993. "Molecular Epidemiology of Group A Streptococcus M Type 1 Infections." *The Journal of Infectious Diseases* 167 (5): 1112–17. https://doi.org/10.1093/INFDIS/167.5.1112.

Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." *EMBnet.Journal* 17 (1): 10. https://doi.org/10.14806/ej.17.1.200.

Massari, Paola, Carol A. King, Alan Yu Ho, and Lee M. Wetzler. 2003. "Neisserial PorB Is Translocated to the Mitochondria of HeLa Cells Infected with Neisseria Meningitidis and Protects Cells from Apoptosis." *Cellular Microbiology* 5 (2): 99–109. https://doi.org/10.1046/j.1462-5822.2003.00257.x.

McClintock, Dayle, Leslie B. Gordon, and Karima Djabali. 2006. "Hutchinson-Gilford Progeria Mutant Lamin A Primarily Targets Human Vascular Cells as Detected by an Anti-Lamin A G608G Antibody." *Proceedings of the National Academy of Sciences of the United States of America* 103 (7): 2154–59. https://doi.org/10.1073/PNAS.0511133103/SUPPL_FILE/11133FIG6.PDF.

McHugh, Domhnall, and Jesús Gil. 2018. "Senescence and Aging: Causes, Consequences, and Therapeutic Avenues." *Journal of Cell Biology* 217 (1): 65–77. https://doi.org/10.1083/JCB.201708092.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." *Genome Research*

20 (9): 1297–1303. https://doi.org/10.1101/gr.107524.110.

Mendez, David, Anna Gaulton, A. Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, et al. 2019. "ChEMBL: Towards Direct Deposition of Bioassay Data." *Nucleic Acids Research* 47 (D1): D930–40. https://doi.org/10.1093/nar/gky1075.

Merideth, Melissa A., Leslie B. Gordon, Sarah Clauss, Vandana Sachdev, Ann C.M. Smith, Monique B. Perry, Carmen C. Brewer, et al. 2008. "Phenotype and Course of Hutchinson–Gilford Progeria Syndrome." *Https://Doi.Org/10.1056/NEJMoa0706898* 358 (6): 592–604. https://doi.org/10.1056/NEJMOA0706898.

Metzgar, David, and Antonella Zampolli. 2011. "The M Protein of Group A Streptococcus Is a Key Virulence Factor and a Clinically Relevant Strain Identification Marker." https://doi.org/10.4161/viru.2.5.16342.

Mgbemena, Victoria, Jesus A. Segovia, Te-Hung Chang, Su-Yu Tsai, Garry T. Cole, Chiung-Yu Hung, and Santanu Bose. 2012. "Transactivation of Inducible Nitric Oxide Synthase Gene by Kruppel-like Factor 6 Regulates Apoptosis during Influenza A Virus Infection." *The Journal of Immunology* 189 (2): 606–15. https://doi.org/10.4049/jimmunol.1102742.

Morgulis, Aleksandr, George Coulouris, Yan Raytselis, Thomas L. Madden, Richa Agarwala, and Alejandro A. Schäffer. 2008. "Database Indexing for Production MegaBLAST Searches." In *Bioinformatics*, 24:1757–64. Oxford Academic. https://doi.org/10.1093/bioinformatics/btn322.

Moriya, Yuki, Masumi Itoh, Shujiro Okuda, Akiyasu C. Yoshizawa, and Minoru Kanehisa. 2007. "KAAS: An Automatic Genome Annotation and Pathway Reconstruction Server." *Nucleic Acids Research* 35 (SUPPL.2): W182. https://doi.org/10.1093/nar/gkm321.

Moskovskich, Anna, Ulrich Goldmann, Felix Kartnig, Sabrina Lindinger, Justyna Konecka, Giuseppe Fiume, Enrico Girardi, and Giulio Superti-Furga. 2019. "The Transporters SLC35A1 and SLC30A1 Play Opposite Roles in Cell Survival upon VSV Virus Infection." *Scientific Reports* 9 (1): 1–11. https://doi.org/10.1038/s41598-019-46952-9.

Muñoz-Espín, Daniel, Marta Cañamero, Antonio Maraver, Gonzalo Gómez-López, Julio Contreras, Silvia Murillo-Cuesta, Alfonso Rodríguez-Baeza, et al. 2013. "Programmed Cell Senescence during Mammalian Embryonic Development." *Cell* 155 (5): 1104–18. https://doi.org/10.1016/J.CELL.2013.10.019.

Musa, Aliyu, Laleh Soltan Ghoraie, Shu Dong Zhang, Galina Glazko, Olli Yli-Harja, Matthias

Bibliography

Dehmer, Benjamin Haibe-Kains, and Frank Emmert-Streib. 2018. “A Review of Connectivity Map and Computational Approaches in Pharmacogenomics.” *Briefings in Bioinformatics* 19 (3): 506–23. https://doi.org/10.1093/bib/bbw112.

Musich, Phillip R., and Yue Zou. 2009. “Genomic Instability and DNA Damage Responses in Progeria Arising from Defective Maturation of Prelamin A.” *Aging (Albany NY)* 1 (1): 28. https://doi.org/10.18632/AGING.100012.

Musser, James M., Vivek Kapur, Sagarika Kanjilal, Uma Shah, Daniel M. Musher, Neil L. Barg, Kenneth H. Johnston, et al. 1993. “Geographic and Temporal Distribution and Molecular Characterization of Two Highly Pathogenic Clones of *Streptococcus Pyogenes* Expressing Allelic Variants of Pyrogenic Exotoxin A (Scarlet Fever Toxin).” *Journal of Infectious Diseases* 167 (2): 337–46. https://doi.org/10.1093/infdis/167.2.337.

Nagalakshmi, Ugrappa, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. 2008. “The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing.” *Science* 320 (5881): 1344–49. https://doi.org/10.1126/SCIENCE.1158441/SUPPL_FILE/1158441_TABLES_S2_TO_S6.ZIP.

Najoshi, Joshi, and Joseph Fass. 2011. “Sickle: A Sliding-Window, Adaptive, Quality-Based Trimming Tool for FastQ Files (Version 1.33) [Software].” 2011. https://github.com/najoshi/sickle.

Napolitano, Francesco, Diego Carrella, Barbara Mandriani, Sandra Pisonero-Vaquero, Francesco Sirci, Diego L Medina, Nicola Brunetti-Pierri, and Diego di Bernardo. 2018. “Gene2drug: A Computational Tool for Pathway-Based Rational Drug Repositioning.” Edited by Jonathan Wren. *Bioinformatics* 34 (9): 1498–1505. https://doi.org/10.1093/bioinformatics/btx800.

Nguyen, Dac-Trung, Stephen Mathias, Cristian Bologa, Soren Brunak, Nicolas Fernandez, Anna Gaulton, Anne Hersey, et al. 2016. “Pharos: Collating Protein Information to Shed Light on the Druggable Genome.” *Nucleic Acids Research* 45: 995–1002. https://doi.org/10.1093/nar/gkw1072.

Nita-Lazar, Mihai, Aditi Banerjee, Chiguang Feng, Mohammed N. Amin, Matthew B. Frieman, Wilbur H. Chen, Alan S. Cross, Lai Xi Wang, and Gerardo R. Vasta. 2015. “Desialylation of Airway Epithelial Cells during Influenza Virus Infection Enhances Pneumococcal Adhesion via Galectin Binding.” *Molecular Immunology* 65 (1): 1–16.

https://doi.org/10.1016/j.molimm.2014.12.010.

Nitzsche, Ramona, Juliane Köhler, Bernd Kreikemeyer, and Sonja Oehmcke-Hecht. 2016. "Streptococcus Pyogenes Escapes Killing from Extracellular Histones through Plasminogen Binding and Activation by Streptokinase." *Journal of Innate Immunity* 8 (6): 589. https://doi.org/10.1159/000448039.

Nix, David A., Samir J. Courdy, and Kenneth M. Boucher. 2008. "Empirical Methods for Controlling False Positives and Estimating Confidence in ChIP-Seq Peaks." *BMC Bioinformatics* 9 (1): 1–9. https://doi.org/10.1186/1471-2105-9-523/FIGURES/4.

Noller, Harry F. 2003. "RIBOSOMAL RNA AND TRANSLATION." *Https://Doi.Org/10.1146/Annurev.Bi.60.070191.001203* 60 (November): 191–227. https://doi.org/10.1146/ANNUREV.BI.60.070191.001203.

Nottingham, Ryan M., Douglas C. Wu, Yidan Qin, Jun Yao, Scott Hunicke-Smith, and Alan M. Lambowitz. 2016. "RNA-Seq of Human Reference RNA Samples Using a Thermostable Group II Intron Reverse Transcriptase." *RNA* 22 (4): 597–613. https://doi.org/10.1261/rna.055558.115.

Ochoa, David, Andrew Hercules, Miguel Carmona, Daniel Suveges, Asier Gonzalez-Uriarte, Cinzia Malangone, Alfredo Miranda, et al. 2021. "Open Targets Platform: Supporting Systematic Drug–Target Identification and Prioritisation." *Nucleic Acids Research* 49 (D1): D1302–10. https://doi.org/10.1093/NAR/GKAA1027.

Okamoto, Shigefumi, Shigetada Kawabata, Ichiro Nakagawa, Yoshinobu Okuno, Toshiyuki Goto, Kouichi Sano, and Shigeyuki Hamada. 2003. " Influenza A Virus-Infected Hosts Boost an Invasive Type of Streptococcus Pyogenes Infection in Mice ." *Journal of Virology* 77 (7): 4104–12. https://doi.org/10.1128/JVI.77.7.4104-4112.2003/ASSET/F74A509A-3DD3-495C-A071-727691B576E4/ASSETS/GRAPHIC/JV0731859008.JPEG.

Okonechnikov, Konstantin, Ana Conesa, and Fernando García-Alcalde. 2016. "Qualimap 2: Advanced Multi-Sample Quality Control for High-Throughput Sequencing Data." *Bioinformatics* 32 (2): 292–94. https://doi.org/10.1093/bioinformatics/btv566.

Orchard, Sandra, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H. Campbell, et al. 2014. "The MIntAct Project - IntAct as a Common Curation Platform for 11 Molecular Interaction Databases." *Nucleic Acids Research* 42 (D1): D358–63. https://doi.org/10.1093/nar/gkt1115.

Bibliography

Oughtred, Rose, Chris Stark, Bobby Joe Breitkreutz, Jennifer Rust, Lorrie Boucher, Christie Chang, Nadine Kolas, et al. 2019. "The BioGRID Interaction Database: 2019 Update." *Nucleic Acids Research* 47 (D1): D529–41. https://doi.org/10.1093/nar/gky1079.

Ounit, Rachid, and Stefano Lonardi. 2016. "Higher Classification Sensitivity of Short Metagenomic Reads with CLARK-S." *Bioinformatics* 32 (24): 3823–25. https://doi.org/10.1093/bioinformatics/btw542.

Ozsolak, Fatih, and Patrice M. Milos. 2011. "RNA Sequencing: Advances, Challenges and Opportunities." *Nature Reviews. Genetics* 12 (2): 87. https://doi.org/10.1038/NRG2934.

Pajuste, Fanny Dhelia, Lauris Kaplinski, Märt Möls, Tarmo Puurand, Maarja Lepamets, and Maido Remm. 2017. "FastGT: An Alignment-Free Method for Calling Common SNVs Directly from Raw Sequencing Reads." *Scientific Reports* 7 (1). https://doi.org/10.1038/s41598-017-02487-5.

Park, Peter J. 2009. "ChIP–Seq: Advantages and Challenges of a Maturing Technology." *Nature Reviews Genetics 2009 10:10* 10 (10): 669–80. https://doi.org/10.1038/nrg2641.

Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. "Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression." *Nature Methods* 14 (4): 417–19. https://doi.org/10.1038/nmeth.4197.

Patro, Robert. 2020. "Salmon: Fast, Accurate and Bias-Aware Transcript Quantification from RNA-Seq Data. - Frequently Asked Questions." 2020. https://combine-lab.github.io/salmon/faq/.

Petrov, Anton, and Soheil Shams. 2004. "Microarray Image Processing and Quality Control." *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology 2004 38:3* 38 (3): 211–26. https://doi.org/10.1023/B:VLSI.0000042488.08307.AD.

Peyvandipour, Azam, Nafiseh Saberian, Adib Shafi, Michele Donato, Sorin Draghici, and Alfonso Valencia. 2018. "A Novel Computational Approach for Drug Repurposing Using Systems Biology." https://doi.org/10.1093/bioinformatics/bty133.

Phan, Tamara, Fatima Khalid, and Sebastian Iben. 2019. "Nucleolar and Ribosomal Dysfunction—A Common Pathomechanism in Childhood Progerias?" *Cells* 8 (6). https://doi.org/10.3390/CELLS8060534.

Piétu, Geneviève, Régine Mariage-Samson, Nicole Adeline Fayein, Christiane Matingou, Eric

Eveno, Rémi Houlgatte, Charles Decraene, et al. 1999. “The Genexpress IMAGE Knowledge Base of the Human Brain Transcriptome: A Prototype Integrated Resource for Functional and Computational Genomics.” *Genome Research* 9 (2): 195–209. https://doi.org/10.1101/GR.9.2.195.

Pilarczyk, Marcin, Mehdi Fazel-Najafabadi, Michal Kouril, Behrouz Shamsaei, Juozas Vasiliauskas, Wen Niu, Naim Mahi, et al. 2022. “Connecting Omics Signatures and Revealing Biological Mechanisms with ILINCS.” *Nature Communications 2022 13:1* 13 (1): 1–13. https://doi.org/10.1038/s41467-022-32205-3.

Pilarczyk, Marcin, Mehdi Fazel Najafabadi, Michal Kouril, Juozas Vasiliauskas, Wen Niu, Behrouz Shamsaei, Naim Mahi, et al. 2019. “Connecting Omics Signatures of Diseases, Drugs, and Mechanisms of Actions with ILINCS.” *BioRxiv*. bioRxiv. https://doi.org/10.1101/826271.

Platzer, Alexander, Julia Polzin, Klaus Rembart, Ping Penny Han, Denise Rauer, and Thomas Nussbaumer. 2018. “BioSankey: Visualization of Microbial Communities Over Time.” *Journal of Integrative Bioinformatics* 15 (4). https://doi.org/10.1515/JIB-2017-0063.

Priebe, Steffen, Jörg Linde, Daniela Albrecht, Reinhard Guthke, and Axel A. Brakhage. 2011. “FungiFun: A Web-Based Application for Functional Categorization of Fungal Genes and Proteins.” *Fungal Genetics and Biology* 48 (4): 353–58. https://doi.org/10.1016/j.fgb.2010.11.001.

PubChem. 2021. “Emetine.” 2021. https://pubchem.ncbi.nlm.nih.gov/compound/10219.

———. 2022. “Estradiol.” 2022. https://pubchem.ncbi.nlm.nih.gov/compound/5757.

Pushpakom, Sudeep, Francesco Iorio, Patrick A. Eyers, K. Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, et al. 2018. “Drug Repurposing: Progress, Challenges and Recommendations.” *Nature Reviews Drug Discovery*. Nature Publishing Group. https://doi.org/10.1038/nrd.2018.168.

Qin, Junjie, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, et al. 2010. “A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing.” *Nature* 464 (7285): 59–65. https://doi.org/10.1038/nature08821.

Rahmani, Elior, Reut Yedidim, Liat Shenhav, Regev Schweiger, Omer Weissbrod, Noah Zaitlen, and Eran Halperin. 2017. “GLINT: A User-Friendly Toolset for the Analysis of High-Throughput DNA-Methylation Array Data.” *Bioinformatics* 33 (12): 1870–72.

https://doi.org/10.1093/bioinformatics/btx059.

Raudvere, Uku, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. 2019. “G:Profiler: A Web Server for Functional Enrichment Analysis and Conversions of Gene Lists (2019 Update).” *Nucleic Acids Research* 47 (W1): W191–98. https://doi.org/10.1093/nar/gkz369.

Reglinski, Mark, and Shiranee Sriskandan. 2015. “Streptococcus Pyogenes.” *Molecular Medical Microbiology*, January, 675–716. https://doi.org/10.1016/B978-0-12-397169-2.00038-X.

Reich, Michael, Ted Liefeld, Joshua Gould, Jim Lerner, Pablo Tamayo, and Jill P. Mesirov. 2006. “GenePattern 2.0.” *Nature Genetics* 38 (5): 500–501. https://doi.org/10.1038/NG0506-500.

Ren, Jie, Kai Song, Minghua Deng, Gesine Reinert, Charles H. Cannon, and Fengzhu Sun. 2016. “Inference of Markovian Properties of Molecular Sequences from NGS Data and Applications to Comparative Genomics.” *Bioinformatics* 32 (7): 993–1000. https://doi.org/10.1093/bioinformatics/btv395.

Riege, Konstantin, Martin Hölzer, Tilman E. Klassert, Emanuel Barth, Julia Bräuer, Maximilian Collatz, Franziska Hufsky, et al. 2017. “Massive Effect on LncRNAs in Human Monocytes During Fungal and Bacterial Infections and in Response to Vitamins A and D.” *Scientific Reports* 7 (January). https://doi.org/10.1038/SREP40598.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. “Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies.” *Nucleic Acids Research* 43 (7): e47. https://doi.org/10.1093/nar/gkv007.

Rivera-Torres, José, Rebeca Acín-Perez, Pablo Cabezas-Sánchez, Fernando G. Osorio, Cristina Gonzalez-Gómez, Diego Megias, Carmen Cámara, et al. 2013. “Identification of Mitochondrial Dysfunction in Hutchinson–Gilford Progeria Syndrome through Use of Stable Isotope Labeling with Amino Acids in Cell Culture.” *Journal of Proteomics* 91 (October): 466–77. https://doi.org/10.1016/J.JPROT.2013.08.008.

Robinson, James T., Helga Thorvaldsdóttir, Aaron M. Wenger, Ahmet Zehir, and Jill P. Mesirov. 2017. “Variant Review with the Integrative Genomics Viewer.” *Cancer Research*. American Association for Cancer Research Inc. https://doi.org/10.1158/0008-5472.CAN-17-0337.

Bibliography

Robinson, Mark D, Davis J Mccarthy, and Gordon K Smyth. 2010. “EdgeR : A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data” 26 (1): 139–40. https://doi.org/10.1093/bioinformatics/btp616.

Roosaare, Märt, Mihkel Vaher, Lauris Kaplinski, Märt Möls, Reidar Andreson, Maarja Lepamets, Triinu Kõressaar, Paul Naaber, Siiri Kõljalg, and Maido Remm. 2017. “StrainSeeker: Fast Identification of Bacterial Strains from Raw Sequencing Reads Using User-Provided Guide Trees.” *PeerJ* 2017 (5): e3353. https://doi.org/10.7717/peerj.3353.

Roth, Bryan L., Estelle Lopez, Shamil Patel, and Wesley K. Kroeze. 2000. “The Multiplicity of Serotonin Receptors: Uselessly Diverse Molecules or an Embarrassment of Riches?” *Neuroscientist*. SAGE Publications Inc. https://doi.org/10.1177/107385840000600408.

Rudewicz, Justine, Hayssam Soueidan, Raluca Uricaru, Hervé Bonnefoi, Richard Iggo, Jonas Bergh, and Macha Nikolski. 2016. “MICADo - Looking for Mutations in Targeted PacBio Cancer Data: An Alignment-Free Method.” *Frontiers in Genetics* 7 (DEC). https://doi.org/10.3389/fgene.2016.00214.

Ryan, Patricia A., and Barbara Juncosa. 2016. *Group A Streptococcal Adherence*. *Streptococcus Pyogenes: Basic Biology to Clinical Manifestations*. University of Oklahoma Health Sciences Center.

Safran, Marilyn, Irina Dalah, Justin Alexander, Naomi Rosen, Tsippi Iny Stein, Michael Shmoish, Noam Nativ, et al. 2010. “GeneCards Version 3: The Human Gene Integrator.” *Database : The Journal of Biological Databases and Curation* 2010. https://doi.org/10.1093/database/baq020.

Sahraeian, Sayed Mohammad Ebrahim, Marghoob Mohiyuddin, Robert Sebra, Hagen Tilgner, Pegah T. Afshar, Kin Fai Au, Narges Bani Asadi, et al. 2017. “Gaining Comprehensive Biological Insight into the Transcriptome by Performing a Broad-Spectrum RNA-Seq Analysis.” *Nature Communications* 8 (1): 1–15. https://doi.org/10.1038/s41467-017-00050-4.

Sam, Elizabeth, and Prashanth Athri. 2019. “Web-Based Drug Repurposing Tools: A Survey.” *Briefings in Bioinformatics*. Oxford University Press. https://doi.org/10.1093/bib/bbx125.

Sayols, Sergi. 2020. “Rrvgo: A Bioconductor Package to Reduce and Visualize Gene Ontology Terms.” 2020. https://ssayols.github.io/rrvgo/.

Bibliography

Scaffidi, Paola, and Tom Misteli. 2008. “Lamin A-Dependent Misregulation of Adult Stem Cells Associated with Accelerated Ageing.” *Nature Cell Biology* 10 (4): 452. https://doi.org/10.1038/NCB1708.

Schmieder, Robert, Robert Edwards, and Alex Bateman. 2011. “Quality Control and Preprocessing of Metagenomic Datasets.” *BIOINFORMATICS APPLICATIONS NOTE* 27 (6): 863–64. https://doi.org/10.1093/bioinformatics/btr026.

Schulz, Marcel H., Daniel R. Zerbino, Martin Vingron, and Ewan Birney. 2012. “Oases: Robust de Novo RNA-Seq Assembly across the Dynamic Range of Expression Levels.” *Bioinformatics* 28 (8): 1086–92. https://doi.org/10.1093/bioinformatics/bts094.

Schwacke, Rainer, Gabriel Y. Ponce-Soto, Kirsten Krause, Anthony M. Bolger, Borjana Arsova, Asis Hallab, Kristina Gruden, Mark Stitt, Marie E. Bolger, and Björn Usadel. 2019. “MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis.” *Molecular Plant* 12 (6): 879–92. https://doi.org/10.1016/j.molp.2019.01.003.

Schweppe, Devin K., Christopher Harding, Juan D. Chavez, Xia Wu, Elizabeth Ramage, Pradeep K. Singh, Colin Manoil, and James E. Bruce. 2015. “Host-Microbe Protein Interactions during Bacterial Infection.” *Chemistry and Biology* 22 (11): 1521–30. https://doi.org/10.1016/j.chembiol.2015.09.015.

Sedlazeck, Fritz J, Philipp Rescheneder, and Arndt Von Haeseler. 2013. “NextGenMap : Fast and Accurate Read Mapping in Highly Polymorphic Genomes” 29 (21): 2790–91. https://doi.org/10.1093/bioinformatics/btt468.

Serçinoğlu, Onur, and Pemra Ozbek Sarica. 2019. “In Silico Databases and Tools for Drug Repurposing.” In *In Silico Drug Design*, 703–42. Elsevier. https://doi.org/10.1016/b978-0-12-816125-8.00024-9.

Shajii, Ariya, Deniz Yorukoglu, Yun William Yu, and Bonnie Berger. 2016. “Fast Genotyping of Known SNPs through Approximate K-Mer Matching.” In *Bioinformatics*, 32:i538–44. Oxford University Press. https://doi.org/10.1093/bioinformatics/btw460.

Shalon, Dari, Stephen J. Smith, and Patrick O. Brown. 1996. “A DNA Microarray System for Analyzing Complex DNA Samples Using Two-Color Fluorescent Probe Hybridization.” *Genome Research* 6 (7): 639–45. https://doi.org/10.1101/GR.6.7.639.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Beno Schwikowski, and Trey Ideker. 2003. “Cytoscape: A Software

Environment for Integrated Models of Biomolecular Interaction Networks." *Genome Research* 13 (11): 2498–2504. https://doi.org/10.1101/gr.1239303.

Shendure, Jay. 2008. "The Beginning of the End for Microarrays?" *Nature Methods* 5 (7): 585–87. https://doi.org/10.1038/nmeth0708-585.

Shirin, H., E. M. Sordillo, T. K. Kolevska, H. Hibshoosh, Y. Kawabata, S. H. Oh, J. F. Kuebler, et al. 2000. "Chronic Helicobacter Pylori Infection Induces an Apoptosis-Resistant Phenotype Associated with Decreased Expression of P27kip1." *Infection and Immunity* 68 (9): 5321. https://doi.org/10.1128/IAI.68.9.5321-5328.2000.

Short, Kirsty R., Jennifer Kasper, Stijn Van Der Aa, Arno C. Andeweg, Fatiha Zaaraoui-Boutahar, Marco Goeijenbier, Mathilde Richard, et al. 2016. "Influenza Virus Damages the Alveolar Barrier by Disrupting Epithelial Cell Tight Junctions." *European Respiratory Journal* 47 (3): 954–66. https://doi.org/10.1183/13993003.01282-2015.

Silge, Julia, and David Robinson. 2016. "Tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *Journal of Open Source Software* 1 (3): 37. https://doi.org/10.21105/JOSS.00037.

Sinensky, M., K. Fantle, M. Trujillo, T. McLain, A. Kupfer, and M. Dalton. 1994. "The Processing Pathway of Prelamin A." *Journal of Cell Science* 107 (1): 61–67. https://doi.org/10.1242/JCS.107.1.61.

Sinha, Jitendra Kumar, Shampa Ghosh, and Manchala Raghunath. 2014. "Progeria: A Rare Genetic Premature Ageing Disorder." *The Indian Journal of Medical Research* 139 (5): 667. /pmc/articles/PMC4140030/.

Sonnhammer, Erik L.L., and Gabriel Östlund. 2015. "InParanoid 8: Orthology Analysis between 273 Proteomes, Mostly Eukaryotic." *Nucleic Acids Research* 43 (D1): D234–39. https://doi.org/10.1093/nar/gku1203.

Stark, Rory, Marta Grzelak, and James Hadfield. 2019. "RNA Sequencing: The Teenage Years." *Nature Reviews Genetics 2019 20:11* 20 (11): 631–56. https://doi.org/10.1038/s41576-019-0150-2.

Steer, Andrew C., Jonathan R. Carapetis, James B. Dale, John D. Fraser, Michael F. Good, Luiza Guilherme, Nicole J. Moreland, E. Kim Mulholland, Florian Schodel, and Pierre R. Smeesters. 2016. "Status of Research and Development of Vaccines for Streptococcus Pyogenes." *Vaccine* 34 (26): 2953–58. https://doi.org/10.1016/J.VACCINE.2016.03.073.

Su, Zhenqiang, Hong Fang, Huixiao Hong, Leming Shi, Wenqian Zhang, Wenwei Zhang,

Yanyan Zhang, et al. 2014. “An Investigation of Biomarkers Derived from Legacy Microarray Data for Their Utility in the RNA-Seq Era.” *Genome Biology* 15 (12): 523. https://doi.org/10.1186/S13059-014-0523-Y/TABLES/8.

Subramanian, Aravind, Rajiv Narayan, Steven M. Corsello, David D. Peck, Ted E. Natoli, Xiaodong Lu, Joshua Gould, et al. 2017. “A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles.” *Cell* 171 (6): 1437-1452.e17. https://doi.org/10.1016/j.cell.2017.10.049.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 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): 15545–50. https://doi.org/10.1073/pnas.0506580102.

———. 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): 15545–50. https://doi.org/10.1073/pnas.0506580102.

Sun, Jianqiang, Tomoaki Nishiyama, Kentaro Shimizu, and Koji Kadota. 2013. “TCC: An R Package for Comparing Tag Count Data with Robust Normalization Strategies.” *BMC Bioinformatics* 14 (1): 1–14. https://doi.org/10.1186/1471-2105-14-219/FIGURES/3.

Supek, Fran, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. 2011. “REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms.” Edited by Cynthia Gibas. *PLoS ONE* 6 (7): e21800. https://doi.org/10.1371/journal.pone.0021800.

Szklarczyk, Damian, Annika L. Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, et al. 2019. “STRING V11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets.” *Nucleic Acids Research* 47 (D1): D607–13. https://doi.org/10.1093/nar/gky1131.

Szklarczyk, Damian, Alberto Santos, Christian Von Mering, Lars Juhl Jensen, and Michael Kuhn. 2016. “STITCH 5: Augmenting Protein-Chemical Interaction Networks with Tissue and Affinity Data.” *Nucleic Acids Research* 44. https://doi.org/10.1093/nar/gkv1277.

Tamura, Koichiro, Daniel Peterson, Nicholas Peterson, Glen Stecher, Masatoshi Nei, and Sudhir Kumar. 2011. “MEGA5: Molecular Evolutionary Genetics Analysis Using

Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods." *Molecular Biology and Evolution* 28 (10): 2731. https://doi.org/10.1093/MOLBEV/MSR121.

Tanoli, Ziaur Rehman, Zaid Alam, Markus Vähä-Koskela, Balaguru Ravikumar, Alina Malyutina, Alok Jaiswal, Jing Tang, Krister Wennerberg, and Tero Aittokallio. 2018. "Drug Target Commons 2.0: A Community Platform for Systematic Analysis of Drug-Target Interaction Profiles." *Database* 2018 (2018): 83. https://doi.org/10.1093/database/bay083.

Tarazona, Sonia, Pedro Furió-Tarí, David Turrà, Antonio Di Pietro, María José Nueda, Alberto Ferrer, and Ana Conesa. 2015. "Data Quality Aware Analysis of Differential Expression in RNA-Seq with NOISeq R/Bioc Package." *Nucleic Acids Research* 43 (21): e140. https://doi.org/10.1093/nar/gkv711.

Tate, John G., Sally Bamford, Harry C. Jubb, Zbyslaw Sondka, David M. Beare, Nidhi Bindal, Harry Boutselakis, et al. 2019. "COSMIC: The Catalogue Of Somatic Mutations In Cancer." *Nucleic Acids Research* 47 (D1): D941–47. https://doi.org/10.1093/nar/gky1015.

Thompson, Julie D., Desmond G. Higgins, and Toby J. Gibson. 1994. "CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice." *Nucleic Acids Research* 22 (22): 4673–80. https://doi.org/10.1093/nar/22.22.4673.

Tilli, C. M.L.J., F. C.S. Ramaekers, J. L.V. Broers, C. J. Hutchison, and H. A.M. Neumann. 2003. "Lamin Expression in Normal Human Skin, Actinic Keratosis, Squamous Cell Carcinoma and Basal Cell Carcinoma." *British Journal of Dermatology* 148 (1): 102–9. https://doi.org/10.1046/J.1365-2133.2003.05026.X.

Tocris. 2020. "Emetine." 2020. https://www.tocris.com/products/emetine-dihydrochloride_7342.

———. 2022a. "Anisomycin." 2022. https://www.tocris.com/products/anisomycin_1290.

———. 2022b. "CHIR 99021." 2022. https://www.tocris.com/products/chir-99021_4423.

———. 2022c. "Kenpaullone." 2022. https://www.tocris.com/products/kenpaullone_1398.

———. 2022d. "LY 294002 Hydrochloride." 2022. https://www.tocris.com/products/ly-294002-hydrochloride_1130.

———. 2022e. "Rapamycin." 2022. https://www.tocris.com/products/rapamycin_1292.

Bibliography

Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R. Kelley, Harold Pimentel, Steven L. Salzberg, John L. Rinn, and Lior Pachter. 2012. "Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks." *Nature Protocols* 7 (3): 562–78. https://doi.org/10.1038/nprot.2012.016.

Trapnell, Cole, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. Van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. 2010. "Transcript Assembly and Abundance Estimation from RNA-Seq Reveals Thousands of New Transcripts and Switching among Isoforms." *Nature Biotechnology* 28 (5): 511. https://doi.org/10.1038/NBT.1621.

Troulé, Kevin, Hugo López-Fernández, Santiago García-Martín, Miguel Reboiro-Jato, Carlos Carretero-Puche, Jordi Martorell-Marugán, Guillermo Martín-Serrano, et al. 2020. "DREIMT: A Drug Repositioning Database and Prioritization Tool for Immunomodulation." *Bioinformatics*, August. https://doi.org/10.1093/bioinformatics/btaa727.

Ullrich, Nicole J., and Leslie B. Gordon. 2015. "Hutchinson–Gilford Progeria Syndrome." *Handbook of Clinical Neurology* 132 (January): 249–64. https://doi.org/10.1016/B978-0-444-62702-5.00018-4.

Ursu, Oleg, Jayme Holmes, Jeffrey Knockel, Cristian G Bologa, Jeremy J Yang, Stephen L Mathias, Stuart J Nelson, and Tudor I Oprea. 2017. "DrugCentral: Online Drug Compendium." *Nucleic Acids Research* 45. https://doi.org/10.1093/nar/gkw993.

Vallenet, David, Alexandra Calteau, Mathieu Dubois, Paul Amours, Adelme Bazin, Mylène Beuvin, Laura Burlot, et al. 2020. "MicroScope: An Integrated Platform for the Annotation and Exploration of Microbial Gene Functions through Genomic, Pangenomic and Metabolic Comparative Analysis." *Nucleic Acids Research* 48 (D1): D579–89. https://doi.org/10.1093/NAR/GKZ926.

Varet, Hugo, Loraine Brillet-Guéguen, Jean Yves Coppée, and Marie Agnès Dillies. 2016. "SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data." *PLOS ONE* 11 (6): e0157022. https://doi.org/10.1371/JOURNAL.PONE.0157022.

Velculescu, Victor E., Lin Zhang, Bert Vogelstein, and Kenneth W. Kinzler. 1995. "Serial Analysis of Gene Expression." *Science* 270 (5235): 484–87. https://doi.org/10.1126/SCIENCE.270.5235.484.

Bibliography

Velculescu, Victor E., Lin Zhang, Wei Zhou, Jacob Vogelstein, Munira A. Basrai, Douglas E. Bassett, Phil Hieter, Bert Vogelstein, and Kenneth W. Kinzler. 1997. "Characterization of the Yeast Transcriptome." *Cell* 88 (2): 243–51. https://doi.org/10.1016/S0092-8674(00)81845-0.

Vêncio, Ricardo Z.N., Tie Koide, Suely L. Gomes, and Carlos A.de B. Pereira. 2006. "BayGO: Bayesian Analysis of Ontology Term Enrichment in Microarray Data." *BMC Bioinformatics* 7 (1): 86. https://doi.org/10.1186/1471-2105-7-86.

Verbruggen, Bas, Lina Gunnarsson, Erik Kristiansson, Tobias Österlund, Stewart F. Owen, Jason R. Snape, and Charles R. Tyler. 2018. "ECOdrug: A Database Connecting Drugs and Conservation of Their Targets across Species." *Nucleic Acids Research* 46 (D1): D930–36. https://doi.org/10.1093/nar/gkx1024.

Vidak, Sandra, and Roland Foisner. 2016. "Molecular Insights into the Premature Aging Disease Progeria." *Histochemistry and Cell Biology* 145 (4): 401. https://doi.org/10.1007/S00418-016-1411-1.

Volodarsky, Dina, Noam Leviatan, Andrei Otcheretianski, and Robert Fluhr. 2009. "HORMONOMETER: A Tool for Discerning Transcript Signatures of Hormone Action in the Arabidopsis Transcriptome." *Plant Physiology* 150 (4): 1796–1805. https://doi.org/10.1104/pp.109.138289.

Wang, Chen, Gang Hu, Kui Wang, Michal Brylinski, Lei Xie, and Lukasz Kurgan. 2016. "PDID: Database of Molecular-Level Putative Protein–Drug Interactions in the Structural Human Proteome." *Bioinformatics* 32 (4): 579–86. https://doi.org/10.1093/bioinformatics/btv597.

Wang, Hongjie, Corinne Ducournau, Kamola Saydaminova, Maximilian Richter, Roma Yumul, Martin Ho, Darrick Carter, Chloé Zubieta, Pascal Fender, and André Lieber. 2015. "Intracellular Signaling and Desmoglein 2 Shedding Triggered by Human Adenoviruses Ad3, Ad14, and Ad14P1." *Journal of Virology* 89 (21): 10841–59. https://doi.org/10.1128/jvi.01425-15.

Wang, Hongjie, Zong Yi Li, Ying Liu, Jonas Persson, Ines Beyer, Thomas Möller, Dilara Koyuncu, et al. 2011. "Desmoglein 2 Is a Receptor for Adenovirus Serotypes 3, 7, 11, and 14." *Nature Medicine* 17 (1): 96. https://doi.org/10.1038/NM.2270.

Wang, Liguo, Shengqin Wang, and Wei Li. 2012. "RSeQC: Quality Control of RNA-Seq Experiments." *Bioinformatics* 28 (16): 2184–85.

https://doi.org/10.1093/bioinformatics/bts356.

Wang, Lingyan, Bishi Fu, Wenjun Li, Girish Patil, Lin Liu, Martin E. Dorf, and Shitao Li. 2017. "Comparative Influenza Protein Interactomes Identify the Role of Plakophilin 2 in Virus Restriction." *Nature Communications* 8 (1): 1–12. https://doi.org/10.1038/ncomms13876.

Wang, Miao, and Randal J. Kaufman. 2016. "Protein Misfolding in the Endoplasmic Reticulum as a Conduit to Human Disease." *Nature* 529 (7586): 326–35. https://doi.org/10.1038/nature17041.

Wang, Yunxia, Song Zhang, Fengcheng Li, Ying Zhou, Ying Zhang, Zhengwen Wang, Runyuan Zhang, et al. 2020. "Therapeutic Target Database 2020: Enriched Resource for Facilitating Research and Early Development of Targeted Therapeutics." *Nucleic Acids Research* 48 (D1): D1031–41. https://doi.org/10.1093/nar/gkz981.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. "RNA-Seq: A Revolutionary Tool for Transcriptomics." *Nature Reviews Genetics*. NIH Public Access. https://doi.org/10.1038/nrg2484.

Wang, Zichen, Alexander Lachmann, Alexandra B. Keenan, and Avi Ma'Ayan. 2018. "L1000FWD: Fireworks Visualization of Drug-Induced Transcriptomic Signatures." *Bioinformatics* 34 (12): 2150–52. https://doi.org/10.1093/bioinformatics/bty060.

Wang, Zichen, Caroline D. Monteiro, Kathleen M. Jagodnik, Nicolas F. Fernandez, Gregory W. Gundersen, Andrew D. Rouillard, Sherry L. Jenkins, et al. 2016. "Extraction and Analysis of Signatures from the Gene Expression Omnibus by the Crowd." *Nature Communications* 7 (September). https://doi.org/10.1038/NCOMMS12846.

Warren, René L., Chen Yang, Benjamin P. Vandervalk, Bahar Behsaz, Albert Lagman, Steven J.M. Jones, and Inanç Birol. 2015. "LINKS: Scalable, Alignment-Free Scaffolding of Draft Genomes with Long Reads." *GigaScience* 4 (1). https://doi.org/10.1186/s13742-015-0076-3.

Wass, Mark N., and Michael J.E. Sternberg. 2008. "ConFunc - Functional Annotation in the Twilight Zone." *Bioinformatics* 24 (6): 798–806. https://doi.org/10.1093/bioinformatics/btn037.

Weber, Michael, Sebastian G. Henkel, Sebastian Vlaic, Reinhard Guthke, Everardus J. van Zoelen, and Dominik Driesch. 2013. "Inference of Dynamical Gene-Regulatory Networks Based on Time-Resolved Multi-Stimuli Multi-Experiment Data Applying

NetGenerator V2.0." *BMC Systems Biology* 7 (1): 1–16. https://doi.org/10.1186/1752-0509-7-1/FIGURES/8.

Westermann, Alexander J., Stanislaw A. Gorski, and Jörg Vogel. 2012. "Dual RNA-Seq of Pathogen and Host." *Nature Reviews Microbiology*. https://doi.org/10.1038/nrmicro2852.

Whirl-Carrillo, M, E M Mcdonagh, J M Hebert, L Gong, K Sangkuhl, C F Thorn, R B Altman, and T E Klein. 2012. "Pharmacogenomics Knowledge for Personalized Medicine." https://doi.org/10.1038/clpt.2012.96.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D', Agostino Mcgowan, Romain François, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/JOSS.01686.

Widmann, Jeremy, Jesse Stombaugh, Daniel Mcdonald, Jana Chocholousova, Paul Gardner, Matthew K. Iyer, Zongzhi Liu, et al. 2012. "RNASTAR: An RNA STructural Alignment Repository That Provides Insight into the Evolution of Natural and Artificial RNAs." *RNA* 18 (7): 1319. https://doi.org/10.1261/RNA.032052.111.

Williams, Claire R., Alyssa Baccarella, Jay Z. Parrish, and Charles C. Kim. 2016. "Trimming of Sequence Reads Alters RNA-Seq Gene Expression Estimates." *BMC Bioinformatics* 17 (1): 103. https://doi.org/10.1186/s12859-016-0956-2.

Wishart, David S, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, et al. 2018. "DrugBank 5.0: A Major Update to the DrugBank Database for 2018." *Nucleic Acids Research* 46. https://doi.org/10.1093/nar/gkx1037.

Wood, Derrick E., and Steven L. Salzberg. 2014. "Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments." *Genome Biology* 15 (3). https://doi.org/10.1186/gb-2014-15-3-r46.

Wu, Douglas C., Jun Yao, Kevin S. Ho, Alan M. Lambowitz, and Claus O. Wilke. 2018. "Limitations of Alignment-Free Tools in Total RNA-Seq Quantification." *BMC Genomics* 19 (1): 510. https://doi.org/10.1186/s12864-018-4869-5.

Wu, Thomas D., and Serban Nacu. 2010. "Fast and SNP-Tolerant Detection of Complex Variants and Splicing in Short Reads." *Bioinformatics* 26 (7): 873–81. https://doi.org/10.1093/bioinformatics/btq057.

Xie, Chen, Xizeng Mao, Jiaju Huang, Yang Ding, Jianmin Wu, Shan Dong, Lei Kong, Ge Gao, Chuan Yun Li, and Liping Wei. 2011. "KOBAS 2.0: A Web Server for Annotation and

Identification of Enriched Pathways and Diseases." *Nucleic Acids Research* 39 (SUPPL. 2): W316. https://doi.org/10.1093/nar/gkr483.

Xu, Guorong, Michael J. Strong, Michelle R. Lacey, Carl Baribault, Erik K. Flemington, and Christopher M. Taylor. 2014. "RNA CoMPASS: A Dual Approach for Pathogen and Host." *PLoS ONE*. https://doi.org/10.1371/journal.pone.0089445.

Yamanishi, Yoshihiro, Masaaki Kotera, Yuki Moriya, Ryusuke Sawada, Minoru Kanehisa, and Susumu Goto. 2014. "DINIES: Drug-Target Interaction Network Inference Engine Based on Supervised Analysis." *Nucleic Acids Research* 42 (W1): 39–45. https://doi.org/10.1093/nar/gku337.

Yan, Linlin. 2021. "Ggvenn: Draw Venn Diagram by 'Ggplot2'. R Package Version 0.1.9." 2021.

Yang, Wanjuan, Jorge Soares, Patricia Greninger, Elena J Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, et al. 2012. "Genomics of Drug Sensitivity in Cancer (GDSC): A Resource for Therapeutic Biomarker Discovery in Cancer Cells." https://doi.org/10.1093/nar/gks1111.

Yates, Andrew D., Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M. Ridwan Amode, et al. 2020. "Ensembl 2020." *Nucleic Acids Research* 48 (D1): D682–88. https://doi.org/10.1093/nar/gkz966.

Young, Matthew D., Matthew J. Wakefield, Gordon K. Smyth, and Alicia Oshlack. 2010. "Gene Ontology Analysis for RNA-Seq: Accounting for Selection Bias." *Genome Biology* 11 (2): R14. https://doi.org/10.1186/gb-2010-11-2-r14.

Yu, Guangchuang, Li Gen Wang, Yanyan Han, and Qing Yu He. 2012. "ClusterProfiler: An R Package for Comparing Biological Themes among Gene Clusters." *OMICS A Journal of Integrative Biology* 16 (5): 284–87. https://doi.org/10.1089/omi.2011.0118.

Yu, Hasun, Sungji Choo, Junseok Park, Jinmyung Jung, Yeeok Kang, and Doheon Lee. 2016. "Prediction of Drugs Having Opposite Effects on Disease Genes in a Directed Network." *BMC Systems Biology* 10 (1): S2. https://doi.org/10.1186/s12918-015-0243-2.

Zakikhany, K., M. A. Degail, T. Lamagni, P. Waight, R. Guy, H. Zhao, A. Efstratiou, R. Pebody, R. George, and M. Ramsay. 2011. "Increase in Invasive Streptococcus Pyogenes and Streptococcus Pneumoniae Infections in England, December 2010 to January 2011." *Eurosurveillance* 16 (5): 19785. https://doi.org/10.2807/ESE.16.05.19785-EN/CITE/PLAINTEXT.

Bibliography

Zerbino, Daniel R. 2010. “Using the Velvet de Novo Assembler for Short-Read Sequencing Technologies.” *Current Protocols in Bioinformatics*. NIH Public Access. https://doi.org/10.1002/0471250953.bi1105s31.

Zhang, Jiajie, Kassian Kobert, Tomá ˇ Flouri, and Alexandros Stamatakis. 2014. “PEAR: A Fast and Accurate Illumina Paired-End ReAd MergeR” 30 (5): 614–20. https://doi.org/10.1093/bioinformatics/btt593.

Zhang, Shu-Dong, and Timothy W Gant. 2009. “SscMap: An Extensible Java Application for Connecting Small-Molecule Drugs Using Gene-Expression Signatures.” https://doi.org/10.1186/1471-2105-10-236.

Zhang, Yang. 2008. “I-TASSER Server for Protein 3D Structure Prediction.” *BMC Bioinformatics* 9 (1): 40. https://doi.org/10.1186/1471-2105-9-40.

Zhang, Zhang, Francesc López-Giráldez, and Jeffrey P. Townsend. 2010. “LOX: Inferring Level Of EXpression from Diverse Methods of Census Sequencing.” *Bioinformatics* 26 (15): 1918. https://doi.org/10.1093/BIOINFORMATICS/BTQ303.

Zhao, Shanrong, Ying Zhang, William Gordon, Jie Quan, Hualin Xi, Sarah Du, David von Schack, and Baohong Zhang. 2015. “Comparison of Stranded and Non-Stranded RNA-Seq Transcriptome Profiling and Investigation of Gene Overlap.” *BMC Genomics 2015 16:1* 16 (1): 1–14. https://doi.org/10.1186/S12864-015-1876-7.

Zheng, Q., and Xiu Jie Wang. 2008. “GOEAST: A Web-Based Software Toolkit for Gene Ontology Enrichment Analysis.” *Nucleic Acids Research* 36 (Web Server issue): W358. https://doi.org/10.1093/nar/gkn276.

Zhong, Xiaoming, Jana Drgonova, Chuan Yun Li, and George R. Uhl. 2015. “Human Cell Adhesion Molecules: Annotated Functional Subtypes and Overrepresentation of Addiction-Associated Genes.” *Annals of the New York Academy of Sciences* 1349 (1): 83. https://doi.org/10.1111/NYAS.12776.

Zhou, Peng, Kevin A.T. Silverstein, Liangliang Gao, Jonathan D. Walton, Sumitha Nallu, Joseph Guhlin, and Nevin D. Young. 2013. “Detecting Small Plant Peptides Using SPADA (Small Peptide Alignment Discovery Application).” *BMC Bioinformatics* 14 (1): 1–16. https://doi.org/10.1186/1471-2105-14-335/FIGURES/5.

Zhou, Xianxiao, Minghui Wang, Igor Katsyv, Hanna Irie, and Bin Zhang. 2018. “EMUDRA: Ensemble of Multiple Drug Repositioning Approaches to Improve Prediction Accuracy.” *Bioinformatics* 34 (18): 3151–59. https://doi.org/10.1093/bioinformatics/bty325.

Bibliography

Zhou, Yingyao, Bin Zhou, Lars Pache, Max Chang, Alireza Hadj Khodabakhshi, Olga Tanaseichuk, Christopher Benner, and Sumit K. Chanda. 2019. "Metascape Provides a Biologist-Oriented Resource for the Analysis of Systems-Level Datasets." *Nature Communications 2019 10:1* 10 (1): 1–10. https://doi.org/10.1038/s41467-019-09234-6.

Zielezinski, Andrzej, Susana Vinga, Jonas Almeida, and Wojciech M. Karlowski. 2017. "Alignment-Free Sequence Comparison: Benefits, Applications, and Tools." *Genome Biology*. BioMed Central Ltd. https://doi.org/10.1186/s13059-017-1319-7.

Zinkernagel, Annelies S, Anjuli M Timmer, Morgan A Pence, Jeffrey B Locke, John T Buchanan, Claire E Turner, Inbal Mishalian, Shiranee Sriskandan, Emanuel Hanski, and Victor Nizet. 2008. "The IL-8 Protease SpyCEP/ScpC of Group A Streptococcus Promotes Resistance to Neutrophil Killing." *Cell Host & Microbe* 4 (2): 170–78. https://doi.org/10.1016/j.chom.2008.07.002.

Curriculum Vitae

Personal Details

Name: Salem Oduro Beffi

Surname: Sueto

Date of Birth: 2nd November 1988

Place of Birth: Accra, Ghana

Nationality: Ghana – Italy

Language: English, Italian, Twi (Ghana)

Email: sueto.pharmatec@gmail.com

GitHub: https://github.com/SalemSueto-BioInfo - https://github.com/salemSueto

LinkedIn: https://www.linkedin.com/in/salem-oduro-beffi-sueto-314bb978/

EDUCATION

Nov 2017 – Nov 2023	PhD Bioinformatics, University of Rostock, Germany
Sep 2016 – Sep 2017	Masters Bioinformatics, University of Glasgow, Scotland
Nov 2011 – Feb 2014	Masters Biotechnology, University of Milano-Bicocca, Italy
Sep 2008 – Nov 2011	Bachelor Biotechnology, University of Milano-Bicocca, Italy

WORK EXPERIENCE

Sep 2022 – Present	Bioinformatician, Scotland Rural College (Edinburgh, U.K.)
Sep 2015 – Sep 2016	Technician, Pharmatec S.R.L. (Milano, Italy)

WORKSHOP

08 Oct 2018 - **2nd NGS Workshop M-V,** Leibniz Institute (Dummerstorf, Germany)

06 Mar 2019 - **Galaxy for linking bisulfite sequencing with RNA sequencing**
de.STAIR (Rostock, Germany)

ACCOMPLISHMENTS

PhD scholarship / doktorandenstipendium (Nov 2017 – Oct 2022)

CERTIFICATE

Academic Writing in Natural Sciences (14/06/2018 – 14/06/2018)
Qualification Program of the Graduate Academy of the University of Rostock

Career Development through Research Funding (12/06/2018 – 12/06/2018)

Qualification Program of the Graduate Academy of the University of Rostock

Rhetoric & Presentation Skills (19/06/2018 – 19/06/2018)

Qualification Program of the Graduate Academy of the University of Rostock

Project Management in Academia (06/09/2018 – 06/09/2018)

Qualification Program of the Graduate Academy of the University of Rostock

Good Scientific Practice (10/10/2018 – 10/10/2018)

Qualification Program of the Graduate Academy of the University of Rostock

PUBLICATIONS

Vector mapping and bloodmeal metabarcoding demonstrate risk of urban Chagas disease transmission in Caracas, Venezuela. 2023. PLOS Neglected Tropical Diseases. https://doi.org/10.1371/journal.pntd.0010613. Maikell Segovia, Philipp Schwabl, Salem Sueto, Candy Cherine Nakad, Juan Carlos Londoño, Marlenes Rodriguez, Manuel Paiva, Martin Stephen Llewellyn, Hernán José Carrasco.

Computational identification of natural senotherapeutic compounds that mimic Dasatinib based on gene expression data. 2022 https://www.biorxiv.org/content/10.1101/2022.05.26.492763v1 Franziska Meiners; Riccardo Secci; Salem Sueto; Georg Fuellen; and Israel Barrantes.

Remarkable Genetic Diversity of Trypanosoma cruzi and Trypanosoma rangeli in Two Localities of Southern Ecuador Identified via Deep Sequencing of Mini-Exon Gene Amplicons. 2020 https://doi.org/10.1186/S13071-020-04079-1 Sánchez Jalil; Sueto Salem; Schwabl Philipp; Grijalva Mario; Llewellyn Martin; and Costales Jaime.