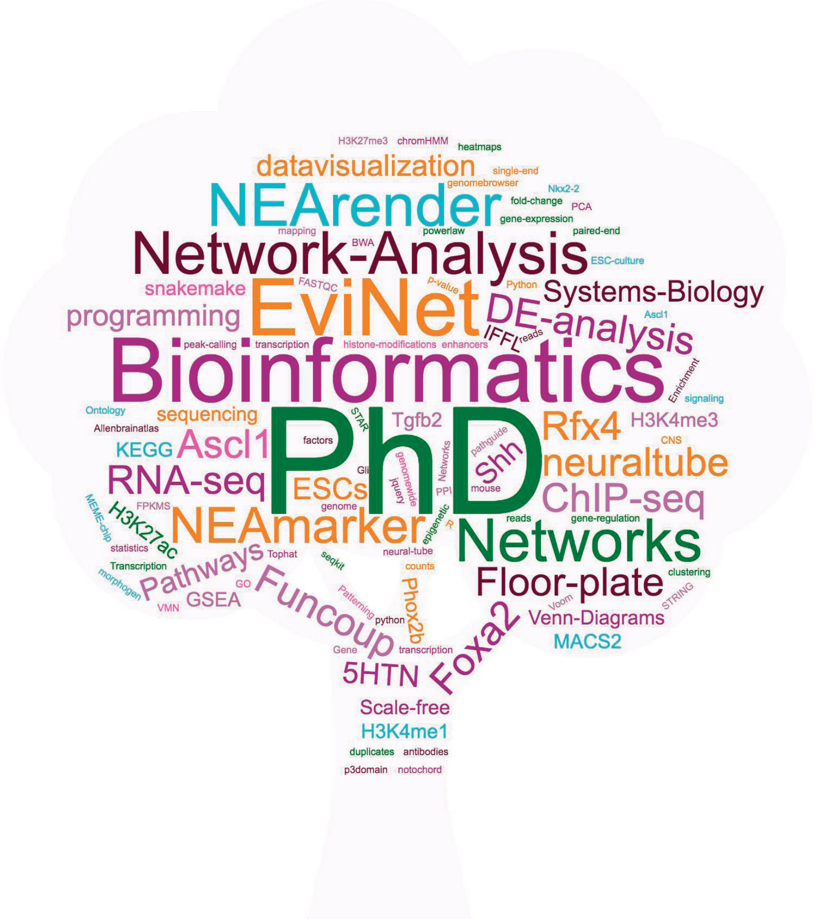


Karolinska Institutet

Thesis for doctoral degree (Ph.D.) 2019

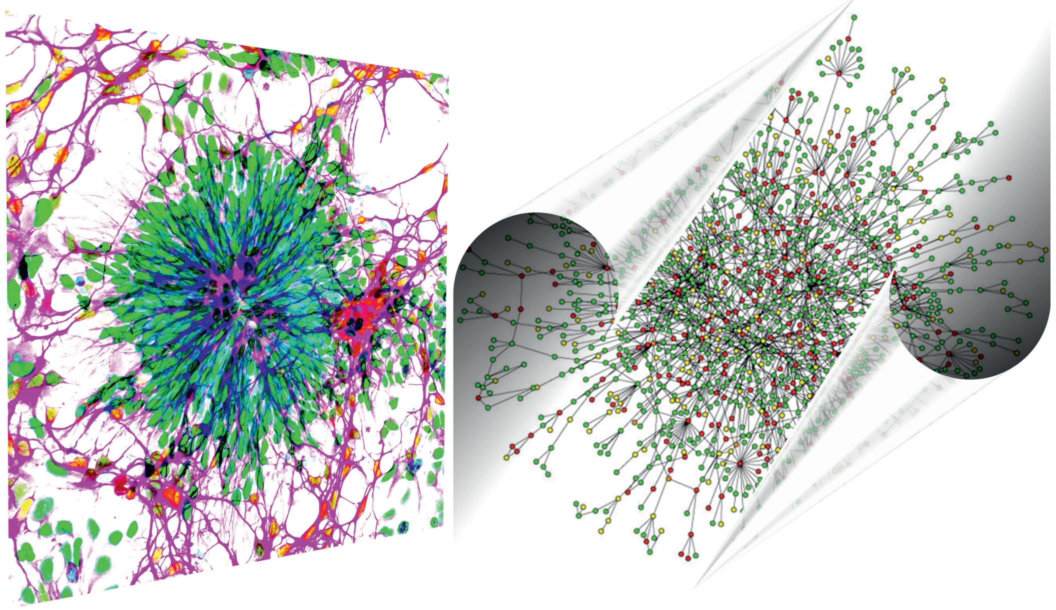
Exploration of Large Molecular Datasets using Global Gene Networks
Computational Methods and Tools

Ashwini Jeggari

Thesis for doctoral degree (Ph.D.)
2019

Exploration of Large Molecular Datasets using Global Gene Networks

Computational Methods and Tools



Ashwini Jeggari

Karolinska Institutet

From Department of Cell and Molecular Biology
Karolinska Institutet, Stockholm, Sweden

EXPLORATION OF LARGE MOLECULAR DATASETS USING GLOBAL GENE NETWORKS

COMPUTATIONAL METHODS AND TOOLS

Ashwini Priya Jeggari

Stockholm 2019

Cover picture: Representation of *in vitro* and *in silico* models
Neural differentiation of ESCs. Credit: Zhanna Alekseenko, Karolinska Institutet, Stockholm, Sweden.
Figure modified for yeast protein-protein interaction network. Credit: Hawoong Jeong, KAIST, Korea.

Published by Karolinska Institutet.
Printed by Eprint AB 2019

ISBN 978-91-7831-491-1

Exploration of Large Molecular Datasets using Global Gene Networks: Computational Methods and Tools

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Ashwini Priya Jeggari

Principal Supervisor:
Prof. Johan Ericson
Karolinska Institutet
Department of Cell and Molecular Biology

Co-supervisor(s):
Dr. Andrey Alexeyenko
Karolinska Institutet
Department of Microbiology, Tumor and Cell Biology

Dr. Zhanna Alekseenko
Karolinska Institutet
Department of Cell and Molecular Biology

Dr. Maria Bergsland
Karolinska Institutet
Department of Cell and Molecular Biology

Opponent:
Adam A Margolin, PhD
Ichan School of Medicine at Mount Sinai
Department of Genetics and Genomic Science

Examination Board:
Prof. Lukas Käll
KTH Royal Institute of Technology
Department of Statistical Biotechnology

Prof. Nico Dantuma
Karolinska Institutet
Department of Cell and Molecular Biology

Dr. Francois Lallemand
Karolinska Institutet
Department of Neuroscience

To my family

ABSTRACT

Defining gene expression profiles and mapping complex interactions between molecular regulators and proteins is a key for understanding biological processes and the functional properties of cells, which is therefore, the focus on numerous experimental studies. Small-scale biochemical analyses deliver high-quality data, but lack coverage, whereas high throughput sequencing reveals thousands of interactions which can be error-prone and require proper computational methods to discover true relations. Furthermore, all these approaches usually focus on one type of interaction at a time. This makes experimental mapping of the genome-wide network a cost and time-intensive procedure.

In the first part of the thesis, I present the developed network analysis tools for exploring large-scale datasets in the context of a global network of functional coupling.

Paper I introduces NEArender, a method for performing pathway analysis and determines the relations between gene sets using a global network. Traditionally, pathway analysis did not consider network relations, thereby covering a minor part of the whole picture. Placing the gene sets in the context of a network provides additional information for pathway analysis, which reveals a more comprehensive picture.

Paper II presents EviNet, a user-friendly web interface for using NEArender algorithm. The user can either input gene lists or manage and integrate highly complex experimental designs via the interactive Venn diagram-based interface. The web resource provides access to biological networks and pathways from multiple public or users' own resources. The analysis typically takes seconds or minutes, and the results are presented in a graphic and tabular format.

Paper III describes NEAmarker, a method to predict anti-cancer drug targets from enrichment scores calculated by NEArender, thus presenting a practical usage of network enrichment tool. The method can integrate data from multiple omics platforms to model drug sensitivity with enrichment variables. In parallel, alternative methods for pathway enrichment analysis were benchmarked in the paper.

The second part of the thesis is focused on identifying spatial and temporal mechanisms that govern the formation of neural cell diversity in the developing brain. High-throughput platforms for RNA- and ChIP-sequencing were applied to provide data for studying the underlying biological hypothesis at the genome-wide scale.

In **Paper IV**, I defined the role of the transcription factor Foxa2 during the specification and differentiation of floor plate cells of the ventral neural tube. By RNA-seq analyses of Foxa2$^{-/-}$ cells, a large set of candidate genes involved in floor plate differentiation were identified. Analysis of Foxa2 ChIP-seq dataset suggested that *Foxa2* directly regulated more than 250 genes expressed by the floor plate and identified *Rfx4* and *Ascl1* as co-regulators of many floor plate genes. Experimental studies suggested a cooperative activator function for *Foxa2* and *Rfx4* and a suppressive role for *Ascl1* in spatially constraining floor plate induction.

Paper V addresses how time is measured during sequential specification of neurons from multipotent progenitor cells during the development of ventral hindbrain. An underlying timer circuitry which leads to the sequential generation of motor neurons and serotonergic neurons has been identified by integrating experimental and computational data modeling.

LIST OF PUBLICATIONS

I. **Ashwini Jeggari**, Andrey Alexeyenko (2017). NEArender: an R package for functional interpretation of ‘omics’ data via network enrichment analysis. BMC Bioinformatics, 18 (Suppl 5):118.

II. **Ashwini Jeggari**, Zhanna Alekseenko, Iurii Petrov, José M Dias, Johan Ericson, Andrey Alexeyenko (2018). EviNet: a web platform for network enrichment analysis with flexible definition of gene sets. Nucleic Acids Research, Volume 46, Issue W1, Pages W163–W170.

III. Marcela Franco, **Ashwini Jeggari**, Sylvain peuget, Franziska Böttger, Galina selivanova, Andrey Alexeyenko (2019). Prediction of response to anti-cancer drugs becomes robust via network integration of molecular data. Scientific Reports, volume 9, Article number: 2379.

IV. **Ashwini Jeggari**, Mariya Kozhevnikova, Christopher W. Uhde, José M. Dias, Zhanna Alekseenko, Katarina Gradin, Elisabet Andersson, Mark D. Borromeo, Jane E. Johnson, Andrey Alexeyenko, and Johan Ericson. Genome-wide characterisation of floor plate transcription reveals cooperative activator function of Foxa2 and Rfx4 and a suppressive role for Ascl1 to spatially constrain floor plate induction in the neural tube. Manuscript.

V. Jose M Dias, Zhanna Alekseenko, **Ashwini Jeggari**, Jannik Vollmer, Mariya Kozhevnikova, Michael P. Matise, Andrey Alexeyenko, Dagmer Iber, Johan Ericson. A Shh/Gli-driven three-node timer device controls temporal identity and fate of neural stem cells. Manuscript.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AGS	Altered Gene Set
AP	Anterior-Posterior
bp	DNA base pairs
CNS	Central Nervous System
DB	Database
DE	Differential Expression
DV	Dorso-Ventral
DDC	Day in Differentiation Condition
ENCODE	Encyclopedia of DNA Elements
FGS	Functional Gene Set
FP	Floor plate
FPKM	Fragments per kilobase and million reads mapped
GEO	Gene Expression Omnibus
GSEA	Gene Set Enrichment Analysis
GO	Gene Ontology
IGV	Integrative Genome Viewer
HTS	High Throughput Sequencing
MN	Motor Neuron
NEA	Network Enrichment Analysis
NGS	Next Generation Sequencing
NW	Network
ORA	Over-Representation Analysis
OLP	Oligodendrocyte Precursor cells
PPI	Protein-Protein Interaction
RP	Roof plate
Shh	Sonic hedge hog
TF	Transcription Factor
TCGA	The Cancer Genome Atlas
5HTN	Serotonergic Neuron

“*Data do not give up their secrets easily. They must be tortured to confess*”

- Jeff Hopper, Bell Labs

1 INTRODUCTION

Different types of specialized cells in multicellular organisms share the same genetic information but differ in their morphology and function, since different gene expression programs are active in the different cell types. Transcription factors (TFs) constitute a large group of proteins that can bind to specific DNA sequences and regulates genes by activating or repressing gene expression. Transcription factors, together with histones and other chromatin-associated proteins, dictate patterns of gene activation in different types of cells (Coulon *et al.*, 2013). Therefore, the challenge here is to understand how these molecular interactions operate in a cell and regulate the gene regulatory network at the systems level. The field of systems biology has evolved to understand these complex biological interactions.

Systems biology is an interdisciplinary field of biology that attempts to use new perspectives in order to comprehensively and systematically analyze complex interactions in biological systems, thereby prioritizing holism over reductionism. It applies a wide range of computational techniques to analyze the data sets of varying sizes and types in order to build descriptive or even predictive models (Tavassoly *et al.*, 2018). Two strategies are employed in the model building:

a) **Top-down approach**: This approach departs from a big picture and then descends to smaller segments. The datasets from high throughput ('-omics') profiling, such as next-generation sequencing, contain vast amounts of information and are meant to describe the genome, proteome, transcriptome etc. However, mining this data requires advanced analytical approaches. Here, modeling techniques, if chosen appropriately, can help to identify key pathways and mechanisms underlying biological processes. This analysis can also provide an insight into the organization, as well as relations among the components (Kitano, 2001; Tavassoly *et al.*, 2018).

b) **Bottom-up approach**: It is mechanistic, contrary to the top-down approach, and starts by modeling cell parts and sub-networks. It usually employs data from small-scale experiments, such as biochemical or molecular biology assays. The experiments can measure key system variables as functions of time or space. Therefore, a bottom-up approach is particularly suited when most of the genes/proteins and their relationships are already known, which enables finding the last few missing pieces to the puzzle or specify the detailed parameters of their interaction. A typical workflow of a bottom-up approach would thus be: i) to build a simulation model that can be used to analyze the dynamical properties of the system by changing parameters that cannot be manipulated in the actual system. ii) to validate the results generated through simulation and its consistency with the experimental data (Kitano, 2001; Tavassoly *et al.*, 2018).

The major part of this thesis is concentrated on a top-down approach, where I applied topological analysis to explore large-scale, omics data in the context of global networks of functional coupling. To this end, I developed network analysis tools for testing underlying biological hypotheses. Paper V shows an example where a bottom-up approach has been employed using mathematical modeling to define timer motif (feed-forward loop) underlying the sequential generation of motor neurons (MNs) and serotonergic neurons (5HTNs) by neural progenitor stem cells in the brainstem.

2 BACKGROUND

2.1 HIGH THROUGHPUT SEQUENCING

"...[*A*] *knowledge of sequences could contribute much to our understanding of living matter"*

- Frederick Sanger

The discovery of DNA in 1869 by Friedrich Miescher, development of classical genetics by Mendel and research on how the human body orchestrates its function led us to this point in history: the human genome sequencing.

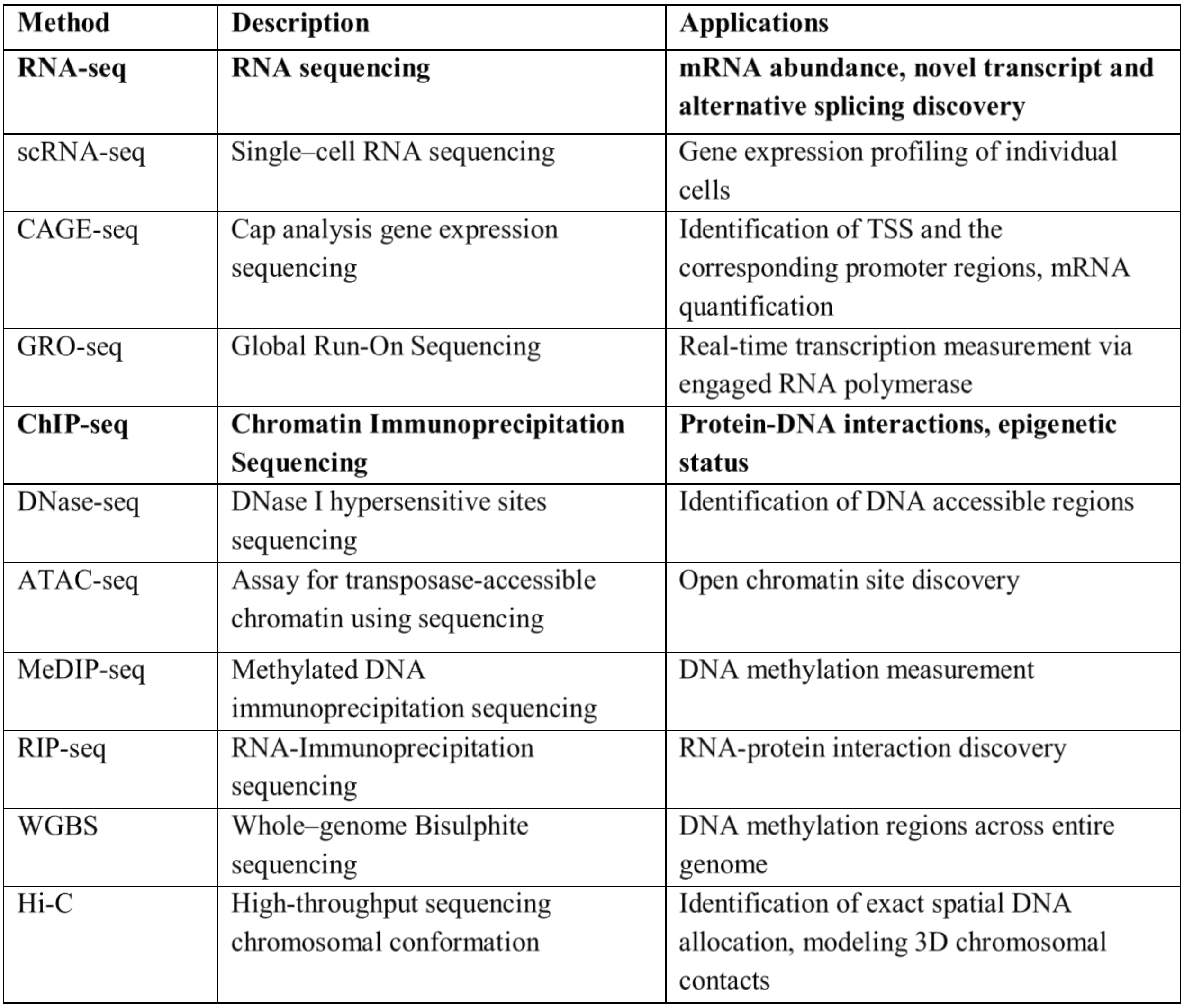
2.1.1 History of genome sequencing

Sequencing (or DNA sequencing) refers to the precise determination of its base pairs (A, T, C, G) in a DNA sample. Around 1977, two methods: chain termination method by Sanger and Coulson (Sanger *et al.*, 1977), chemical cleavage procedure developed by Maxam and Gilbert (Maxam *et al.*, 1977) transformed the field of genomics. The first whole genome of an organism (Bacteriophage ΦX174) was sequenced in 1977 (Sanger *et al.*, 1977), which used gel electrophoresis and manual calling bases for DNA sequencing. After years of developments, Sanger sequencing introduced capillary electrophoresis, which allowed to sequence longer fragments with automated base calls. The first automated DNA Sanger sequencing was introduced by Applied Biosystems (invented by Lyold M. Smith) in 1987 (Cook Deegan *et al.*, 2014). The Sanger (or first-generation) sequencing was widely used for three decades, but the cost and time were major stumbling blocks to sequence complex genomes.

Since 2005, Next Generation Sequencing (NGS) (or second-generation) technologies, namely 454, ABI SOLID, Illumina arrived at the market (Shendure *et al.*, 2008). With such advancements as speed, rapid sequencing of millions of nucleotides (reads) in a single run, low cost, better genome coverage and accuracy, the NGS technologies have in no time incorporated into the modern research world (Shendure *et al.*, 2017). This generation uses either sequencing by synthesis (454, Illumina) or sequencing by ligation (ABI/SOLID). The NGS technologies employ a PCR amplification step to obtain a DNA library of a sufficient amount for loading into the sequencer. Technically, this step is expensive in terms of time and money. Another caveat of NGS technologies is shorter reads (50-700 bp) (Genohub), which creates problems while assembling larger, mostly diploid genomes, especially in the repetitive regions or for very similar gene alleles (Roberts *et al.*, 2013). Promising to overcome these drawbacks, third-generation (or long-read) sequencers based on single-molecule real-time sequencing (SMRT) approach from PacBio, Illumina True-Seq Synthetic

Long-Read, and the Oxford Nanopore technologies entered the market. They are based on single-molecule templates and can sequence longer fragments (between 5000-15000bp, with some reads exceeding 100000 bp).

High throughput sequencing techniques have been widely used to study different '-omes' such as genomes, transcriptomes, proteomes, and metabolomes. Numerous sequencing protocols were developed to address various technical biological questions (*Table 1*). Among these, I focus in this thesis on the RNA- and ChIP-sequencing (Papers II, IV, V). Respective large-scale datasets have been derived, analyzed and applied to address the concrete biological questions. In parallel, research consortia such as ENCODE (ENCODE Project Consortium *et al.*, 2012), TCGA (Cancer Genome Atlas Research Network *et al.*, 2013), GEO (Barrett *et al.*, 2012), FANTOM (Lizio *et al.*, 2015) etc. characterized biological systems at multiple levels by generating omics datasets with various platforms and made them publicly available. For Paper I and Paper III in this thesis, I also integrated datasets from such open-resource projects.



Method	**Description**	**Applications**
RNA-seq	**RNA sequencing**	**mRNA abundance, novel transcript and alternative splicing discovery**
scRNA-seq	Single–cell RNA sequencing	Gene expression profiling of individual cells
CAGE-seq	Cap analysis gene expression sequencing	Identification of TSS and the corresponding promoter regions, mRNA quantification
GRO-seq	Global Run-On Sequencing	Real-time transcription measurement via engaged RNA polymerase
ChIP-seq	**Chromatin Immunoprecipitation Sequencing**	**Protein-DNA interactions, epigenetic status**
DNase-seq	DNase I hypersensitive sites sequencing	Identification of DNA accessible regions
ATAC-seq	Assay for transposase-accessible chromatin using sequencing	Open chromatin site discovery
MeDIP-seq	Methylated DNA immunoprecipitation sequencing	DNA methylation measurement
RIP-seq	RNA-Immunoprecipitation sequencing	RNA-protein interaction discovery
WGBS	Whole–genome Bisulphite sequencing	DNA methylation regions across entire genome
Hi-C	High-throughput sequencing chromosomal conformation	Identification of exact spatial DNA allocation, modeling 3D chromosomal contacts

Table 1: *List of few sequencing approaches developed to identify epigenetic and transcriptional regulation (source: Enseqlopedia).*

2.2 BIOLOGICAL NETWORKS

Networks are indispensable in representing complex relationships between the molecular components of a cell that govern various biological functions. At a conceptual level, a network is a graph (G) where *V* vertices (nodes) represent molecular components (genes, proteins etc.) that are connected by *E* edges, i.e. functional links, between pairs of nodes. The nodes and edges together, thus form a network *G (V, E)*.

There exist numerous versions and representations of species-specific (e.g. human) global networks. Nodes and edges can be differently "colored" according to certain properties shared by subsets of nodes or edges. Depending on the nature of interactions and the purpose of network modeling, the graphs can be directed (when link direction is defined) or undirected (undefined direction). Truly undirected network edges would be physical interactions representing protein complex formation (*Figure 1A*). For comparison, in gene regulatory networks, the information flow should be directed from a transcription factor to the regulated gene (*Figure 1B*). Additionally, edges may have weight attributes specifying either statistical confidence or interaction strength (*Figure 1C*). A common usage of signaling networks is quantitative modeling with e.g. differential equation tools. This bottom-up approach, however, requires precisely knowing all/most of the nodes and edge components in the sub-network, while the latter is very limited in size.

The network modeling can also be classified by specific topological patterns. Bipartite graphs consist of two node sets so that edges are considered only across, i.e. between nodes from the different sets, but not within a set. This approach is used in network enrichment analysis described further in chapter 2.3.2.4 (*Figure 1D*). Alternatively, collections of multiple gene sets, such as Gene Ontology, can be hierarchically ordered into Directed Acyclic Graphs (DAGs) where specificity of the biological function increases down the DAG tree (*Figure 1E*). Contrary to the bipartite graphs mentioned above, nodes in DAGs necessarily overlap, while edges are formed by functional annotations assigned externally (e.g. from experiments) rather than via gene-gene links of the global network. In addition, the feed-forward loop (FFL) (or the three-node motif) is a pattern that is often observed in gene regulatory networks. Two of the FFL nodes represent transcription factors *a* and *b* (*Figure 1F*), regulating each other, whereas both then jointly affect a target gene *c* (R.Milo *et al.*, 2002). An FFL is coherent when the direct effect *a*→*c* is the same as the net indirect effect *a*→*b*→*c,* in which both links can be either positive or negative (*Figure 1G*). Otherwise, it is referred to as an incoherent feed-forward loop (IFFL) (*Figure 1H*).

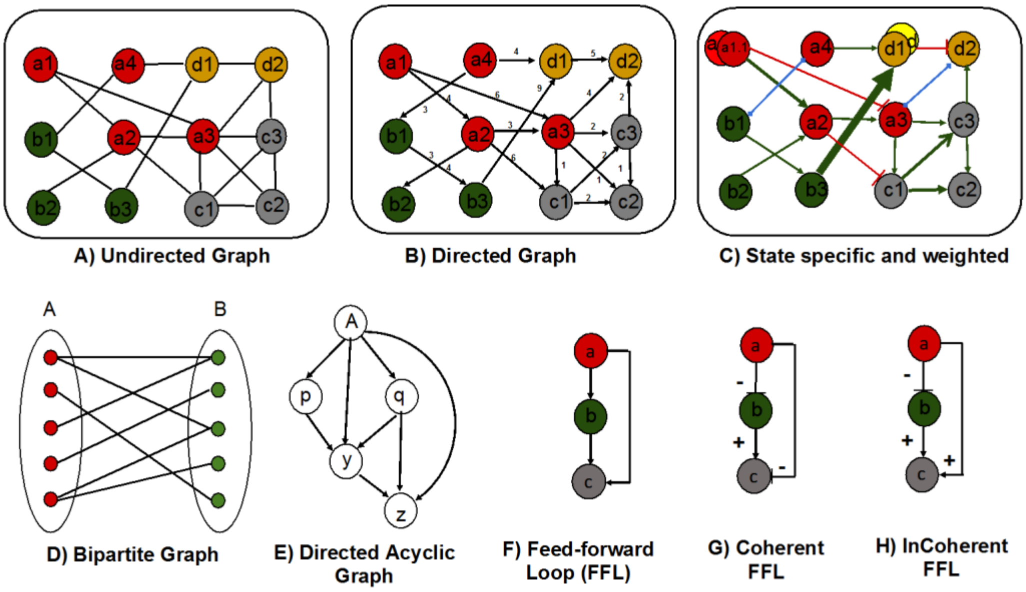


Figure 1: *Types of network representation. Nodes with different attributes are color-coded (red, green, grey, orange) and labeled alphabetically (a, b, c, d). A) Undirected graph B) Directed version of the same graph; edge labels denote confidence scores. C) Directed graph where node state and edge weight attributes are used for color coding: red-repression; green-activation; blue- no effect. D) Bipartite graph of gene nodes where A and B are gene set attributes. E) Directed acyclic graph where the nodes are sets of genes of related function; arrows denote narrowing functional annotation. F) Feed-forward loop G) Coherent feed-forward loop H) Incoherent feed-forward loop.*

2.2.1 Network Topology

The number of links a node has to the other nodes in the network, is called *node degree (k)*. The probability that a randomly chosen node has degree *k* defines the degree distribution *P(k)* of a network. Nodes with high degrees are often referred to as *hubs*. The degree distribution allows to distinguish different network classes. For most of the known versions of the global biological network such as yeast-to-hybrid protein interaction networks (PPIs), networks predicted from data integration (FunCoup) etc., the degree distribution follows the power law, meaning that many nodes in the network have few links and few nodes have many links (i.e. are hubs). Networks possessing this property are termed *scale-free* networks (Barabasi *et al.*, 1999).

$$\boldsymbol{P_{(k)} \sim k^{-\gamma}} \qquad \text{- } \textit{(power-law distribution)}$$

Here γ refers to the degree exponent, with its value for most networks between 2 and 3 ($2 < \gamma < 3$). The justification of why a biological network should have a power-law distribution or a scale-free property was explained by the preferential attachment model (Barabasi *et al.*, 1999). The proposed modeling (growth process) started with a smaller network and expanded it by creating new nodes and connecting them to already existing nodes with a probability proportional to the node degrees of the latter. Thus, highly connected nodes were more likely to gain links to new nodes, which allowed their node degrees to increase faster (i.e., preferential attachment). In the context of PPIs, the phenomenon gene duplication[1] (paralogy), which is wide-spread in the complex eukaryotic genomes, ought to be the key mechanism supporting the proposed model (Barabási *et al.*, 2004). Another notable property of scale-free networks is robustness, i.e. tolerance to attacks. It means that removing a randomly picked node in the scale-free network would have a negligible effect, whereas deleting a hub with high betweenness-centrality[2] would collapse the network topology – and thereby functionality. By simulating attacks on protein nodes in PPIs, one could predict the candidate drug targets (Yu *et al.*, 2007; Azevedo *et al.*, 2015).

2.2.2 Data integration and biological network prediction

The development of high-throughput interaction assays (such as yeast 2-hybrid, co-immunoprecipitation, mass-spectrometry) and curated databases has generated high-quality datasets for a considerable number of organisms – first and foremost for the human and its mammal models. The associations between genes and/or proteins are derived from these datasets. The datasets thus serve information for rather an accurate prediction of edges in the true biological network. Before the edge prediction, most methods convert raw experimental data into quantitative scores. Each type of experimental evidence, such as co-expression, co-occurrence in the subcellular domains, orthologous, phylogenetic profiling, protein domain interactions, literature reports etc. – would require a special association metric (linear correlation, mutual entropy etc.). Further, the association scores are benchmarked, i.e. compared by their prevalence in well known (gold standard, well-annotated) sets of functional associations, such as e.g. KEGG pathway maps or literature-derived interactions versus non-interacting gene pairs. This procedure generates a likelihood space with regard to functional coupling. Finally, a summary likelihood value is reported for each putative edge. STRING reports interaction scores, which report estimated edge confidence given all the available evidence for the gene pair (Szklarczyk *et al.*, 2015). Alternatively in FunCoup, these values, combined in a naïve Bayesian network as log-linear sums of individual (up to 50) likelihood values, are called Final Bayesian scores (FBS) (Alexeyenko *et al.*, 2009; Schmitt *et al.*, 2014).

2.2.3 Biological Network Databases

Using different evidence types, data sets, and prediction algorithms, several resources were developed during the last 10-15 years. *Table 2* summarizes some of the most popular databases, which were collectively used for Paper II in this thesis.

[1] Gene duplication in the genome produces identical proteins that would in the beginning interact with the same protein partners and may later diverge - both functionally and topologically.

[2] Propensity of a node to be found on a shortest path between two randomly picked nodes

Database	Description	Supporting species	Evidences	Algorithm and scores	Reference
STRING Search Tool Retrieval of Interacting Genes/Proteins	Protein-protein interactions containing both physical and functional associations	5090 species	Gene Neighbourhood, gene fusion, gene cooccurrence, co-expression, databases, Text mining	Probabilistic confidence score	https://string-db.org/ (Szklarczyk *et al.*, 2015)
GeneMANIA Multiple Association Network Integration Algorithm	Gene function prediction	6 species (human, mouse, yeast, fly, plant and roundworm)	Genetic and protein interactions, pathways, co-expression, co-localization and domain similarity	Gaussian field label propagation	https://genemania.org/ (Warde-Farley *et al.*, 2010)
FunCoup Functional Coupling	Prediction of functional association between genes/proteins	17 species	Protein-protein interactions, coexpression, subcellular localization, metabolic and signaling interactions, phylogenetic associations	Naive Bayesian integration	http://funcoup.sbc.su.se/ (Alexeyenko *et al.*, 2009; Schmitt *et al.*, 2014)
CORUM Comprehensive resource of mammalian protein complexes	Reference dataset of mammalian protein complex information (complexome)	human, mouse, rat	Highly quality data for protein complex function, localization, subunit composition from individual experiments and literature references	Manual curation	https://mips.helmholtz-muenchen.de/corum/# (Ruepp et al. 2007; Giurgiu et al. 2019;)
BioGrid Biological General Repository for Interaction Datasets	Database access for protein, genetic, chemical interactions	71 species	Gene-protein interactions, Protein-drug interactions, gene-phenotype and gene-gene interactions	Interaction Management Systems	https://thebiogrid.org/ (Stark *et al.*, 2006; Oughtred *et al.*, 2019)
I2D Interlogous Interaction Database	Known and predicted protein-protein interactions	5 species (human, rat, mouse, fly, roundworm)	Domain-domain cooccurrence, gene co-expression, GO terms	Orthology domain co-occurrence, GO similarity	http://ophid.utoronto.ca/ophidv2.204/ (Brown *et al.*, 2005, 2007)
Innate DB Innate immune response database	Interactions and signaling responses involved in mammalian innate immunity	3 species (human, mouse, cattle)	Protein-DNA, protein-protein interactions collected from various databases such as MINT, Intact, BIOGRID, BID, DIP	Manual curation from literature studies. Bovine interactions are predicted via orthology.	https://www.innatedb.co (Breuer *et al.*, 2013)
Pathway Commons	Integrates biological pathway and molecular interaction data from 9 public databases	human	Protein-protein interactions from low-throughput or high-throughput studies aggregated from various resources such as Reactome, NCI Pathways, PhosphoSite, HumanCyc etc.	Manual and computational prediction	http://www.pathwaycommons.org/ (Cerami *et al.*, 2011)

Table 2: *Summary of biological network databases used in paper II (evinet.org)*

2.2.4 Applications of biological networks

Analysis using biological networks have been applied to a number of different topics in life science research. Besides supporting and illustrating experimentally derived results, networks have also been used to study outcomes helping to interpret highly complex systems. Furthermore, the predicted associations have been used as input for sophisticated methods evaluating network properties and setting them within a biological context. Four examples where applications of biological networks have been applied are described below.

Network browsing: retrieval and investigation of small sub-networks

Biological networks are mostly used by the scientific community as a look-up context to find genes/gene sets of interest and identify possible interaction partners. This can help designing experiments, revealing new insight into experimental outcomes or providing additional information for drawing conclusions. Databases like FunCoup and STRING (*Table 2*) have been designed to provide high usability and ease of access to biological networks. Tools like Cytoscape (Shannon *et al.*, 2003) has been developed to provide platforms to further analyze and visualize biological networks (*Figure 2*).

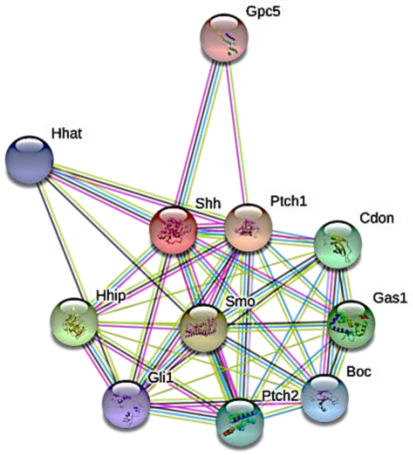


Figure 2*: Network query for 'Shh' gene in STRING database reveals its interaction partners. Links represent the associations between the nodes with various evidence codes.*

Gene and pathway annotations

Network-based pathway annotation uses the networks as additional evidence sources to reveal relations between pathways and gene sets. This form of application is discussed in Section 2.3.2.4.

Discovery and evaluation of novel functional modules

Modules are a group of highly interconnected genes; which are assumed to perform similar functions (Tornow *et al.,* 2003). Distinct algorithms have been developed to identify functional modules in protein interaction networks. For example, methods like MGClus (Merge Gain Clustering) (Frings *et al.*, 2013) determine clusters with a strongly interconnected neighborhood in a given biological network. Weighted correlation network analysis (WGCNA) finds modules of highly correlated genes using the eigengene network methodology (i.e. considering the module eigengene or an intramodular hub gene) in order to relate modules to one another and to external sample traits (Langfelder *et al.*, 2008). DIAMOND uses connectivity significance to identify the full disease module around a set of known disease proteins (Ghiassian *et al.,* 2015).

Gene prioritization with regard of causing a disease or as potential biomarkers

Gene prioritization algorithms assume an input set of known disease (or pathway, or functional term) genes as a reference and attempt to identify related genes by ranking them as candidate members to the given reference set via links in the biological network. These approach in finding associated genes based on the network connectivity is called “guilt-by-association” (GBA) (Erten *et al.*, 2011; Guney *et al.,* 2012; Winter *et al.*, 2012). The paradigm considers network adjacency as evidence of functional relatedness. The common problem with using the GBA approach is that good benchmarking results are usually obtained only for best characterized (i.e. already well known) genes (Gillis *et al.*, 2012). Also, in the absence of additional biological knowledge, GBA leads to prohibitively high false-positive rates. Indeed, each functional set is likely to acquire hundreds of high scoring gene candidates based on network connectivity scores *per se*. In this context, involving independent experimental evidence from dedicated studies allows reducing false discovery rates to acceptable levels (Hong *et al.*, 2010; Reynolds *et al.*, 2010; Bennet *et al.*, 2011)

2.3 BIOLOGICAL PATHWAYS

A biological pathway is a process-oriented set of molecules that interact while performing certain actions in the cell or extracellular space, such as eye sensitivity to direct sunlight or repairing damaged tissue. Any disruptions in these pathways would either be lethal or lead to diseases such as cancer, neural degeneration, or diabetes. According to the National Human Genome Research Institute (NIH 2015), most well-known biological pathways can be classified as involved in metabolism, gene regulation, and signal transduction.

Metabolic pathways: A cascade of biochemical reactions occurring within a cell, where the metabolites are the intermediates in the reaction catalyzed by enzymes, and the product of an enzymatic reaction acts as a substrate for another reaction. In a graph representation, nodes represent biochemical reactions and edges describe compounds driving that reaction. In a metabolic pathway, genes encode enzymes that drive the biochemical reaction and edges refer to the reaction or reaction products.

Gene regulatory pathways: Interactions between transcription factors, DNA and recruitment of co-factors in the cell that directly govern the process of transcription. Also, other proteins such as chromatin remodelers, histone acetyltransferases, histone deacetylases, kinases and methylases are essential for gene regulation.

Signal transduction pathways: A cascade of molecular events, by which a physical or chemical signal is transmitted through a cell to invoke a cellular response. Here one distinguishes between first and second messengers. First messengers are molecules like ligands binding to the cell membrane while second messengers (such as cyclic-AMP, calcium, nitric oxide) are chemical relays which carry out the intracellular signal.

2.3.1 Pathway Databases

A pathway database contains high confidence datasets in a highly organized form for convenient retrieval and usage (Zhang *et al.*, 2012). Currently, *pathguide* (http://www.pathguide.org/), an encyclopedia of pathway databases, reports 702 biological pathways and molecular interaction related online resources. Although regularly updated, the information provided in these resources is by no means complete and may contain false pathway members. In the following section, I discuss the most well-known resources, where functional gene sets (FGS) for network enrichment analysis (evinet.org) are readily available.

Gene Ontology (GO)

GO is an accessible resource, which provides a structured and controlled vocabulary for describing gene and gene product functions. The primary motivation behind the initial efforts of GO consortium was based on the hypothesis that similar genes (paralogs and orthologs) often have conserved functions across species. So, the integration of information from all organisms in one central repository enabled knowledge sharing and inferring functionalities for newly discovered genes. An ontology consists of a set of well-defined terms with well-defined relationships between the terms. In GO, the terms are categorized into three non-overlapping ontologies such as; Biological Process (BP), Molecular Function (MF) and Cellular Component (CC). The biological process refers to a biological objective to which the gene or gene product contributes (e.g., signal transduction, cell growth, and maintenance). Molecular Function defines the biochemical activities of a gene product (e.g. catalysis, transporter activity). Cellular component relates to the location in the cell where a gene product is active (e.g., Nucleus, Cytoplasm) (Ashburner *et al.*, 2000). GO terms are structured as a directed cyclic graph, where nodes represent GO terms and relationships between terms indicates the edges (as mentioned in *Figure 1F*). The initial paralogy approach has developed into broad usage of increasingly available experimental evidence. The 18 codes in GO stand for evidence from literature, experimental, database, or computational methods to date. It is essential to note that most GO terms (98%) are annotated without curators (du Plessis *et al.*, 2011), and are recurrent with child terms in the DAG. Using GO Slims was proposed for reducing the full collection to a few general/universal terms and is optimal for providing a view over the range of functions or processes in a given organism.

Kyoto Encyclopedia of Genes and Genomes (KEGG)

KEGG is a collective database resource consisting of 18 databases dealing with biological pathways, genes and genomes, diseases, drugs and enzymes (Kanehisa *et al.*, 2000). In general, the database provides for the most pathways, a manually curated map which can also include and link to other pathways and is downloadable in XML format. KEGG database is regularly updated and as of May 2019, supporting 5923 species, including the four species (human, mouse, rat and plant) available in evinet.org.

Molecular Signatures Database (MSigDB)

MSigDB is one of the most popular repositories of gene sets initially developed for use with Gene Set Enrichment Analysis (Subramanian *et al.*, 2005) but later employed by many similar approaches. The latest version of MSigDB consists of eight collections (C1-C7, and H), which include genes grouped by their location in the human genome (C1), canonical pathways and experimental signatures curated from publications (C2), genes sharing conserved *cis*-

regulatory motifs up- or downstream of their coding sequences (C3), clusters of genes co-expressed in microarray compendia (C4), genes grouped according to GO categories (C5), signatures of oncogenic pathway activation (C6), and a large collection of immunological conditions (C7). The hallmark gene sets (H) are derived by summarizing many MSigDB gene sets to represent 50 most specific, well-defined and the least overlapping biological processes and in addition displays a coherent expression of member genes in human samples. The MSigDB gene sets are reviewed, curated, and manually annotated by the MSigDB curators (Liberzon *et al.*, 2015). These annotations are available for human only.

Reactome

Reactome is a free, open-source, manually curated and peer-reviewed database providing information about the biological pathway, proteins, reactions, small molecules, and drugs. All shreds of evidence are tracked by primary literature, making the annotations more reliable (Joshi-Tope *et al.*, 2004). It is one of the popular resources, where information from its database is extensively cross-referenced to different online resources such as NCBI Gene, STRING, Pubmed literature database etc. Currently, it supports human along with 15 other non-human species. However, the annotations from human pathways are projected onto other species based on orthology (Reactome, 2019).

WikiPathways

WikiPathways is a resource for biological pathways maintained by and for the scientific community. The idea here is that any registered researcher can browse, create, or edit a pathway, such that the information is useful for other researchers (Pico *et al.*, 2008). Currently, it supports 25 species (including human, mouse, rat and plant) and the database is updated monthly.

Protein Analysis Through Evolutionary Relationships (PANTHER)

PANTHER is an online resource for a comprehensive protein evolutionary and functional classification of proteins (and their genes) (Mi *et al.*, 2019). The classification of proteins is done according to the groups of related evolutionary proteins (at the family level) and those related proteins with similar functions into subfamilies. Like in GO, the proteins are categorized either into molecular function or biological processes based on the functional context, i.e., the function of a protein at the biochemical level (e.g. protein kinase) or the cellular level (e.g. mitosis). PANTHER also comprises Pathways which explicitly specifies the relationships between the interacting molecules. For our functional gene set collection in evinet.org, we utilized the data from PANTHER Pathways for human, which contains information about 177 regulatory and metabolic pathways. In general, the database comprehends information about 132 species.

MetaCyc

MetaCyc is a resource for metabolic pathways involved in primary and secondary metabolism and contains information about associated metabolites, reactions, enzymes and genes. Annotations are highly curated based on literature studies and experimental validations. It is one such resource where the information is linked to other databases such as NCBI-Gene, STRING, KEGG, Interpro etc. Currently, the database contains 2722 pathways and supports 3009 species including human (Caspi *et al.*, 2018).

NetPath

NetPath is a resource of signal transduction pathways. NetPath provides detailed maps of several immune signaling pathways, which include approximately 1600 reactions annotated from the literature and more than 2800 instances of transcriptionally regulated genes, all linked to over 5500 published articles. NetPath allows biomedical scientists to visualize, process and manipulate data about signaling pathways (Kandasamy *et al.*, 2010). The information is only available for human.

Pathway Interaction Database (PID)

PID is a joint initiative of the US National Cancer Institute and Nature Publishing Group. The database provides information about curated and peer-reviewed molecular signaling pathways, regulatory events, and key cellular processes. PID differs from REACTOME, in that it is focused on signaling and regulatory pathways and does not attempt to cover metabolic processes or generic mechanisms such as transcription, translation, etc. (Schaefer *et al.*, 2009). The annotations are available for human only.

Small Molecule Pathway database (SMPdb)

SMPDB contains small molecule pathways found in human. The pathways provide information on human metabolic, drug metabolism, drug action, physiological activity and metabolic disease pathways. The database contains information for human-only (Jewison *et al.*, 2014).

2.4 ENRICHMENT/PATHWAY ANALYSIS

"Imperfect prediction, despite being imperfect, can be valuable for decision-making purposes"

- Michael Kattan

With the onset of omics era, large volumes of biomolecular data are being generated by modern research and screening projects. Typically, this should expose reduced sets of genes or proteins which have changed their state or abundance in a certain experimental or pathological condition. I refer here to such sets as Altered Gene Sets, AGS. Potentially, each of the genes in these lists may acquire a specific role in research or a clinical application. However, considering these genes as a plain list fails to provide mechanistic insights into the underlying biology being studied. On the other hand, selecting only familiar genes tend to leave less characterized ones outside the users' interest. To overcome these challenges and to extract meaningful information from the data, researchers came up with the idea of linking AGS to functionally annotated sets, such as pathways[3] (further referred to as Functional Gene Sets, FGS). An essential advantage of coupling FGS to AGS is that, it is often simpler and more relevant to see relations to known functions presented as a wholesome property of an earlier characterized gene set, than to *de novo* perceive roles of individual gene/proteins or the whole AGS list (Khatri *et al.*, 2012). Such annotation via using functional relations is known as enrichment analysis. The detection of enrichment of a certain functional category is feasible, given the FGS is sufficiently large – according to requirements of the chosen enrichment method.

A plethora of algorithms has been developed in the past years to perform this task. Broadly, they are categorized into three classes: (i) Over-representation analysis on shorter (less-than-complete) lists (ii) Functional class scoring using full gene lists, ranked by a score (iii) methods using network edges – usually topology within known pathways (Huang *et al.*, 2009; Khatri *et al.*, 2012). However, a number of recent methods attempted to integrate information from biological networks into the enrichment analysis, giving rise to a new class called network enrichment analysis (NEA).

2.3.2 Classification of Enrichment Analysis Methods

2.3.2.1 *Over-Representation Analysis (ORA)*

ORA is the state-of-the-art approach, initially developed to analyze microarray datasets. The principle behind this class of tools is to quantify the gene overlap between the AGS and FGS (traditionally from GO, KEGG). The input AGS gene list is obtained by applying certain cutoff

[3] Gene sets whose function is well characterized and available in public databases (as described in chapter 2.3.1) e.g., GO, KEGG

thresholds, for example, a set of differentially expressed (DE) genes. Thereafter, for each pathway (or a functional category) in FGS, input genes that belong to the pathway are counted. The same is repeated with the background gene lists (e.g., non-DE genes). In the end, the assessment is done by applying a statistical test such as Fisher's exact or hypergeometric test) to evaluate the significance of the overlap between two gene sets, which produces a score of enrichment for the given AGS-FGS pair. The most popular tools of this class are DAVID (Huang *et al.*, 2007), BINGO (Maere *et al.*, 2005) and GORILLA (Eden *et al.*, 2009) etc.

2.3.2.2 *Functional Class Scoring (FCS)*

FCS methods accept the full list of genes and are independent of the choice of a cutoff. All FCS methods vary in their framework, but the general principle remains the same (Subramanian *et al.*, 2005; Barbie *et al.*, 2009; Tarca *et al.*, 2012). First, gene-level statistics evaluate the importance/significance of each gene within an experimental condition or an experimental contrast. Second, gene-level statistics measured for all genes that belong to a given pathway are aggregated into the single, cumulative pathway score, which may or may not be assessed for significance with regard to probability of null hypothesis through permutation tests. The most popular method of this class is Gene Set Enrichment Analysis (Subramanian *et al.*, 2005):

Gene Set Enrichment Analysis (GSEA)

In this approach, genes from the experimental data are ranked (*L*) based on their correlation between expression levels of two experimental conditions. Then enrichment score is calculated for a given FGS as a running statistic, i.e. by descending along the ranked list *L*. When the next gene from *L* is a hit (member of the FGS), the score increases and decreases otherwise. The significance of enrichment score is calculated (using Kolmogorov–Smirnov (KS) statistic) in comparison to the permuted data, which is generated by swapping the samples condition labels. Finally, the enrichment scores for each gene set are normalized and the p-values are adjusted for multiple testing.

GSEA is used for analyzing gene rankings from multiple samples with replicates. However, while dealing with a single sample (generally the case with patient samples), this method cannot be applied. Barbie et al. (Barbie *et al.*, 2009) proposed a 'single sample' extension of GSEA (ssGSEA). The method is similar to GSEA, but the gene list is ranked (*L*) by their absolute expression in a single sample from high to low. Next, instead of using the Kolmogorov-Smirnov statistics for estimating the score profiles, the scores are calculated using the Empirical Cumulative Distribution Functions (ECDF) for the genes that belong to the FGS versus the complement (genes that do not belong to FGS). The enrichment scores are calculated for each AGS-FGS pair. The sample sizes normalize enrichment scores and reports significant estimates in the end. The GSEA and ssGSEA methods are benchmarked in Paper III.

2.3.2.3 *Pathway Topology (PT)*

Both ORA and FCS methods consider the number of genes in a pathway or gene correlations to identify the significant pathways and ignores interactions between genes (Khatri *et al.*, 2012). This limitation is addressed by PT methods, which incorporate the topology of the gene network to estimate the significant statistics (Tarca *et al.*, 2009). PT methods primarily rely on detailed maps from pathway databases such as KEGG (Kanehisa *et al.*, 2000) or Reactome (Joshi-Tope *et al.*, 2004). The limitation of PT algorithms is that they only work if a detailed, accurate pathway map is known. Tools such as Impact Factor (IF) (Draghici *et al.*, 2007), SPIA (Tarca *et al.*, 2009) and METACORE (Clarivate Analytics) belongs to this class of methods.

Signaling Pathway Impact Analysis (SPIA)

SPIA (Tarca *et al.*, 2009) measures the pathway perturbation taking into account known intra-pathway topology. Impact analysis is carried out in two steps: calculation of p-values from PNDE (an enrichment analysis of differentially expressed genes in a given pathway, so that either of ORA or FCS methods can be applied) and PPERT, which compares the perturbation factor for a particular gene to that of all other genes in the given pathway. These two channels of evidence are combined into a global probability value, which is then used for pathway ranking and testing the null hypothesis (i.e. the one that the pathway is not significantly perturbed in a given condition). In Paper III, SPIA is benchmarked along with a range of other methods.

2.3.2.4 *Network Enrichment Analysis (NEA)*

A major drawback of ORA and FCS methods is that functional annotations are highly incomplete, this means that the overlap with known pathways is often minimal, resulting in a large number of false negatives (i.e., low coverage). Here, the statistical assumption is that all genes are independent of each other and equally important for the analysis (Subramanian *et al.*, 2005). However as emphasized above, genes within a pathway are known to interact with each other. The PT methods could somewhat improve the situation by using known interactions between genes within a pathway (Khatri *et al.*, 2012). In order to precisely follow the ideas of modern cell biology and systems biology, the pathway-based analysis can be improved by using global networks as an additional information input.

A more advanced network-based approach is to analyze the network cross-connectivity between AGS and FGS, for which methods of network enrichment analysis have been proposed (Alexeyenko *et al.,* 2010, 2012; Glaab *et al.*, 2012; Mccormack *et al.*, 2013; Jeggari *et al.*, 2017). Here, one assumes that a pathway is enriched if a significant number of network edges are found between any genes of AGS and any genes of FGS. The approach is based on the fundamental assumption that the network consists of functional associations between genes of the same types that may be found within a pathway. The performance of these methods depends mainly on two factors. First, it is the quality of the network– if it has low coverage (high false-negative rate of edges) or poor biological relevance (respectively, high false-positive rate), then it will not provide enough statistical power to detect enrichment. Second, it is the relevance of the statistical model, meaning that the ability of a method to distinguish spurious from biologically relevant observations by estimating enrichment significance in an unbiased manner. The NEA methods differ in the method of calculating the enrichment scores, in that it accounts for node degrees of the AGS and FGS genes, and by the amount of computational time required to estimate the null model.

EnrichNet

The EnrichNet method first maps the AGS onto a global network. The network nodes corresponding to the AGS genes are used as seed nodes to compute the network distances between AGS and FGS. The distance scores are calculated between the genes in AGS and the pathways in FGS, using the random walk with restart algorithm. The network-based association score (Xd-score) is relative to the average distance to all pathways and represents a positive or negative deviation from the average distance. The Xd-score is correlated with a classical over-representation scores (q-values) and is presented in a regression plot, for setting an empirical user cutoff (Glaab *et al.,* 2012). While this approach reports the enrichment scores, it does not assess the statistical significance of enrichment.

How to determine the statistical properties of a given biological network?

When determining the statistical properties of a biological network, the choice of an appropriate network null model is indispensable. In network analysis, it was proposed to generate the null (reference) model by randomizing the real biological network through systematically swapping edges between the node pairs (Maslov et al. 2002; McCormack et al. 2013). This has been shown to accurately preserve the node degree distribution, scale-freeness and other first-order topological properties of the network. Patterns (such as an AGS-FGS relation) in the actual network are compared to those in the randomized network to calculate the mean expected number of links and variance, as well as respective p-values.

The reference model is crucial for the given reason. As discussed in chapter 2.2, most biological networks follow the scale-free property. Therefore, a hub gene could have some links to genes of a certain module simply by chance, whereas for a gene with a modest number of connections, the same number of links would indicate a significant biological pattern (Mccormack *et al.*, 2013). Without considering the node degree, it does not make sense to treat network hubs equal to sparsely connected genes. Consecutively, genes with very different node degrees present in AGS and FGS should contribute differently to the enrichment score and its significance.

NEA-2012 and CrosstalkZ

The advantage of the NEA-2012 (Alexeyenko et al. 2012) and Crosstalkz (McCormack et al. 2013) compared to EnrichNet (Glaab *et al.*, 2012) is that the assessment of enrichment significance accounts for the genes' node degrees. Calculation of enrichment scores in both algorithms is done by randomizing the network. Multiple runs of network randomizations are performed by edge rewiring in the original network, by which mean and standard deviation of the number of connecting links is estimated and are further used for calculating the z-scores and p-values. The statistical assumption to calculate the network enrichment score is considered to be normal distributed under "true null", i.e. assumption of no enrichment between the AGS and FGS. Both methods evaluated statistical significance by z-scores but differed in the randomization algorithms. NEA-2012 represented the network as a binary adjacency matrix, which is symmetrical, with margin representing the degree distribution of nodes. The randomization is achieved through permutation of matrix elements. NEA-2012 is capable of processing the fully ranked lists as like FCS, which significantly increases the CPU load and might not be essential to consider full gene lists for most of the applications. Further, Crosstalkz (Mccormack *et al.*, 2013) implemented and benchmarked four network randomization strategies: edge permutation, node permutation, edge assignment and edge assignment with second-order preservation. The edge permutation randomly swapped edges in the network. This alternative is used for the calculation of mean, standard deviation, z, and p as described above. The node permutation swapped node labels between nodes with similar degree. The edge assignment started with an unconnected (empty) network and randomly added nodes according to the node degree distribution of the original network until all the node degrees are recovered. Node degree of the neighbouring nodes is preserved by edge assignment with second-order conservation.

McCormack et al. demonstrated that in this randomization procedure higher-order topological properties, such as the propensity of high-degree nodes to avoid connections with other high-degree nodes, could still be biased. The removal of higher-order topological biases is not always justified and the decision to apply this over-randomization should depend on a particular research question. Also, multiple network randomizations while dealing with bigger networks or large AGS and/or FGS collections are very CPU-intense tasks. Another limitation by these methods (Alexeyenko et al. 2012 and McCormack et al. 2013) is that the test statistic is based on a normal approximation for the reference distribution, which is an integer and non-negative by its nature. Mccormack et al. showed that approximating by the normal distribution would be inaccurate when the expected number of links between AGS and FGS is small – which may lead to high false rates for smaller FGS. These limitations were taken into account and addressed in the new implementation of NEArender.

NEArender

The NEArender algorithm evaluates the enrichment statistics based on the assumption that the underlying network edges are binomial distribution. The calculations are briefed in the Methods and discussed in Paper I.

2.4 MECHANISMS OF DEVELOPING NERVOUS SYSTEM

The vertebrate central nervous system (CNS) is the most complex and intriguing tissue in the human body, comprising hundreds or even thousands of functionally distinct types of neurons that establish selective synaptic connectivity with each other and to form functional neural circuits. The diversity of neurons is identifiable by their morphology, functional properties, connectivity, mode of neurotransmission and gene expression patterns (Osseward *et al.*, 2019). The different neuronal subtypes are generated from mitotically active immature neural stem cells (NSCs) at specific positions and at defined time points during development of the neural tube, i.e. embryonic anlagen of the brain and spinal cord. The CNS also contains oligodendrocytes and astrocytes, which are non-neuronal support cells that are generated subsequent to neurons in the developing CNS. Oligodendrocytes form myelin sheets that enwrap and insulate axonal processes that facilitate the rapid conduction of electrical impulses. Oligodendrocytes also produce trophic growth factors that support neuronal survival. Astrocytes, in turn, provide structural support, modulate synaptic activity and contribute to maintain the blood-brain barrier (Rowitch *et al.*, 2010).

2.4.1 Neural Induction

The vertebrate CNS is derived from the dorsal ectoderm at gastrula stages of the early developing embryo and involves activation or deactivation of several signaling pathways, including FGFs, Wnts and BMPs (Claudio D. Stern, 2005). The neural plate initially forms an epithelial sheet that starts to invaginate and its dorsal edges merge and detach from the ectoderm to form the neural tube (*Figure 3A*). The neural tube which is located below the epidermis on the “back-side” of the developing embryo. The neural tube is subdivided along two major axes, the anterior-posterior (AP) axis (also termed rostro-caudal) that runs from the head to tail, and the dorsal-ventral (DV) axis from the back to belly (*Figure 3B*). The anterior part of the neural plate/neural tube, which is fated to give rise to a different subdivision of the brain is specified first, while the caudal spinal cord tissue is progressively added along with the caudal extension of the body axis (Sasai *et al.*, 2014). Already at the early neural plate stage, NSCs at different positions are exposed to different patterning signals that impose unique positional identities to cells along the AP- and DV-axis of the neural tube, and enables cells to differentiate into distinct subtypes of neurons when NSCs leave the cell cycle and differentiate into post-mitotic neurons. Neurogenesis, in turn, involves the activity of Notch signaling pathway and the activity of a family of basic helix-loop-helix (bHLH) transcription factors. This activity further regulates the balance of cells to choose either between undifferentiated neural precursors or differentiating into neurons (Louvi *et al.*, 2006). Thus, mechanisms that mediate positional patterning of NSCs are integrated with pan-neuronal factors regulating neurogenesis.

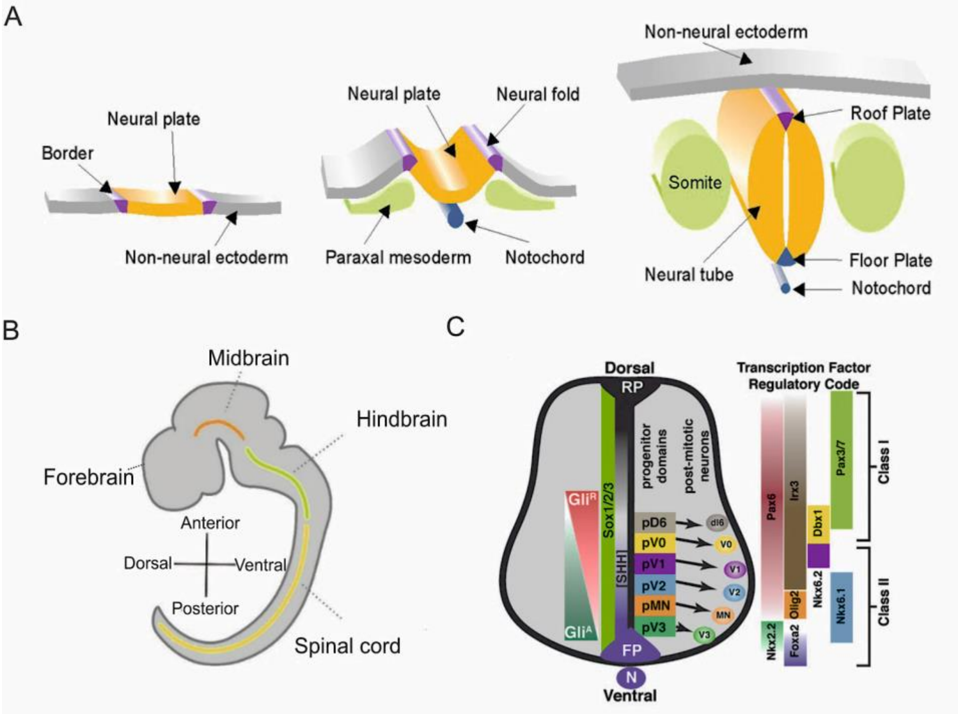


***Figure 3**: A) Formation of the neural tube. Neural progenitors originate from ectoderm as a sheet of cells called neural plate. During development the axial and paraxial mesoderm cells, develops into notochord and somite, causing the folding of the neural tube (Adapted from Sanja Kurdija). B) Mouse embryo where large subdivisions of the CNS are indicated. Anterior-Posterior (AP) patterning and Dorso-Ventral (DV) patterning. C) Dorso-ventral patterning of the neural tube. Distinct progenitor domains are formed along the DV axis. Transcriptional regulatory code of class I (repressed) and class II (induced) TFs. Adapted from (Peterson et al., 2012)*

2.4.2 Anterior-Posterior patterning of the neural tube

Patterning of cells along the AP axis is initiated at early stages and underlies the regionalization of the neural tube into forebrain (FB), midbrain (MB), hindbrain (HB) and spinal cord (SC) regional territories (*Figure 3B*). After this, secondary signaling centers are established at specific boundaries of gene expression that enables further diversification (or patterning) of cells along the AP-axis (Wurst *et al.*, 2001; Lim *et al.*, 2007). During neural induction, the newly formed NSCs acquire a FB identity by default unless exposed to signals that impose a progressively more posterior identity to NSCs (MB, HB and SC, respectively). A dynamic and combination of spatiotemporal signaling by FGF from the caudally regressing node; Wnt and retinoic acid (RA) signaling from the paraxial mesoderm or the somite induces progressively caudal MB, HB and SC-identities, and with an increased requirement for prolonged exposure or higher concentration of Wnt signaling induces caudal-most spinal cord identities (Nordström *et al.*, 2006).

2.4.3 Dorsal-Ventral patterning of the neural tube

The neural tube is also patterned along the DV-axis in response to locally acting morphogen signals emerging from signaling centers located at the dorsal and ventral extremes of the neural tube, which diffuse and form opposing concentration-gradients along the DV-axis (*Figure 3C).* In the dorsal neural tube, BMP signals initially produced by the ectoderm abutting the dorsal edge of the neural plate and later by roof plate cells at the dorsal midline of the neural tube diffuse into the dorsal neural tube and induce the specification of different subtypes of dorsal interneurons depending on concentration and/or time of BMP exposure, in part through local induction of bHLH-type or homeodomain (HD) transcription factor proteins that promote different dorsal fates of differentiation (Liem *et al.*, 1997; Chesnutt *et al.*, 2004; Louvi *et al.*, 2006). Wnt signaling has also been implicated in patterning and growth of cells in the dorsal neural tube (Ulloa *et al.*, 2009).

In the ventral neural tube, cells are patterned in response to the graded morphogen activity of Shh, initially produced by the notochord underlying the forming neural tube and later by the floor plate (FP) induced at the ventral midline of the neural tube (Roelink *et al.*, 1995). Apart from inducing the FP, the graded activity of Shh accounts for the induction of five cardinal progenitor domains, each generating different classes of motor neurons (MNs) or subtypes of interneurons (Ericson *et al.*, 1997; Wijgerde *et al.*, 2002). Patterning of cells in the ventral neural tube by graded Shh signaling is a paradigm form of morphogen signaling in neural pattern formation (*Figure 3C).* Below I discuss the Shh pathway and regulation of the downstream network that translates the graded information of Shh into discrete ventral progenitor domains downstream of Shh in further detail.

Together, patterning of cells along the DV-axis of the neural tube engage at two opposing signaling activities that also seems to intersect with each other, as Shh-induced ventral domains extend dorsally when BMP signaling is inhibited (Liem *et al.*, 2000), while ventral fates are suppressed in response to ectopic BMP or Wnt signaling (Liem *et al.*, 1997; Ulloa *et al.*, 2009). In addition to these signaling activities, RA signaling by somites promotes intermediate cell fates (Pierani *et al.*, 1999) possibly by counteracting the ventral Shh and dorsal BMP gradients at intermediate positions of the neural tube (Oosterveen *et al.*, 2013).

2.4.4 The Shh signaling pathway

Shh signaling functions as a true morphogen[4] in the patterning of the ventral neural tube, but an important question remains how the extracellular concentration gradient of the Shh signaling is translated into the intracellular specification of multiple cell identities?

[4] Signaling molecule that acts directly on the cells to produce specific cellular response depending on its concentration gradient.

The Shh receptor/transduction complex consists of two transmembrane proteins including Patched1 (Ptc1), to which Shh binds, and Smoothened (Smo) which initiates the intracellular Shh signaling cascade. Ptc1 regulates the activity of Smo. In the absence of Shh, Ptc1 is localized in the cilium[5] and prevents the accumulation of Smo in this structure. Conversely, binding of Shh to Ptc1 results in the removal of Ptc1 from the cilium and concomitant accumulation of Smo (Corbit *et al.*, 2005; Haycraft *et al.*, 2005; Rohatgi *et al.*, 2007). Ultimately, Smo regulates the activity of bifunctional Gli (Gli1-3) transcription factors. In the absence of Shh, Gli2 and Gli3 proteins are processed to generate a transcriptional repressor (GliR). While in the presence of Shh, both proteins are stabilized in their full-length activator form (GliA). Gli1 only works as an activator and is induced by Gli2 and Gli3 downstream of Shh (*Figure 4*). As a result, the extracellular gradient of Shh is translated into opposing intracellular gradients of Gli activator and Gli repressor activities along the DV axis of the neural tube (Fuccillo *et al.*, 2006).

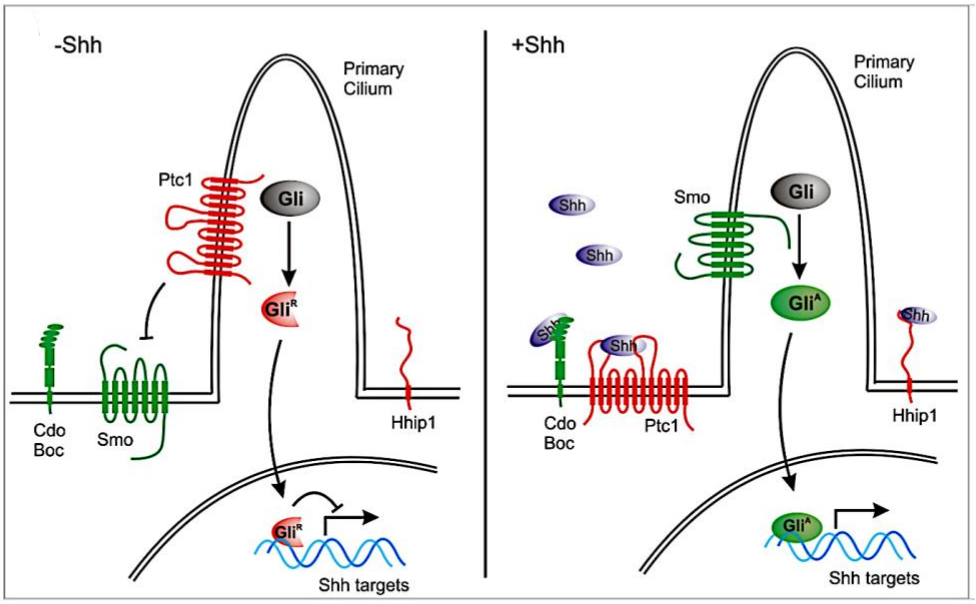


***Figure 4**: Schematic illustration of Shh signaling pathway A) In the absence of Shh, Ptc1 localizes to the cilium and inhibits the activity of Smo in the cilium. Under these conditions, Gli proteins are completely degraded or truncated into repressor form (GliR) that translocate to the nucleus and inhibits the transcription of target genes. Binding of Shh to Ptc1 releases inhibition of Smo. Ptc1 is removed from the cilium with the concomitant ciliary accumulation of Smo. The activation of Smo inhibits proteolytic processing of Gli proteins resulting in the accumulation of GliA (activator form), which translocates to the nucleus and activates the target genes (Adapted from JM Dias).*

Graded Shh signaling regulates the regional expression of a group of transcription factors, that are characterized by the presence of homeodomain (HD) DNA binding motifs or

[5] Cilia are extensions of cell membrane that contain a core microtubule structure and exhibit intra-flagellar transport. Primary cilia play an important role in the Shh signal transduction.

bHLH sequences. Class I genes such as *Pax6, Irx3, Dbx1* and *Dbx2* are repressed, while class II genes, including *Foxa2, Nkx2.2, Nkx2.9, Olig2, Nkx6.1* and *Nkx6.2,* are induced by Shh-Gli signaling (*Figure 3C*). Different thresholds of Shh signaling are required for the repression or activation of individual class I and class II genes, resulting in a nested expression pattern of these genes along the DV axis in the spinal cord (or hindbrain). Furthermore, class I and class II genes cross-repress each other to establish the p0, p1, p2, pMN, p3 and FP domains along the neural tube. In turn, each progenitor domain gives rise to distinct neuronal (V0, V1, V2, V3 interneurons and MNs) and non-neuronal FP subtypes (*Figure 3C*) (Briscoe *et al.*, 2000; Jessell, 2000; Dessaud *et al.*, 2008).

2.4.5 Sequential specification of motor neurons and serotonergic neurons by Nkx2.2^{+} NSCs in the ventral hindbrain

A pool of NSCs in the p3 domain, which is located dorsal to the FP cells expresses the homeodomain TF Nkx2.2 in response to high morphogen gradients of Shh signaling. During development, this progenitor domain sequentially generates visceral motor neurons (vMNs), serotonergic neurons (5HTNs) and oligodendrocyte precursors (OLPs) (Pattyn *et al.*, 2003; Vallstedt *et al.*, 2005). During the period of vMN neurogenesis, these progenitor cells express the paired homeobox-like TF Phox2b which acts as a temporal effector protein. The ON/OFF-status of Phox2b expression determines whether young Nkx2.2^{+} NSCs will select early MN-fate or if cells should terminate MN-production and begin to generate late-born 5HTNs respectively. Additionally, Tgfβ2 acts as an important temporal switch signal that triggers the vMN-to-5HTN fate switch by suppressing Phox2b expression and imposes “age” upon cells by constraining their developmental potential (Pattyn *et al.*, 2003; Dias *et al.*, 2014). In this process, Shh signaling induces Phox2b and Tgfβ2, but Tgfβ2 is induced with a temporal delay relative to Phox2b (Dias *et al.*, 2014). However, it is unclear about the molecular mechanism underlying the late onset of Tgfβ2 and how the activity of Shh and Tgfβ2 pathways may be functionally interconnected in this temporal differentiation process.

2.4.6 Source of material for temporal lineage studies

In vivo studies in model organisms provide an important basis of our understanding of cell fate determination in the CNS. However, quantitative limitations (very few cells of the desired type) prohibit high-throughput characterization of these processes. On the other hand, Embryonic Stem Cells (ESC) can differentiate into any cell type and provide an unlimited resource. ESCs are derived from the inner cell mass (ICM) of preimplantation embryos after the formation of a blastocyst. Further, these cells can be removed from its normal embryonic environment and cultured under appropriate conditions to generate ESCs, which can differentiate to all cell lineages (Nishikawa *et al.*, 2007).

The establishment of FP, p3 domains and the sequential specification of MNs, 5HTNs and OLPs by Nkx2.2$^+$ neural progenitors can be recapitulated in mouse ESC cultures by transiently exposing differentiating cells to all-trans retinoic acid (RA) and the Shh agonist Hh-Ag1.3 (SAG). The temporal transitions of specification states and potency of ESC-derived Nkx2.2$^+$ neural progenitors in this process have been well defined (Dias *et al.*, 2014). This *in vitro* differentiation paradigm provides an unlimited source of material to apply for RNA and ChIP-sequencing.

3 METHODS CONSIDERED

3.1 RNA SEQUENCING

RNA sequencing (RNA-seq) is an experimental technique that uses high throughput sequencing to reveal the presence and quantity of RNA in a biological sample at a given time-point. With nearly ~15000 references in PubMed, it is one of the most cited high-throughput sequencing methods. Several groups first published the method in 2008 (Marioni *et al.*, 2008; Mortazavi *et al.*, 2008; Nagalakshmi *et al.*, 2008; Sultan *et al.*, 2008; Wilhelm *et al.*, 2008). Before RNA-seq, the transcriptomics field was first revolutionized by the microarray technology. However, the latter turned out to be hampered by complex normalization procedures and limitations in detecting low abundance transcripts. Later on, digital transcript-counting approaches such as Serial Analysis of Gene Expression (SAGE) (Velculescu *et al.*, 1995), massively parallel signature sequencing (MPSS) (Harbers *et al.*, 2005) overcame to a certain extent inherent limitations of the arrays, but were expensive and could not be used for differential splicing analysis (Sultan *et al.*, 2008).

RNA-seq aims at sequencing the RNA content of cells. One crucial aspect of the experimental design is the RNA extraction protocol itself. In a cell, ribosomal RNA (rRNA) constitutes up to 90% of total RNA, while containing only 1-2% of messenger RNA, that we are interested in for gene expression quantification. The alternative, depending on the targeted study, is between whether to enrich for the mRNA fraction using poly-A selection or to deplete rRNA. For example, the polyA strategy is based on the fact that mRNA almost always contains a polyA tail that can be used to assay protein-coding transcripts only (Griffith *et al.*, 2015). After RNA extraction, the assessment of RNA quality is determined by the RNA Integrity Number (RIN), a measure of RNA degradation in the sample (Schroeder *et al.*, 2006). Typically, RIN scores ≥ 8 are considered good enough for transcriptome analysis. The RNA quality from RIN scores and the total amount of starting RNA are considered to be critical for subsequent steps such as library preparation, sequencing and computational analysis.

The next step is the library preparation, where mRNAs are fragmented to smaller pieces of RNA to enable sequencing. Currently, all the available sequencing technologies require a DNA template for sequencing (e.g.: Illumina protocol rely on “sequencing by synthesis”). Therefore, it is not possible to directly sequence RNA transcripts with the same protocols. RNA-Seq method relies on an enzyme called reverse transcriptase (mainly found in viruses), to reverse-transcribe RNA into cDNA. The cDNA is fragmented and the fragments shorter than 500 bp are selected. Then platform-specific adapter sequences, DNA barcodes and amplifying the DNA for sequencing conclude this step. Adding DNA barcodes to the ends of each fragment enables multiplexing. It means various libraries are pooled and sequenced simultaneously in a single sequencing run, facilitating efficient use of DNA sequencing machine. Library concentration is assessed using Bioanalyzer or qRT-PCR. Finally, the cDNA library is ready for sequencing, usually performed in a sequencing core facility or by a service provider.

Depending upon the individual research requirements, various NGS protocols are chosen to sequence the sample libraries. This includes factors such as sequencing depth, read-length, single-end, or paired-end reads, and the choice depends on the biological questions one is aiming to address from sequencing data. For example, the sequencing depth of 5-million mapped reads are sufficient to quantify accurately medium to highly expressed genes, whereas up to 100-million mapped reads to precisely quantify low expressed genes and transcripts. The cheaper and short single-end reads are generally sufficient for studying gene-expression levels with available reference genomes – whereas expensive and longer paired-end reads are preferable for de-novo transcript discovery or isoform expression analysis or in-case of poorly annotated transcriptomes (Sims *et al.*, 2014). In my experience from dealing with both single and paired-end sequencing data, single-end read sequencing is more cost-effective, but while dealing with PCR duplicates or short reads, which often mapped to multiple locations, it turned to generate spurious hits. On the other end, paired-end reads provided a more precise mapping to the reference genome, which reduced the number of multi-mapped reads. It is much easier to identify PCR duplicate reads, as fragment length and fragment locations of two paired ends can be easily estimated/calculated.

3.1.1 Computational processing of sequencing read data

The sequencing of cDNA fragments from the DNA sequencer machine produces millions of reads. The reads pooled during the library preparation step are now separated, according to sample origin in our case – using the tagged DNA barcodes (demultiplexing). The demultiplexed reads are then aligned to the reference genome/transcriptome. Since only mRNAs are extracted with a polyA tail, the aligner tries to map the majority of the reads to exonic regions of the reference genome. Therefore, specialized algorithms that are capable of handling splice junctions are used for this purpose. The widely used and publicly available programs/tools are HISAT (Kim *et al.*, 2015), Tophat (Trapnell *et al.*, 2009) and STAR (Dobin *et al.*, 2013). It is good to assess the quality of sequenced samples before investing time in the data post-processing. Some of the important quality measures include: a) Phred-scores (Q), for measuring the accuracy of DNA bases generated by sequencing machine; usually Q>30 is considered for good-quality reads b) presence of adapter or biological sample contamination c) over-amplification of PCR fragments (duplicates) (Patel *et al.*, 2012). One of the most commonly used tools to perform initial quality measurement on sequenced data is FASTQC (Andrews S. (2010)). Another direct judgement could be by visualizing the distribution of aligned reads to the reference genome in a genome browser (e.g., IGV). It gives the clear distinction of reading alignments across exons, intron and intergenic regions which can tell about pre-mRNA fraction (more reads in introns). Hereafter, making sense of RNA-Seq data depends on the research question. For example, RNA-seq data can be used for the detection and quantification of alternative splicing (Wang *et al.*, 2008) and estimation of total RNA abundance in the cell (Gaidatzis *et al.*, 2015). However, the primary objective in most of the biological studies is profiling, in order to determine differential gene expression between biological samples.

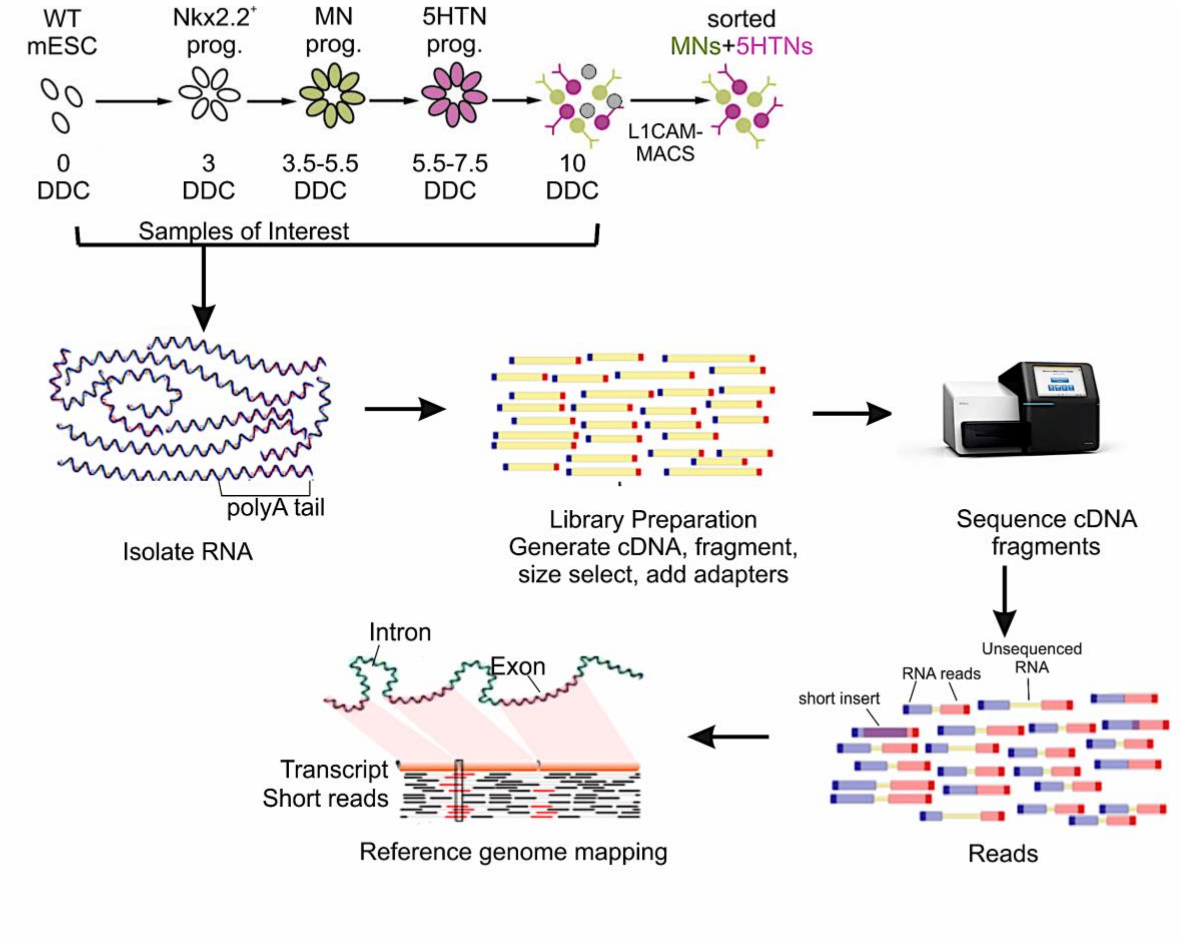


***Figure 5**: Major steps followed for performing RNA-seq. Figure adapted from (Griffith et al., 2015)*

Once the reads are aligned to the reference genome, the task is to summarize and aggregate reads over biological units (such as exons, transcripts or genes). The most common approach is to count the number of reads overlapping the exons in a gene (i.e. quantify as read counts) (Anders *et al.*, 2014). Since the number of reads observed for a given gene is proportional not just to the gene expression level but also to its gene length and the sequencing depth of the library, dividing each read count by the corresponding library size (in millions) yields counts per million (CPM) (Chen *et al.*, 2014) gives a simple measure of reading abundance that can be compared across different sizes. Standardizing further by gene length (in kilobases) gives a metric called FPKM (Fragments per kilobase and million mapped reads) (Trapnell *et al.*, 2010) (*Figure 7*).

The calculated metric values (x=counts or CPMs or FPKMs) convey the abundance of a gene/transcript in a particular experimental condition ($x>0$). However, performing differential expression (DE) analysis, i.e. identifying the gene expression changes across experimental conditions, one should employ statistical tools. This means taking a table of summarized count data and performing statistical testing between samples of interest. While fitting the correct statistical model to the data is an essential step before making inferences about differentially expressed genes. Earlier with microarrays, the intensities values were log-transformed and analyzed as normally distributed random variables. The problem with the RNA-seq data is that the read counts possess unequal variance even upon log-transformation. Namely, count values positively correlate with variance, which introduces bias into DE significance estimates, such as p-values. To overcome this problem, accurate statistical modeling (or complete elimination) of the mean-variance relationship is the key to design a powerful method of analysis. Many models based on the negative-binomial distribution(Anders *et al.*, 2013), generalized linear models (Smyth *et al.*, 2014) has been presented for RNA-seq data.

In our work, we used `voom` transformation from the limma R package. In this approach, the normalized read counts are also log-transformed considering sequencing depths (log CPM values as discussed above). The authors demonstrated that the mean-variance relationship is often rather complex and therefore suggested fitting the joint distribution with a non-linear function, which is then used to compute gene-wise weights for each observation (since the variance is modeled at the observational level, this method is dubbed as voom). Further, the differential expression was estimated with limma functions such as `lmFit` (to fit the linear model), `eBayes` (to compute moderated t-statistics for the hypothesis that the log2-fold change is zero), and `topTable` (utility function to summarize `eBayes` fit).

In comparison with other DE methods, this approach requires at least three samples per condition to gain sufficient power to detect DE genes and is relatively unaffected by outliers (Soneson *et al.*, 2013). To obtain good statistics, having several biological replicates (multiple samples that come from the same type of cells) for each experimental condition plays a vital role. RNA-seq cost has been lately drastically reduced, while experimental designs constrained to one or two replicates would impose difficulties in applying the statistical models. However in clinical cohorts, patient samples are rarely available and are usually represented with single samples (Cancer Genome Atlas Research Network, 2011). Although such designs usually end up in a regression-based estimation (where patients serve single points in the regression model), I also present an approach (Paper I) that can increase the robustness of non-replicated analyses of differential gene expression.

3.2 CHIP SEQUENCING

Chromatin (DNA and associated proteins) Immuno (use of antibodies to target specific protein) Precipitation (enrichment assay, where the total pool of chromatin is enriched only with the protein of interest) followed by high-throughput sequencing (ChIP-seq).

ChIP-seq is a powerful experimental approach to identify the genome-wide binding sites of protein-DNA interactions. With the appropriate antibodies, this technique can be used to locate transcription factors binding to specific DNA sites or to capture the chromatin states involved in transcription regulation, notably histone proteins undergoing modifications, such as methylation and acetylation (Solomon *et al.*, 1988; Barski *et al.*, 2007; Johnson *et al.*, 2007) (*Figure 6*).

The first step is to extract the chromatin from the cells under the experimental condition. To preserve the DNA-protein interactions the cells are fixed with formaldehyde, followed by cell lysis and fragmentation of the DNA to ~300-600 bp, using sonication. A part of the chromatin material is used as input, with the fixation reversed. The rest of the chromatin is incubated with magnetic beads conjugated with a specific antibody to immunoprecipitate the target protein, whereas the non-captured DNA is eluted. Furthermore, the protein-DNA complexes are reverse linked by incubation at 65°C for at least 5 hours. As just DNA is required for sequencing, RNA and proteins are digested with RNases and proteinase respectively. Subsequently, ChIP-seq libraries are constructed from the purified and precipitated DNA (*Figure 6*). The sequencing of sample libraries proceeds in the same way as discussed for RNA-seq.

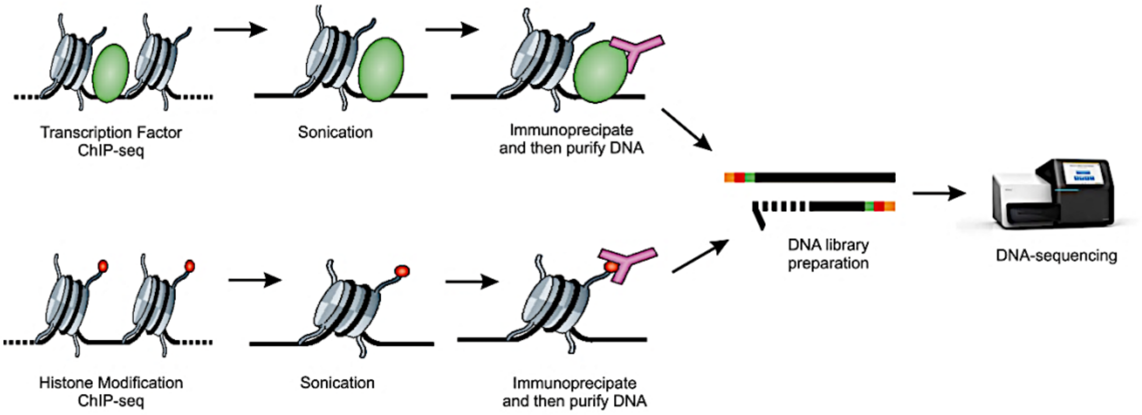

Transcription Factor ChIP-seq
Sonication
Immunoprecipitate and then purify DNA
DNA library preparation
DNA-sequencing
Histone Modification ChIP-seq
Sonication
Immunoprecipitate and then purify DNA

***Figure 6**: Major steps followed for performing transcription factor and histone modifications ChIP-sequencing*

3.2.1 Analyzing genome-wide ChIP seq data

To standardize ChIP-seq experiments and bioinformatics data analyses, ENCODE and modENCODE consortia have developed guidelines (Landt *et al.*, 2012; Marinov *et al.*, 2014). Unlike RNA-Seq, simpler and faster aligner programs such as Bowtie (Langmead *et al.*, 2009) or BWA (Li *et al.*, 2009) can be used to align chip sequenced short reads (~35-100bp) to the reference genome. Some genomic regions exhibit a higher density of mapped reads. This pileup of reads in the enriched regions is referred to as "*peaks*". Depending on the type of interactions, these peaks either might appear either 'narrow' or 'broad'. For example, transcription factor proteins bind to the DNA at specific locations (4-24bp) depending on the sequence motifs and are thus relatively narrow. On the contrary, histone marks (*Table 3*) usually present on nucleosomes covering long stretches of DNA (~100 to 10000bp), corresponding to broader peaks. Different algorithms can achieve the identification of these peak regions (a.k.a. peak calling). I have used MACS2 (Zhang *et al.*, 2008; Feng *et al.*, 2012) with customized parameter settings to call narrower or broader peaks. The reproducibility between biological replicates is assessed by the IDR method (Li *et al.*, 2011). Various downstream statistical and bioinformatics analysis can be applied after the identification of peaks, such as motif analysis to construct putative transcription factor binding sites (Heinz *et al.*, 2010; Machanick *et al.*, 2011) or identification of genes associated with peaks (Zhu *et al.*, 2010). The most common way to represent the chip-seq analysis results is to generate the signal intensity profiles along the genome (Ramírez *et al.*, 2014), which could then be visualized through genome browsers such as by Santa Cruz UCSC browser (Karolchik *et al.*, 2009) or IGV (*Figure 7*).

Histone Mark	**Description**
H3K4me3	*Active Transcription*
H3K27ac	*Active Enhancer*
H3K4me1	*Enhancer*
H3K27me3	*Repression*

***Table 3**: Histone Modifications included in this study*

3.2.2 Identification of chromatin states

Mapping of epigenetic marks such as histone modifications provides a powerful tool for genome annotation, for identifying presumptive regulatory regions, and/or estimate their current activity. Similarly to the TF ChIP-seq approach, individual histone marks can be studied in isolation through the identification of narrow or broader peaks (Zhang *et al.*, 2008; Feng *et al.*, 2012). However, additional information can be gained by summarizing the combinatorial pattern of multiple histone marks. Such pattern identification known as ‘chromatin states’ detection allows capturing known classes of genomic elements such as promoters, enhancers, transcribed and repressed regions etc. Methods using Hidden Markov Models (HMM) have been developed in recent years in order to infer combinatorial states expressed as patterns of presence/absence of underlying histone modifications (Ernst *et al.*, 2017). The java command-line version of the ChromHMM program was used to analyze the histone modifications datasets. *Table 3* summarizes the histone marks considered for Paper IV.

Sequencing Analysis Pipeline

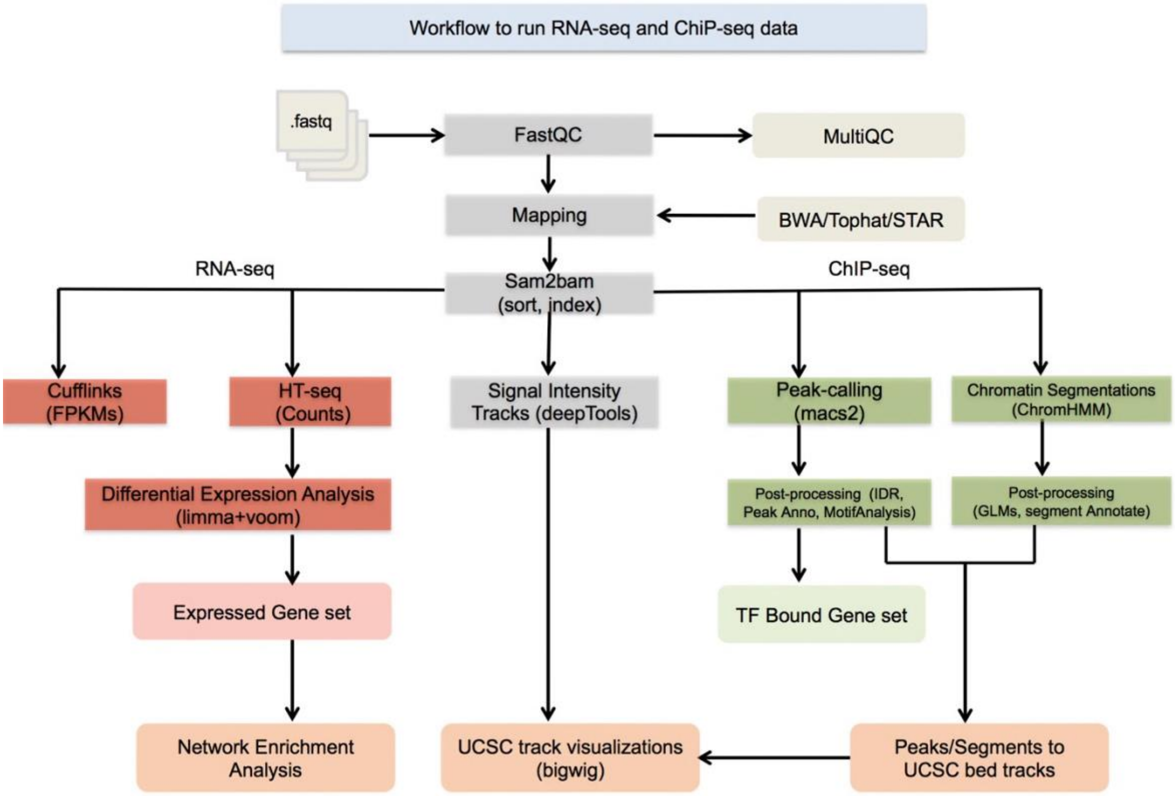


***Figure 7**: Workflow of RNA-seq and ChIP-seq data analyses.*

3.3 NEArender

NEArender tool was implemented as an R package (Jeggari *et al.*, 2017) and integrated the core NEA functionality with various auxiliary functions for preparation and evaluation of the input components as well as output interpretation. Three following input components are required by NEArender package in order to calculate network enrichment statistics.

3.3.1 Altered Gene Set (AGS)

AGS are sets of genes derived from a particular current experiment (or a pathological condition) of which function has yet to be characterized. The number of genes included in each AGS would be either data-driven (inclusion of all significant genes) or pre-defined by the user (listing N top-ranking ones regardless of significance). In the package, function `sample2ags` can create AGSs from sample columns of an R matrix using one of the five available algorithms described in (Jeggari *et al.*, 2017). To deal with mutational datasets (usually data is represented as binary, Wildtype vs. Mutant), a special function `mutations2ags` is implemented to allow direct creation of AGS as full sample-specific sets of mutated genes.

3.3.2 Functional Gene Set (FGS)

FGS is a set of genes whose common function is already characterized and is provided by either a knowledge-driven database, or literature, or by a custom expert curation. Any of the public knowledge resources discussed in chapter 2.3.1 can be considered. NEArender can import a collection of multiple FGS from a TAB-delimited file listing all members of each functional category using function `import.gs`.

3.3.3 Network (NW)

NW represents the global network of functional associations between genes/proteins. Any of the networks from chapter 2.2.3 can be utilized, as well as a new, custom network submitted as a network file. The latter can be imported using function `import.net`. We note that the parametric calculation chosen by NEArender for the reasons of speed and absence of bias from smaller gene sets might produce correct estimates of the null-model AGS-FGS connectivity only in scale-free networks. Therefore, in order to evaluate the topology of a chosen network for scale-freeness and second-order dependencies, R package is equipped with auxiliary functions `connectivity` and `topology2nd` respectively. To test the ability of a chosen network to perform well in a NEA, the function `benchmark` is employed. It implements 'guilty-by-association' approach, i.e. tests networks by their ability to convey membership of known FGS via network enrichment statistics. It systematically executes a series of multiple individual tests for each member of FGS (as positive cases) and other nodes of NW (randomly picked genes with node connectivity values matching to the FGS members – as negative cases). For each gene, the procedure tests the null hypothesis of the gene not being an FGS member. This procedure produces true positive (TP), false positives (FP), true negatives (TN) and false negative (FN) rates if the positive/negative gene cases receive a NEA score above or below a certain threshold, respectively. The counts of alternative test outcomes TN, FP, TN, FP allow plotting of a ROC curve with function `roc`, by applying sequential cutoff values of the NEA z-statistic.

3.3.4 Calculation of network enrichment scores

Upon the above three input components prepared, network enrichment analysis can be run using the core function `nea.render`. The enrichment statistics is based on the chi-square statistics estimated with a binomial (i.e. 2-class) formula, where AGS-FGS enrichment was compared against the reference, i.e. the "nonAGS-nonFGS" connectivity (Equation 1). We note that this approach is focused on considering only AGS-FGS relations and therefore ignores indirect links (i.e. network paths of length>1) and cannot evaluate higher-order topologies as well as not capable of estimating other popular network statistics.

$$\chi^2 = \frac{(n_{\text{AGS-FGS}} - \hat{n}_{\text{AGS-FGS}})^2}{\hat{n}_{\text{AGS-FGS}}} + \frac{(!n_{\text{AGS-FGS}} - !\hat{n}_{\text{AGS-FGS}})^2}{!\hat{n}_{\text{AGS-FGS}}} \quad \text{--------------------------} \; (1)$$

where !n means "complement to n", i.e. all global network edges that did not belong to $N_{AGS-FGS}$. The number of links expected under the true null, i.e. by chance, was determined by:

$$\hat{n}_{\text{AGS-FGS}} = \frac{N_{AGS} * N_{FGS}}{2 * N_{total}}$$

N_{AGS} and N_{FGS} refer to the sums of connectivity values (node degrees) of member nodes of AGS and FGS, respectively, $N_{AGS-FGS}$ is the number of network edges between AGS and FGS genes, and N_{total} is the number of edges in the whole network.

Since `nea.render` accepts multiple AGSs and multiple FGSs as input vectors, its output is matrices (outer products) of the relevant statistics: the chi-square score, the corresponding normally distributed z-score (preferred for linear modeling in downstream analyses), the p- and q (p-value adjusted for multiple testing) values as well as the numbers of links connecting AGS and FGS. Function `nea.render` can engage multiple CPUs by using R package 'parallel'. Meanwhile, function `gsea.render` can compute trivial ORA statistics using the same input, except NW and produce output in a format maximally unified with `nea.render`.

4 RESULTS AND DISCUSSIONS

4.1 Paper I - NEArender: an R package for functional interpretation of ‘omics’ data via network enrichment analysis

Pathway annotation tools are indispensable to provide biological insights to the experimental datasets. As discussed in the pathway analysis session, gene enrichment analysis (GEA) tools traditionally considered an overlap between the experimental list and a pathway while ignoring the functional associations. Network enrichment analysis (NEA) implemented the strategy of integrating biological networks into GEA, thereby improving the sensitivity of GEA methods. It provides statistical scores to assess whether the members of AGS are enriched in a given FGS or not, considering the edges from the global network. Compared to GEA, the pathway score matrix from NEA is less sparse due to its higher sensitivity and therefore can be efficiently used in downstream analyses for e.g. predicting disease outcomes and phenotype modeling. To achieve this, the employed method should, beyond the higher statistical power, estimate the biological relevance by rendering the original data into the space of pathway scores in a faster, more convenient, and less biased procedure. Earlier NEA algorithms such as NEA-2012 and CrossTalk (Alexeyenko *et al.*, 2012; Mccormack *et al.*, 2013) estimated the probability of a null AGS-FGS model based on randomizations of the global network. By doing so, these versions suffered from problems associated with the statistical model and excessive computational time.

Considering the limitations of various pathway analysis tools as discussed above in Paper I, we introduce a new NEArender algorithm, which is based on a parametric approach using a chi-square formula for the null model estimation. This method produces unbiased estimates only in scale-free networks. However, networks that are artificially constructed from, for example, ChIP-seq based collections of transcription factor binding events, do not fit the power-law distribution and the network randomization approach needs to be employed while analyzing such networks. Here, the approach is to compare the p-value distributions between the previous implementations of network randomization (NRZ, an equivalent of NEA-2012) versus the current chi-square binomial (CSB, i.e. NEArender) approach. To evaluate this, network randomization runs with NRZ (N=3; 10; 30; 100; 300) was performed and compared the correlation of p-values with those from CSB. The estimates of enrichment from the NRZ procedure were apparently biased and could sufficiently converge to respective CSB values for smaller gene set sizes only at a very large, impractical number of randomizations, such as N(rand)=100…300. As mentioned above, these randomizations require excessive time. Calculating 100 randomizations on a Linux server using Perl software in NZB mode takes around 2300 min, while the parametric mode requires only less than 15 min using the same Perl software, with the CPU usage of 300-400MB RAM in both instances. Another possibility to minimize the randomization task also is to reduce the number of AGS and/or FGS. However, for predicting features in phenotype modeling where large AGS and FGS collections are required, these could negatively impact the analyses. Using the R package NEArender in its solely parametric mode required less than 5 min of processing time and used less than 200 MB RAM; thus, considering saving the computational time.

Additionally, the behavior of NEA in comparison to the default differential expression analysis of individual genes and ORA is investigated by simulating lack of experimental replicates (e.g. patient samples in clinical cohorts). The potential use of ORA and NEA is the ability to measure enrichment and summarize individual genes to the level of pathways and biological processes. However, this feature suggests a potential increase in the robustness of conclusions in experiments that lack replicates. The robustness of using single-gene expression values between ORA and NEA under replicated versus non-replicated design has been evaluated by the cell transcriptome data from FANTOM5 CAGE-sequencing dataset of the normal human tissue samples (Lizio *et al.*, 2015). The models were estimated via correlations of DE p-values from different sample sets or single samples obtained through DE analysis on raw gene profiles, ORA and NEA. The latter two were run on the same AGS and FGS collections. The results demonstrated that in the absence of replicates, the performance of NEA, i.e. preservation of significant findings was superior over the gene-wise analysis and ORA in terms of robustness. NEA has also enabled higher sensitivity compared to ORA. However, the advantages of sensitivity and robustness of NEA should not justify avoiding replicated designs. Furthermore, this approach is unlikely to identify an individual deregulated gene while being optimal in summarizing biological information at the pathway level.

In summary, the NEArender algorithm allowed avoiding the network randomization step while providing an unbiased parametric estimation of network connectivity. Output pathway scores in the simple matrix format can be integrated into bioinformatics pipelines as input for predicting disease outcomes and phenotype modeling. One such practical application is discussed in Paper III. On the other hand, the most popular pathway annotation tool in current research is DAVID (Huang *et al.*, 2007), which considers the gene overlap to known pathways as evidence. As discussed previously, though this approach has a lower sensitivity (true positive rate), many research studies rely on it. The reasons may be attributed to its simplicity, usability or affordable computational time. Although, NEArender thoroughly optimized in terms of computational time and high sensitivity, users still require prior knowledge of R programming. Thus, the major advantage of DAVID (Huang *et al.*, 2007) and similar resources is the availability of a user-friendly online interface. Therefore, a web interface to NEArender called EviNet (Jeggari *et al.*, 2018) is implemented, which is discussed in Paper II.

4.2 Paper II: EviNet: a web platform for network enrichment analysis of flexibly defined gene sets

The purpose of the web resource EviNet (Evidence-based Network Enrichment Test) is to provide a programming-free and user-friendly interface for exploring novel, experimentally defined gene sets. EviNet possesses unique functionalities for obtaining enrichment statistics integrated with interactive visualizations at the level of AGS-FGS and underlying subnetworks.

The web interface is created using jQuery functionality, while the HTML and JavaScript code is dynamically provided by Perl scripts via AJAX interface. The back-end employs PostgreSQL database engine and custom R scripts for accessing e.g. the core functionality of NEArender package and Venn diagrams. EviNet requires three input components, as described in the methods section 3.3. The users can submit their AGS and FGS either by directly pasting the gene list or uploading it as a file in the AGS menu tab. In addition, large FGS collections are available from the server by selection from the precompiled menu. A similar menu provides access to precompiled network versions for a few eukaryotic organisms, including human.

A unique feature of EviNet is the dynamic redefinition of AGS by changing the confidence and fold-change thresholds. This idea was developed due to the need in analyzing RNA-seq datasets to obtain DE gene lists with different, often very complex, combinations of experimental contrasts. Many biologists would be interested in overlapping DE lists from different experimental conditions by accounting for values associated with potential biological significance.

To address this, a Venn diagram functionality is implemented to quickly visualize and select gene lists that emerge from the relationships between experimental conditions. The server-side script reads the user-provided input file with DE values and generates lists of genes that satisfy each of the contrast-specific sets of filtering conditions. Each such list corresponds to one ellipse in the Venn diagram. Further, all possible overlaps in Venn diagrams (3, 7, and 15 in 2, 3 and 4-contrast analyses, respectively) are accompanied by corresponding gene lists, which pop up on the screen upon mouse clicks at the intersection areas. The list tables also contain DE values and can be investigated by sorting, gene ID search etc. Users can change filtering criteria, followed by the regeneration of the Venn diagram and the gene lists. Finally, the user chooses an arbitrary number of intersection gene lists for NEA. The lists will be treated as AGSs, while the user can proceed to the successive FGS and NW tabs, and then execute NEA. Figure I in the paper summarizes the workflow implementation of EviNet.

FGS collected from various resources (chapter 2.3.1) is readily available to the users. As these resources are updated on a timely basis, available back-end scripts are able to regenerate these files to provide updated versions. The upload option for FGS terms and collections creates a flexible environment for defining user-specific FGS. For example, FGS related to particular terms such as neuron or axon can be extracted from the full collection of GO-BP file, and the subset of terms can be used as a custom FGS collection. The global networks from various databases, mentioned in *Table 2* are available, too. The output results are presented in both visual and tabular formats using jQuery libraries. A project management system is also implemented for the registered users, where both input files and analysis results can be privately shared between project members. Each user can log in and use the project space with uploaded files stored in a separate directory with a unique project name.

A practical demonstration of EviNet is shown in this paper, by analyzing the transcriptome dynamics upon mouse embryonic stem cell (mESC) differentiation. Mouse embryonic stem cells are pluripotent stem cells, which can give rise to ectodermal, endodermal, and mesodermal cell lineages when cultured in appropriate differentiation conditions. Although mESCs were treated with the same morphogen cocktail, a large diversification occurred in the cell population. Both cell identities in this population and the signaling cascades that govern this process are unknown. Therefore, this aspect was studied by analyzing the transcriptome differences between mESCs in non-differentiating and differentiating conditions. Multiple differential expression criteria were combined for performing network analysis on the resulting gene lists. The analysis detected both known and revealed novel signaling genes and pathways that would be potentially important for maintaining stem cell pluripotency, differentiation, and diversification towards lineage specification.

4.3 PAPER III: Prediction of response to anti-cancer drugs becomes robust via network integration of molecular data

Regardless of the research field (cancer genomics, developmental biology, neuroscience), most biomedical studies aim is to understand the disease mechanisms and to discover novel biomarkers. Despite the widening range of high-throughput platforms and exponential growth of generated data volumes, biomarker discovery from large-scale data is a challenging task. For example, in cancer research, several endeavors to collect and analyze tumor specimens from cancer patients have resulted in the development of large public resources such as TCGA (Cancer Genome Atlas Research Network *et al.*, 2013) and CCLE (Barretina *et al.*, 2012). These consortia generated data from *in vitro* cancer cell lines and tumor samples using a range of omics assays (such as exome sequencing, copy number variation, DNA methylation, and RNA-seq). TCGA also provided information on clinical data (such as patient survival rate, administration of anti-cancer treatments etc.). These resources have created an unprecedented opportunity to study the underlying oncogenic molecular signatures for various cancer types. However, this is not an easy task. The mere feasibility of finding valid biomarkers from a large pool of potentially useful molecular signatures has been impeded by statistical and biological challenges such as excessive data dimensionality, imperfect analytical tools, the heterogeneity of cancer genomes, and the downstream diversity of regulation and expression patterns (Crystal *et al.,* 2014). Apparently, any 'omics' data analysis would face such challenges. While the NEArender algorithm is capable of rendering any molecular profile into the pathway space (applications presented in Paper I and Paper II), here the focus is on the biomarker discovery. The logic behind selecting the large datasets is to validate the usefulness of the newly developed algorithm and assess its robustness in other applications, including basic research data.

The additional challenge here is to identify patient sub-categories responsive to a treatment rather than one-dimensional drug sensitivity or survival analyses. A practical method should profile individuals across the cohort so that the profiles can be fit to clinical variables and covariates. Therefore, a crucial feature for biomarker discovery would be the ability to assign scores to individual samples rather than to derive feature-pathway associations from the whole data collection. A more challenging task is to identify the conservation of associated pathway-level features between the *in vitro* drug screens and the clinical application of the same drugs.

In Paper III, we present the new algorithm NEAmarker, a method for finding sensitive and robust biomarkers at the pathway level. Here large omics datasets are summarized into pathway scores, which are then used for evaluating the drug correlations. This strategy can also be implemented via any other enrichment analysis method (discussed in chapter 2.3.2), which is capable of reporting enrichment scores. However, the statistical power of different pathway analysis methods to detect the correlation with the related drug screens needs to be tested.

Previously, GSEA enrichment scores were used to analyze correlations between drug sensitivity and molecular features. The pathway enrichment scores represented correlates of drug sensitivity over the whole screened collections rather than characterized individual cell lines (Haibe-Kains *et al.,* 2013). SPIA, an approach that considers pathway topology into account is capable of calculating sample-specific pathway scores. However, their scores were based on gene expression values, which excluded the usage of other data types (e.g. point mutation datasets). Considering these specific features and accounting for their complexity, applicability to different experimental designs and the ability to analyze individual samples rather than the whole cohort, we modified the usage and output from the tools and evaluated a number of other enrichment analysis methods (ORA, GSEA, ssGSEA, ZGSEA, EGSEA and SPIA) that could be potentially useful in the proposed framework. The data analysis procedure included the method-specific steps for sample/patient characterization, enrichment analysis, and phenotype modeling.

To test NEAmarker along with other alternative enrichment methods, we took advantage of publicly available omics datasets (such as gene expression, copy number variation and point mutation) for cancer cell lines (*in vitro*) and primary tumor samples (clinical) from CCLE and TCGA respectively. The large datasets were processed in order to obtain sample/cell-line specific AGS. Running enrichment analyses on the AGS lists versus an FGS collection (mostly KEGG) allowed transforming the original, gene-wise omics data spaces into lower-dimensional pathway spaces (AGSxFGS score matrices). Further, the *in vitro* screens and clinical follow-up observations provided information on anti-cancer drug response in cell lines and patients, respectively. These variables, together with AGSxFGS matrix scores, were finally used to evaluate the drug versus feature correlations.

The benchmark results proved the superiority of NEA in comparison to the original gene profiles and benchmarked enrichment methods both within and across the *in vitro* and clinical domains in terms of- (i) level of correlation with drug sensitivity in cancer cell lines (ii) consistency of the discovered correlates in independent drug screens (iii) ability to explain the differential survival of patients and (iv) ability of the *in vitro* correlates to predict survival of patients who received the same drug. A new screen of four anti-cancer compounds validated the performance of the multivariate models of cancer cell sensitivity.

The poor performance of the original gene profiles and alternative enrichment methods could be explained by the excessive dimensionality of the former and reduced sensitivity of the latter. In addition, the ability to use smaller and hence more specific AGSs could have provided extra advantage of NEA over ORA and GSEA. On the other hand, NEA could also deteriorate on AGS of insufficient size when using sparser networks (around 10^4 - 10^5 edges) and networks with many missing nodes. These potential limitations were known from previous research and we tried to avoid them in the present work by using the denser network from data integration. Future implementation of NEA might adopt advantages of the alternative enrichment methods by employing full gene lists (as in GSEA) and intra-pathway topology (as in SPIA).

4.4 PAPER IV: Genome-wide characterization of floor plate transcription reveals cooperative activator function of Foxa2 and Rfx4 and a suppressive role for Ascl1 to spatially constrain floor plate induction in the neural tube

Graded Shh signaling by the axial mesoderm and the floor plate (FP) underlies the positional specification of five ventral progenitor domains and induction of the floor plate at the ventral midline (Jessell, 2000) (*Figure 3C).* The winged-helix transcription factor (TF) Foxa2 is induced in ventral midline cells in direct response to Shh signaling (Placzek *et al.*, 2005) and the forced expression of Foxa2 is sufficient to induce ectopic FP cells in the neural tube (Sasaki *et al.*, 1994). Foxa2 is required for the early specification of axial mesoderm (Weinstein *et al.*, 1994) and it has therefore been difficult to resolve if Foxa2 is required for FP differentiation. In addition to the FP, Foxa2 is also expressed at low levels in neurogenic p3-progenitors located immediately dorsal to the FP, and which are defined by their expression of the homeodomain protein Nkx2.2 (Jacob *et al.*, 2007). Once induced by Shh, Foxa2 promotes its expression through positive feedback and activates expression of Shh differentiating FP cells (Sasaki *et al.*, 1994, 1996; Metzakopian *et al.*, 2012). Shh and Foxa2 thereby form a positive feedback circuitry that must be interrupted over time to prevent a continuous spread of FP induction and to allow the establishment of the overlying Nkx2.2$^+$ p3-domain. This is at least partly achieved through temporal modulation in the responsiveness of cells to Shh over time (Lek *et al.*, 2010; Dias *et al.*, 2014; Sasai *et al.*, 2014) but how this is regulated at the molecular level remain poorly resolved.

To define the role of Foxa2 during the specification of FP cells and Nkx2.2$^+$ p3-progenitors, we examined the fate of differentiating Foxa2$^{-/-}$ embryonic stem cells (ESCs) exposed to the Shh-agonist SAG, and could show that Foxa2 is absolutely required for FP differentiation. We compared the transcriptome profiles of Foxa2$^{-/-}$ ESCs versus wild-type cells and defined 405 genes that are differentially expressed between these two conditions. The identified gene list includes already known FP markers such as *Shh, Arx, Corin, Slits*. Network enrichment analysis on the full list of identified FP genes shows enrichment for pathways such as axon-guidance, cell-adhesion, signal transduction. It is known that FP expresses secreted and transmembrane proteins that regulate the growth of commissural axons that cross the midline (Brose and Tessier-Lavigne, 2000). The subnetwork of the FP genes associated with the axon-guidance pathway reveals many known and novel genes. One such module shows the cluster of genes such as Slit ligands (*Slit1*, *Slit2*, *Slit3*) interacting with its receptors Roundabouts (*Robo1, 2, 3*)(Long *et al.*, 2004). Three other genes, namely *Negr1, App* and *Dlg4* contains the highest node degrees (hub genes) connecting many other FP genes in the subnetwork. Also, this subnetwork of axon-guidance could help to understand how members of Shh pathway genes are also involved in this process. For instance, the FP genes (*Cdon, Shh, Gpc3, Bmp7, Wnt4, Sfrp2*) have functional links associated with other known Shh pathway genes (Figure 2B in the Paper IV manuscript).

Foxa2 TF ChIP-seq data analysis revealed that Foxa2 directly binds in the proximity of at least 250 genes expressed by the FP. We also compared our dataset of Foxa2-bound regions in FP and Nkx2.2$^+$ NSCs with previously published Foxa2 ChIP-seq data (Metzakopian *et al.*, 2012) in midbrain progenitors that generate dopamine neurons. We found that only ~43% overlap of Foxa2 bound regions overlapped between the two datasets, indicating that the genome-wide binding profile of Foxa2 differs significantly between different subtypes of Foxa2-expressing cells in the ventral neural tube.

Histone modification data adds another dimensionality to define the chromatin states, identification of presumptive regulatory regions and estimate their changing activity across different conditions/cell types. By defining the chromatin states of four histone marks H3K4me1, H3K4me3, H3K27ac and H3K27me3 in wild-type and Foxa2$^{-/-}$ cells isolated at 3.5 and 5.5 DDC using ChromHMM (Ernst *et al.*, 2017), we looked at the acquisition of epigenetic state (promoters, active enhancers, transcribed or repressed genes) in/around the regions bound by Foxa2. The chromatin state analysis predicted that Foxa2-bound regions associated with FP expressed genes were scored as active enhancers in wild-type samples at 5.5DDC. The vast majority of these regions were transformed into weak enhancers or acquired a quiescent state in Foxa2$^{-/-}$ cells at 5.5DDC, supporting that Foxa2 acts as a master regulator by directly binding and regulating/activating hundreds of genes during the specification and differentiation of FP cells.

Motif enrichment analysis on a subset of Foxa2-bound regions associated with FP genes identified the binding sites for the transcription factors Rfx4 and Ascl1. Also, our analysis of Rfx4 mutant mice suggests that Rfx4 works together with Foxa2 to promote expression of Foxa2-bound FP genes during the differentiation of FP cells. Ascl1 is not expressed by FP cells but is expressed in dorsally abutting Nkx2.2^{+} NSCs, implying that Ascl1 contributes to constrain FP induction through direct suppression of FP genes. Analysis of Ascl1 ChIP-seq data indicates that *Ascl1* binds directly to a cohort of FP genes, including *Foxa2*, and analysis of Ascl1$^{-/-}$ cells indicate that an increased fraction of ESC-derived NSCs adopts a FP fate. Conversely, *Foxa2* binds to *Ascl1* and analysis of epigenetic marks indicate that the *Ascl1* loci acquire a more active state in Foxa2$^{-/-}$ NSCs. Thus, this data suggests that a cross-repressive interaction between *Foxa2* and *Ascl1* in balancing the positional specification of FP cells and Nkx2.2^{+} NSCs.

To summarize, using genome-wide data we were able to identify factors and a mechanism involved in restricting the induction of FP fate in the developing neural tube. The study provides an increased understanding of the gene regulatory network underlying FP differentiation and identifies a novel role for *Ascl1* as a suppressor of FP fate downstream of Shh.

4.5 PAPER V: A SHH/GLI-driven three-node timer motif controls temporal identity and fate of neural stem cells

Temporal patterning of neurons contributes to the generation of neural cell diversity at all axial levels of the CNS, but how time is encoded in these processes has not been resolved in any temporal lineage of the CNS (Kohwi *et al.*, 2013; Syed *et al.*, 2017). To define the composition and functional properties of time-measuring gene regulatory networks, we examined a relatively well-defined lineage in the ventral hindbrain that sequentially produces motor neurons (MNs), serotonergic neurons (5HTNs) and oligodendrocyte precursors (OPCs) (Pattyn *et al.*, 2003; Vallstedt *et al.*, 2005). The lineage is induced by Shh and defined by the expression of the TF Nkx2.2 (Pattyn *et al.*, 2003). Previous studies have shown that young Nkx2.2^{+} NSCs co-express early- and late-acting fate determinants (Pattyn *et al.*, 2000, 2003, 2004; Jacob *et* *al.*, 2007; Dias *et al.*, 2014), but the activity of the TF Phox2b predominates by specifying MN fate (Pattyn *et al.*, 2000; Dias *et al.*, 2014). Once Phox2b is downregulated or genetically ablated; MN production is terminated and 5HTNs are generated by default (Pattyn *et al.*, 2003) suggesting that Phox2b functions as temporal effector output. Activators of Phox2b have not been defined, but a self-sustained and temporally delayed activation of Tgfβ2 operates as an extrinsic signal that triggers MN-to-5HTN fate switch by repressing Phox2b (Dias *et al.*, 2014). Thus, Phox2b and Tgfβ2 constitute important regulatory components of a timer circuitry, but to understand how time is set by the network, it is necessary to define activators of Phox2b and resolve how the temporally gated activation of Tgfβ2 is mechanistically implemented.

We hypothesized that activators driving Phox2b are progressively lost over time since Phox2b becomes downregulated and cells undergo an MN-to-5HTN fate switch in the absence of Tgfβ2 signaling, but this occurs on a delayed temporal schedule (Dias *et al.*, 2014). To identify candidate activators, we defined the transcriptome of ESC-derived Nkx2.2^{+} NSCs isolated at different time points by RNA-sequencing and defined genes that were extensively downregulated during the phase of MN-production. By this approach, we identified Gli1-3 as candidate activators of Phox2b and could show that ongoing Shh signaling was required for sustained expression of Phox2b in the Nkx2.2^{+} lineage *in vitro*.

Biochemical analyses revealed that the progressive downregulation of *Gli1-3* genes was translated into a parallel temporal decline of GliA (Gli2A+Gli3A+Gli1) and GliR (Gli2R+Gli3A) activities over time, and that the amount of GliR generated at a given time was determined by the level of *Gli2/3* transcription and not by Ptch1-mediated feedback inhibition of the Shh pathway. To define the effect of constant GliA input on temporal output, we generated mice in which Gli1 was constitutively expressed in the Nkx2.2^{+} lineage (termed Gli1^{ON} mice). Unexpectedly, our analysis revealed that constant GliA input had only a minor effect on Phox2b expression, and cells underwent a MN-to-5HTN fate switch an almost normal temporal schedule. Importantly, we found that *Tgfβ2* was notably upregulated in Gli1^{ON} mice, suggesting a feedforward activation of *Tgfβ2* by GliA. This suggested a three-node circuitry forming an incoherent feedforward loop (IFFL), whereby GliA activates Phox2b but also the suppressive Tgfβ-node negatively regulating Phox2b. In strong support for this, we could show that the window of MN production was dramatically extended when Gli1 was overexpressed and when the Tgfβ pathway concurrently inactivated. Our genetic analyses in vivo further established that Tgfβ2 predominates over the Shh pathway by suppressing Phox2b even if cells express GliA at levels sufficient to sustain *Phox2b* transcription, and provided experimental support to the notion that delayed MN-to-5HTN switch in *Tgfbr1* mutant mice (Dias *et al.*, 2014) reflects the downregulation of Phox2b due to depletion of GliA.

The fact that GliA promotes expression of both Phox2b and Tgfβ2 raised the key question of how activation of the Tgfβ2-node is circumvented at early stages when GliA expression in cells peak? Detailed analyses of wild type and Gli1^{ON} mice revealed that young progenitors which express *Gli2/3* are refractory for *Tgfβ2* induction by GliA, while *Tgfβ2* responded in a GliA dose-dependent manner in old cells that ceased to express *Gli2/3* and thereby lost capacity to produce GliR. We could also show by the epistasis experiment that GliR acted dominant-negative over GliA by suppressing *Tgfβ2* expression induced in response to forced expression of GliA and Nkx2.2. Collectively, these data suggested a GliR inhibitor-titration regulation of Tgfβ2, whereby a high GliR-sensitivity prohibits GliA-mediated gene activation until GliR has been titrated, thereby establishing a delayed activation of the Tgfβ-node. In other genetic experiments, we could also show that Gli1 was required for the late induction of *Tgfβ2* and for prompt suppression of Phox2b and termination of MN-production. This establishes that the GliA threshold required for *Tgfβ2* induction is higher than the threshold necessary to sustain Phox2b expression, and show that the feedforward activation of Gli1 by Gli2/3 mediate a function to boost GliA input late in the temporal differentiation process. Collectively, our data outline a three-node IFFL circuitry in which GliA promotes *Phox2b* expression and MN fate, but also accounts for a delayed activation of a suppressive Tgfβ-node that triggers an MN-to-5HTN fate switch by repressing Phox2b. Since the amount of GliR generated is coupled to Gli transcription, an altered decay-rate of Gli genes will change time-output by regulatory circuitry and therefore, conceptually explains how time is encoded by the circuitry.

Our biochemical analyses reveal notable fluctuation of Gli1 and Gli2 in cells at a given time point examined, revealing the noisy expression of Gli proteins in temporal lineage progression. Data further suggest that the production of GliA and GliR are coupled and thereby inflexible, and the theoretical part of our study suggest that decay of Gli proteins alone cannot counterbalance noise. However, our computational analyses reveal that the diffusible and self-activating properties of Tgfβ2, in combination with hysteresis, produces prompt suppression of *Phox2b* and a coordinated switch at the population level. Integration of the Tgfβ-node into the Shh/Gli-driven circuitry thereby acts to counterbalances noise and generates a more precise timer mechanism as compared to a timer-based only on Gli decay. The community features mediated by Tgfβ2 are not attainable with temporal networks based exclusively on intrinsic transcriptional regulators. Our study, therefore, provides a functional basis for the intrinsically programmed activation of extrinsic switch signals in temporal patterning processes of the vertebrate CNS.

5 CONCLUSIONS AND FUTURE PERSPECTIVES

The first part of the thesis introduced the network-based pathway analysis (NEA) as a key tool to comprehend a genome-wide pattern of interactions between multiple genes and their products that would emerge from high throughput datasets. Available biological network resources, as well as the most popular approaches in the pathway analysis field, have been discussed. The most important difference between the network-based and network-free analyses is that, the latter can only employ and study genes which already belong to the functional term. On the contrary, NEA can identify and use those that are both in and around, including potentially novel, never annotated genes. Previously developed network-based methods suffered from problems with the statistical model correctness and/or excess computational time. The scope of this thesis includes the development of the NEArender tool, which integrates biological components for the network analysis while employing– instead of the tedious network randomization runs – a quick analytical, parametric calculation to assess the statistical significance of functional associations between AGS and FGS. The specific features of NEArender in its current implementation are that, (i) it considers AGS of only limited length rather than full ranked gene lists, (ii) accounts for only direct links between AGS and FGS genes (iii) ignores intra-FGS and intra-AGS edges and (iv) disregards edge attributes and their directionality. Integrating this information is likely to give more insights into the role of certain interactions and leads to better performance. Therefore, future work should aim at implementing respective features into NEArender.

Even though NEArender has been optimized for speed and is suitable for integration into larger bioinformatics pipelines, the operation requires knowledge of R scripting and command-line interface. Therefore, a user-friendly web interface called EviNet has been developed, which provides a fast and flexible solution for performing network analysis online. In the current world of pathway analysis, commercial databases such as Ingenuity Pathway Analysis (IPA) are attractive for the biological community – mainly due to their proprietary knowledge databases, design interface and data protection. In comparison, EviNet is transparent in both the public database usage and the algorithmic details and - being still in its infancy, it incorporates a number of valuable and often unique features. EviNet supports using multiple alternative biological networks and functional gene sets collected from various public databases. In addition, the program incorporates a data management system and has the flexibility in integrating highly complex experimental designs using interactive Venn-diagram features. However, there is still room for future developments. For instance, EviNet integrates the data from multiple database resources and the output presented might contain heterogeneous annotation contents. Integrating an algorithm that could group the FGS annotations into similar clusters like in DAVID (Huang *et al.*, 2007) functional annotation clustering could reduce the burden of redundant terms and make the biological interpretation more focused. Also, integrating ancillary tools for gene ID format conversion, differential expression analysis and respective data visualization in various forms such as heatmaps, 2D plots etc. could enhance the usefulness of EviNet and is to be developed in the future implementations.

With the expanding range of high-throughput platforms and enormous volumes of data generation, the need for computational tools is essential. However, the tools should also generate results (biomarkers discovered from large-scale data) that are possible to validate in a clinical application. Previously developed network and pathway analysis tools were rarely applied to this task. The newly developed NEAmarker present a proof of concept as well as a practical application of using pathways enrichment scores from NEArender in drug response prediction. The method integrates data from multiple omics platforms in order to model drug sensitivity with enrichment variables. In this paper, the comparison has also been made with conventional analysis of original gene profiles and pathway enrichment methods. The analysis results showed that the poor performance of the individual gene analysis is due to excess data dimensionality, whereas, other alternative pathway enrichment methods might lack sensitivity. The ability to summarize information scattered over the network, and thereby use smaller and more specific AGS has provided an advantage to NEA, but the analysis could be deteriorated by using the same AGS and FGS input with sparser networks. Employing full gene lists as in GSEA, or intra-pathway topology as in SPIA, while considering the above-mentioned limitations might help NEA to evolved towards better performance.

The second part of the thesis is focused on identifying the regulatory factors that are governed by mechanisms that operate in space and over time, which account for the formation of neural cell diversity in developing brain. Genome-wide sequencing techniques such as ChIP-seq and RNA-seq aided in depicting the underlying epigenome and transcriptional regulation events. In Paper IV, we identified the novel role for Ascl1 as a suppressor of floor plate fate. In Paper V, Gli proteins were identified as the candidate activators of Phox2b expression and MN-fate and also accounted for the late onset of Tgfβ2, which executes the MN-to-5HTN fate switch by suppressing Phox2b. Further computation modeling on the input components enabled to set the timer motif at which the temporal fate switch is prompted. Resolving the molecular mechanisms and identifying the regulatory factors underlying the temporal control of neurogenesis and progenitor potency is of central importance for overall understanding of neural development. This can provide future tools for cell-reprogramming technologies that aim to develop specific subtypes of neural cells from human Embryonic Stem Cells (hESCs) or induced Pluripotent Stem Cells (iPSCs) (*Mertens et al. 2016*).

With the advent of genome sequencing technologies, we have learned how to read our own genome and with CRISPR-CAS9, we created a tool for writing our own instructions. A major task remaining for the future is to identify novel biomarkers to determine the underlying molecular signatures of diseases detrimental to human health. The development of novel computational approaches and refinement of the existing dataset analysis tools are essential to progress towards this goal. In this thesis, an attempt has been made to develop a robust bioinformatics toolkit for network enrichment analysis. The software has been benchmarked and tested extensively to validate its usefulness with large datasets. Besides, integrating various bioinformatics approaches into custom analyses of large-scale datasets contributed to the field of developmental biology by helping to identify the regulatory factors.

6 ACKNOWLEDGEMENTS

I am glad that I have been surrounded by many wonderful people who let me grow higher in life and have been there to lift me up when I am falling down.

First of all, thank you **Johan Ericson** for accepting me into your lab. It's been a great journey for me. Thank you for involving me in various projects and teaching me developmental biology concepts. You have given me all the freedom to portray my own ideas with independence. Irrespective, you have also participated in what I explore and have guided me with your expertise. Thank you for providing me with all high-throughput datasets, access to world-class conferences and courses which made me attain knowledgeable skillsets to nurture my future career.

Andrey Alexeyenko, you have played a significant role in my research career right from my master's thesis. Thank you for guiding me all these years to enhance not just my bioinformatics and programming skills, but also focusing on many aspects throughout this Ph.D. Though we didn't share the workspace, the way you organize and schedule helped me to get papers from you at the right time. You have been a great supervisor and thank you for giving me the right push with your insights when I needed the most.

Members of JE lab

Zhanna: You are a charming person whom I have met in my life. Thank you for introducing me to the field, showing me some of your interesting in vitro cultures and making me feel fascinated about stem-cell biology. Thank you for all your support, care and love Zhanna! **Jose:** Your simplifications and illustrations to address complex questions helped me to clearly establish my data analysis framework and also structure my programs accordingly. Thank you for all the brilliant discussions and sharing knowledge. **Katarina**: You are one such calm person whose smile never fades with a welcoming personality. I could find an easy solution for many things I wondered about. Thank you for always listening to me and recommending me with all your valuable suggestions. **Masha,** I still remember our first journey to Bodrum and many more throughout this Ph.D. I enjoyed sharing the thoughts in science, as well as Ph.D struggles with you. I wish you too accomplish your Ph.D soon and have a beautiful life in the UK. Good luck with your future endeavors! **Lizzy,** energy and charisma surround you all the time. I will never forget your extra efforts to provide me with appropriate medical help. I am extremely grateful for your care and have always enjoyed chatting with you. **Svitlana**, thank you for exposing me to the collection of ESC culture protocol and also for all those pleasant conversations. **Chris,** thank you for sharing your brilliant insights and enjoyed collaborating with you on a few projects as well as exchanging our domain knowledge. **Olga,** Thank you for all those friendly chats.

Maria Bergsland and **Johan Reimegård,** thank you for participating in my queries related to ChIP-seq data analysis. **Carsten Daub,** thank you for being my mentor and all your advice. **Matti Nikkola:** Thank you for your guidance throughout my Ph.D and engaging me in teaching activities. **Linda Lindhall**, **Lina Petterson**, **Christine Lindberg** and **Margaret Ulander**, thank you for taking care of all the administration work and letting me focus on my research. **Anethe Mansen** and **Angelo de Milto**, for offering me the internship opportunity.

CMB, Biomedicum and other friends

Anoop, Elaiya, Matti and **Alberto**. Thank you for your company and friendly chats during lunch. **Anoop** and **Elaiya**, your research experiences speak a lot and thank you for always enlightening me about career opportunities and insights. **Mauricio**, my bioinformatics solo friend at CMB. Thank you for all the good times in several courses, knowledge sharing and your friendly support during this Ph.D life. Good luck, Mau! **Katrin Mangold,** Sweet and enthusiastic person. Thank you for always keeping your heart open for the kind and warm hugs. **Soniya Savant,** still remembers how our first conversation had started. Thank you **Soniya** and **Sandeep,** for your caring words and friendly conversations. Good luck with your new home and for your aspiring careers. **Simona Hanke, Anna Middleton and Bettina Reichenbach**, lunches are lively with your talks. Enjoyed all your conversations over some of our lunches and thank you for your company gals. **Divya Nagaraj,** thank you for sharing a positive attitude and making me assure that things go well. I hope to hear some more interesting stories in the future**. Shahul**, thank you for involving me in activities and your parties at Lappis. Good luck with your new career. **Him Jaiswal,** Thank you for your kindness and advices. **Anushree,** Glad I met you, thank you for all those beautiful stories and discussions. **Varsha**: Thank you for being a good friend, never I will forget the wonderful time with you in Japan. **Srivathsa**, always felt happy to share my knowledge with you and helping out with bioinformatics, as well as other friendly chats. I wish something dramatic happens with your phenotype, as well as in your life and everything turns exciting. Good luck, Sri! **David Gro,** you are a fantastic course organizer and it was an enjoyable experience in Idrefjäll. Thank you! Also, thanks to **Tatiana, Vera Shirokova**, **Iv Mayor**, **Kim**, **Christina Kantzer**, **Goncalo, Benjamin ka-cheuk, Eduardo Guimaraes, David, Daniel Holl, Helena Silva, Divya Thiagarajan, Simon Kebede and Iurii Petrov**. **530 group: Shailesh** (skiing partner), **Constance, Chenhong and Dörte:** Fun hanging around with you and discussing science outside of KI.

Senthil, thank you for your support during many instances, especially all those battles with accommodation searches and move-outs. I am glad that we grew up together as bioinformaticians in our careers and expanded our bits of knowledge by all our amazing discussions. Thank you for optimizing some of my pipelines and teaching me the tricks and tacks of being a good programmer. Thank you for everything you have done on this long journey of my life in Sweden. Long live our friendship. **Lakshmi**, my comfort-zone talking friend. Always remember our endless talks at my home and thank you for listening to me. **Revathy,** thank you for filling me with your authentic Tamil food always. Ekambar, Lakshmi, Revathi, Vinoth and Senthil, I will never forget all our get-togethers, food parties and your fantastic company. Thank you all. **Varun and pratyusha:** Thank you for your insights on science. **Dilshad:** Tumko aur tumare friendship ko jail na aasan nahi hai. But I enjoyed every bit of hangouts and Stockholm misses you Dilu.

Proofreaders: Andrey, Johan, Anoop, Zhanna, Katta, Jose and Chandhu: for your constructive remarks and inputs from various expertise domains. Thank you for your time and valuable comments.

Zhanna, Andrey, Lizzy, Svitlana, Katrin and Johan Sten: Thank you for traveling to India for my wedding. Your presence has created a special page in my wedding diaries and I treasure those memorable moments for my life.

Biggest assets of my life-family

My parents: **Indira** and **Devendar**: You dreamt of my goals and I'm what I'm today because of you. You have accomplished my works when I'm lazy and taught me what I should aspire in my life. You gave me all the freedom in this world to be like a bird and heartfully respected my choices. Thank you for believing me, driving me all my lazy, crazy times and being there for me at all times. Thank you for all your constant support and love. **Vanima and Anand appa**: Blessed to find you, like my other parents. Thank you for your steady support and encouragement to pursue what I wished for. **Anusha:** You are the one who dragged me to Sweden and there for me and encouraged me to start my Ph.D. **Sharath**, you are one who invokes positive vibes whenever I miss home and being what I am. Thank you both for all your unlimited care. Excited about the little one. **Tharun:** Thank you for all your support and believing in me. **Vamshi** (my techie guru): With the blink of an eye, you have always been supportive in many ways. Thank you! Finally, but not last **Chandhu:** all this wouldn't be possible without your utmost love and support. Your words boost me every day up, especially the last few months were crazy. Thank you for giving me all my time and waiting with loads of patience for me to accomplish my objectives. Thank you for all your efforts in reading my thesis, even though it's all Greek to you. Thanks for motivating me to stay focused and hearing me out when I am depressed. Hereafter I am all yours to spend our exciting lives together forever ❤.

7 REFERENCES

Alexeyenko, A. *et al.* (2010) 'Dynamic zebrafish interactome reveals transcriptional mechanisms of dioxin toxicity', *PLoS ONE*, 5(5).

Alexeyenko, A. *et al.* (2012) 'Network enrichment analysis: extension of gene-set enrichment analysis to gene networks.', *BMC bioinformatics*, 13(1), p. 226.

Alexeyenko, A. and Sonnhammer, E. L. L. (2009) 'Global networks of functional coupling in eukaryotes from comprehensive data integration', *Genome research*, 19(6), pp. 1107–1116.

Anders, S. and Huber, W. (2013) 'Differential expression of RNA-Seq data at the gene level – the DESeq package'.

Anders, S., Pyl, P. T. and Huber, W. (2014) 'HTSeq – A Python framework to work with high-throughput sequencing data', *bioRxiv*, p. 2824.

Andrews S. (2010) (no date) *FastQC: a quality control tool for high throughput sequence data*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed: 11 May 2019).

Ashburner, M. *et al.* (2000) 'Gene Ontology: tool for the unification of biology', *Nature Genetics*. Nature Publishing Group, 25(1), pp. 25–29.

Azevedo, H. and Moreira-Filho, C. A. (2015) 'Topological robustness analysis of protein interaction networks reveals key targets for overcoming chemotherapy resistance in glioma', *Scientific Reports*. Nature Publishing Group, 5(1), p. 16830.

Barabasi, A.-L. and Albert, R. (1999) 'Emergence of scaling in random networks', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 286(5439), pp. 509–12.

Barabási, A.-L. and Oltvai, Z. N. (2004) 'Network biology: understanding the cell's functional organization', *Nature Reviews Genetics*, 5(2), pp. 101–113.

Barbie, D. A. *et al.* (2009) 'Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1', *Nature*. Nature Publishing Group, 462(7269), pp. 108–112.

Barretina, J. *et al.* (2012) 'The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity', *Nature*. Nature Publishing Group, 483(7391), pp. 603–607.

Barrett, T. *et al.* (2012) 'NCBI GEO: archive for functional genomics data sets—update', *Nucleic Acids Research*. Narnia, 41(D1), pp. D991–D995.

Barski, A. *et al.* (2007) 'High-Resolution Profiling of Histone Methylations in the Human Genome', *Cell*. Cell Press, 129(4), pp. 823–837.

Bennet, A. M. *et al.* (2011) 'Genetic Association of Sequence Variants Near AGER/NOTCH4 and Dementia', *Journal of Alzheimer's Disease*, 24(3), pp. 475–484.

Breuer, K. *et al.* (2013) 'InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation', *Nucleic Acids Research*, 41(D1), pp. D1228–D1233.

Briscoe, J. *et al.* (2000) 'A homeodomain protein code specifies progenitor cell identity and neuronal fate in the ventral neural tube.', *Cell*. Elsevier, 101(4), pp. 435–45.

Brown, K. R. and Jurisica, I. (2005) 'Online Predicted Human Interaction Database', *Bioinformatics*. Narnia, 21(9), pp. 2076–2082.

Brown, K. R. and Jurisica, I. (2007) 'Unequal evolutionary conservation of human protein interactions in interologous networks', *Genome Biology*. BioMed Central, 8(5), p. R95.

Cancer Genome Atlas Research Network (2011) 'Integrated genomic analyses of ovarian carcinoma', *Nature*, 474(7353), pp. 609–615.

Cancer Genome Atlas Research Network, J. N. *et al.* (2013) 'The Cancer Genome Atlas Pan-Cancer analysis project.', *Nature genetics*. NIH Public Access, 45(10), pp. 1113–20.

Caspi, R. *et al.* (2018) 'The MetaCyc database of metabolic pathways and enzymes', *Nucleic Acids Research*. Narnia, 46(D1), pp. D633–D639.

Cerami, E. G. *et al.* (2011) 'Pathway Commons, a web resource for biological pathway data', *Nucleic Acids*

Research, 39(Database issue), pp. D685-690.

Chen, Y., Shi, W. and Smyth, G. K. (2014) ‘voom: precision weights unlock linear model analysis tools for RNA-seq read counts ∗’, (May 2013), pp. 1–31.

Chesnutt, C. *et al.* (2004) ‘Coordinate regulation of neural tube patterning and proliferation by TGFβ and WNT activity’, *Developmental Biology*, 274(2), pp. 334–347.

Clarivate Analytics (no date) *MetaCore*. Available at: https://clarivate.com/products/metacore/.

Claudio D. Stern (2005) ‘Neural induction: old problem, new findings, yet more questions.’, *Development (Cambridge, England)*. The Company of Biologists Ltd, 132(9), pp. 2007–2021.

Cook Deegan and Mullan, R. (2014) ‘Origins of the Human Genome Project’. University of Washington.

Corbit, K. C. *et al.* (2005) ‘Vertebrate Smoothened functions at the primary cilium’, *Nature*, 437(7061), pp. 1018–1021.

Coulon, A. *et al.* (2013) ‘Eukaryotic transcriptional dynamics: from single molecules to cell populations’, *Nature Reviews Genetics*. Nature Publishing Group, 14(8), pp. 572–584.

Crystal, A. S. *et al.* (2014) ‘Patient-derived models of acquired resistance can identify effective drug combinations for cancer’, *Science*, 346(6216), pp. 1480–1486.

Dessaud, E., McMahon, A. P. and Briscoe, J. (2008) ‘Pattern formation in the vertebrate neural tube: a sonic hedgehog morphogen-regulated transcriptional network.’, *Development (Cambridge, England)*, 135(15), pp. 2489–503.

Dias, J. M. *et al.* (2014) ‘Tgfβ Signaling Regulates Temporal Neurogenesis and Potency of Neural Stem Cells in the CNS’, *Neuron*, pp. 927–939.

Dobin, A. *et al.* (2013) ‘STAR: ultrafast universal RNA-seq aligner’, *Bioinformatics*, 29(1), pp. 15–21.

Draghici, S. *et al.* (2007) ‘A systems biology approach for pathway level analysis’, *Genome Research*. Cold Spring Harbor Laboratory Press, 17(10), pp. 1537–1545.

Eden, E. *et al.* (2009) ‘GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists.’, *BMC bioinformatics*. BioMed Central, 10, p. 48.

ENCODE Project Consortium *et al.* (2012) ‘An integrated encyclopedia of DNA elements in the human genome.’, *Nature*, 489(7414), pp. 57–74.

Enseqlopedia (no date). Available at: http://enseqlopedia.com/enseqlopedia/ (Accessed: 15 July 2019).

Ericson, J. *et al.* (1997) ‘Pax6 controls progenitor cell identity and neuronal fate in response to graded Shh signaling.’, *Cell*, 90(1), pp. 169–80.

Ernst, J. and Kellis, M. (2017) ‘Chromatin-state discovery and genome annotation with ChromHMM’, *Nature Protocols*. Nature Publishing Group, 12(12), pp. 2478–2492.

Erten, S. *et al.* (2011) ‘DA DA: Degree-Aware Algorithms for Network-Based Disease Gene Prioritization’, *BioData Mining*, 4(1), p. 19.

Feng, J. *et al.* (2012) ‘Identifying ChIP-seq enrichment using MACS.’, *Nature protocols*. Nature Publishing Group, 7(9), pp. 1728–40.

Frings, O., Alexeyenko, A. and Sonnhammer, E. L. L. (2013) ‘MGclus: network clustering employing shared neighbors’, *Molecular bioSystems*, 9(7), pp. 1670–1675.

Fuccillo, M., Joyner, A. L. and Fishell, G. (2006) ‘Morphogen to mitogen: the multiple roles of hedgehog signalling in vertebrate neural development’, *Nature Reviews Neuroscience*. Nature Publishing Group, 7(10), pp. 772–783.

Gaidatzis, D. *et al.* (2015) ‘Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation’, *Nature Biotechnology*. Nature Publishing Group, 33(7), pp. 722–729.

Genohub (no date) *Next Generation Sequencing Instrument Guide*. Available at: https://genohub.com/ngs-instrument-guide/ (Accessed: 11 May 2019).

Ghiassian, S. D., Menche, J. and Barabási, A.-L. (2015) ‘A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome’, *PLOS Computational Biology*. Edited by A. Rzhetsky. Public Library of Science, 11(4), p.

e1004120.

Gillis, J. and Pavlidis, P. (2012) '"Guilt by Association" Is the Exception Rather Than the Rule in Gene Networks', *PLoS Computational Biology*, 8(3).

Giurgiu, M. *et al.* (2019) 'CORUM: the comprehensive resource of mammalian protein complexes—2019', *Nucleic Acids Research*. Narnia, 47(D1), pp. D559–D563.

Glaab, E. *et al.* (2012) 'EnrichNet: network-based gene set enrichment analysis', *Bioinformatics*, 28(18), pp. i451–i457.

Griffith, M. *et al.* (2015) 'Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud Introduction to RNA Sequencing'.

Guney, E. and Oliva, B. (2012) 'Exploiting Protein-Protein Interaction Networks for Genome-Wide Disease-Gene Prioritization', *PLoS ONE*, 7(9).

Haibe-Kains, B. *et al.* (2013) 'Inconsistency in large pharmacogenomic studies', *Nature*. Nature Publishing Group, 504(7480), pp. 389–393.

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

Haycraft, C. J. *et al.* (2005) 'Gli2 and Gli3 Localize to Cilia and Require the Intraflagellar Transport Protein Polaris for Processing and Function', *PLoS Genetics*, 1(4), p. e53.

Heinz, S. *et al.* (2010) 'Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.', *Molecular cell*, 38(4), pp. 576–89.

Hong, M.-G. *et al.* (2010) 'Genome-wide pathway analysis implicates intracellular transmembrane protein transport in Alzheimer disease', *Journal of Human Genetics*. Nature Publishing Group, 55(10), pp. 707–709.

Huang, D. W. *et al.* (2007) 'The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists.', *Genome biology*. BioMed Central, 8(9), p. R183.

Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2009) 'Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists', *Nucleic Acids Research*. Oxford University Press, 37(1), pp. 1–13.

Jacob, J. *et al.* (2007) 'Transcriptional repression coordinates the temporal switch from motor to serotonergic neurogenesis', *Nature neuroscience*, 10(11), pp. 1433–9.

Jeggari, A. *et al.* (2018) 'EviNet: a web platform for network enrichment analysis with flexible definition of gene sets.', *Nucleic acids research*. England, 46(W1), pp. W163–W170.

Jeggari, A. and Alexeyenko, A. (2017) 'NEArender: an R package for functional interpretation of "omics" data via network enrichment analysis', *BMC Bioinformatics*. BioMed Central, 18(S5), p. 118.

Jessell, T. M. (2000) 'Neuronal specification in the spinal cord: inductive signals and transcriptional codes', *Nature Reviews Genetics*, 1(1), pp. 20–29.

Jewison, T. *et al.* (2014) 'SMPDB 2.0: Big Improvements to the Small Molecule Pathway Database', *Nucleic Acids Research*, 42(D1), pp. D478–D484.

Johnson, D. S. *et al.* (2007) 'Genome-wide mapping of in vivo protein-DNA interactions.', *Science (New York, N.Y.)*, 316(5830), pp. 1497–502.

Joshi-Tope, G. *et al.* (2004) 'Reactome: a knowledgebase of biological pathways', *Nucleic Acids Research*. Narnia, 33(Database issue), pp. D428–D432.

Kandasamy, K. *et al.* (2010) 'NetPath: a public resource of curated signal transduction pathways', *Genome Biology*, 11(1), p. R3.

Kanehisa, M. and Goto, S. (2000) 'KEGG: Kyoto Encyclopedia of Genes and Genomes', *Nucleic Acids Research*, 28(1), pp. 27–30.

Karolchik, D., Hinrichs, A. S. and Kent, W. J. (2009) 'The UCSC Genome Browser', *Current protocols in bioinformatics*, Chapter 1, p. Unit1.4-Unit1.4.

Khatri, P., Sirota, M. and Butte, A. J. (2012) 'Ten years of pathway analysis: Current approaches and outstanding challenges', *PLoS Computational Biology*.

Kim, D., Langmead, B. and Salzberg, S. L. (2015) 'HISAT: a fast spliced aligner with low memory requirements’, *Nature methods*. NIH Public Access, 12(4), p. 357.

Kitano, H. (2001) *Foundations of Systems Biology edited by*. 1st ed.

Kohwi, M. and Doe, C. Q. (2013) ‘Temporal fate specification and neural progenitor competence during development’, *Nature Reviews Neuroscience*. Nature Research, 14(12), pp. 823–838.

Landt, S. G. *et al.* (2012) ‘ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.’, *Genome research*, 22(9), pp. 1813–31.

Langfelder, P. and Horvath, S. (2008) ‘WGCNA: an R package for weighted correlation network analysis’, *BMC Bioinformatics*. BioMed Central, 9(1), p. 559.

Langmead, B. *et al.* (2009) ‘Ultrafast and memory-efficient alignment of short DNA sequences to the human genome’, *Genome Biology*. BioMed Central, 10(3), p. R25.

Lek, M. *et al.* (2010) ‘A homeodomain feedback circuit underlies step-function interpretation of a Shh morphogen gradient during ventral neural patterning.’, *Development (Cambridge, England)*, 137(23), pp. 4051–60.

Li, H. and Durbin, R. (2009) ‘Fast and accurate short read alignment with Burrows-Wheeler transform’, *Bioinformatics (Oxford, England)*. 2009/05/18. Oxford University Press, 25(14), pp. 1754–1760.

Li, Q. *et al.* (2011) ‘Measuring reproducibility of high-throughput experiments’, *The Annals of Applied Statistics*, 5(3), pp. 1752–1779.

Liberzon, A. *et al.* (2015) ‘The Molecular Signatures Database (MSigDB) hallmark gene set collection’, *Cell Systems*, 1(6), pp. 417–425.

Liem, K. F., Jessell, T. M. and Briscoe, J. (2000) ‘Regulation of the neural patterning activity of sonic hedgehog by secreted BMP inhibitors expressed by notochord and somites’, *Development*, 4866(127), pp. 4855–4866.

Liem, K. F., Tremml, G. and Jessell, T. M. (1997) ‘A Role for the Roof Plate and Its Resident TGFβ-Related Proteins in Neuronal Patterning in the Dorsal Spinal Cord’, *Cell*. Cell Press, 91(1), pp. 127–138.

Lim, Y. and Golden, J. A. (2007) ‘Patterning the developing diencephalon’, *Brain Research Reviews*, 53(1), pp. 17–26.

Lizio, M. *et al.* (2015) ‘Gateways to the FANTOM5 promoter level mammalian expression atlas’, *Genome Biology*. BioMed Central, 16(1), p. 22.

Long, H. *et al.* (2004) ‘Conserved roles for Slit and Robo proteins in midline commissural axon guidance.’, *Neuron*, 42(2), pp. 213–23.

Louvi, A. and Artavanis-Tsakonas, S. (2006) ‘Notch signalling in vertebrate neural development’, *Nature Reviews Neuroscience*, 7(2), pp. 93–102.

Machanick, P. and Bailey, T. L. (2011) ‘MEME-ChIP: motif analysis of large DNA datasets’, *Bioinformatics (Oxford, England)*. 2011/04/12. Oxford University Press, 27(12), pp. 1696–1697.

Maere, S., Heymans, K. and Kuiper, M. (2005) ‘BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks’, *Bioinformatics*, 21(16), pp. 3448–3449.

Marinov, G. K. *et al.* (2014) ‘Large-scale quality analysis of published ChIP-seq data.’, *G3 (Bethesda, Md.)*, 4(2), pp. 209–23.

Marioni, J. C. *et al.* (2008) ‘RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays’, *Genome Research*, 18(9), pp. 1509–1517.

Maslov, S. *et al.* (2002) ‘Specificity and Stability in Topology of Protein Networks’, *Science*, 296(5569), pp. 910–913.

Maxam, A. M. and Gilbert, W. (1977) ‘A new method for sequencing DNA.’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 74(2), pp. 560–4.

Mccormack, T. *et al.* (2013) ‘Statistical Assessment of Crosstalk Enrichment between Gene Groups in Biological Networks’, *PLoS ONE*, 8(1).

Metzakopian, E. *et al.* (2012) ‘Genome-wide characterization of Foxa2 targets reveals upregulation of floor plate genes and repression of ventrolateral genes in midbrain dopaminergic progenitors.’, *Development (Cambridge, England)*, 139(14), pp. 2625–34.

Mi, H. *et al.* (2019) ‘Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0)’, *Nature Protocols*. Nature Publishing Group, 14(3), pp. 703–721.

Mortazavi, A. *et al.* (2008) ‘Mapping and quantifying mammalian transcriptomes by RNA-Seq’, *Nature Methods*. Nature Publishing Group, 5(7), pp. 621–628.

Nagalakshmi, U. *et al.* (2008) ‘The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing’, *Science*, 320(5881), pp. 1344–1349.

NIH (no date) *Biological Pathways Fact Sheet | NHGRI, 2015*. Available at: https://www.genome.gov/about-genomics/fact-sheets/Biological-Pathways-Fact-Sheet (Accessed: 15 May 2019).

Nishikawa, S.-I., Jakt, L. M. and Era, T. (2007) ‘Embryonic stem-cell culture as a tool for developmental cell biology’, *Nature reviews. Molecular cell biology*, 8(6), pp. 502–507.

Nordström, U. *et al.* (2006) ‘An Early Role for Wnt Signaling in Specifying Neural Patterns of Cdx and Hox Gene Expression and Motor Neuron Subtype Identity’, *PLoS Biology*. Edited by R. Nusse. Public Library of Science, 4(8), p. e252.

Oosterveen, T. *et al.* (2013) ‘SoxB1-driven transcriptional network underlies neural-specific interpretation of morphogen signals.’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 110(18), pp. 7330–5.

Osseward, P. J. and Pfaff, S. L. (2019) ‘Cell type and circuit modules in the spinal cord’, *Current Opinion in Neurobiology*. Elsevier Current Trends, 56, pp. 175–184.

Oughtred, R. *et al.* (2019) ‘The BioGRID interaction database: 2019 update’, *Nucleic Acids Research*. Narnia, 47(D1), pp. D529–D541.

Patel, R. K. and Jain, M. (2012) ‘NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data’, *PLoS ONE*. Edited by Z. Liu. Public Library of Science, 7(2), p. e30619.

Pattyn, A. *et al.* (2000) ‘Control of hindbrain motor neuron differentiation by the homeobox gene Phox2b.’, *Development (Cambridge, England)*, 127(7), pp. 1349–58.

Pattyn, A. *et al.* (2003) ‘Coordinated temporal and spatial control of motor neuron and serotonergic neuron generation from a common pool of CNS progenitors.’, *Genes & development*, 17(6), pp. 729–37.

Pattyn, A. *et al.* (2004) ‘Ascl1/Mash1 is required for the development of central serotonergic neurons.’, *Nature neuroscience*, 7(6), pp. 589–95.

Peterson, K. A. *et al.* (2012) ‘Neural-specific Sox2 input and differential Gli-binding affinity provide context and positional information in Shh-directed neural patterning.’, *Genes & development*, 26(24), pp. 2802–16.

Pico, A. R. *et al.* (2008) ‘WikiPathways: Pathway Editing for the People’, *PLoS Biology*. Public Library of Science, 6(7), p. e184.

Pierani, A. *et al.* (1999) ‘A Sonic Hedgehog–Independent, Retinoid-Activated Pathway of Neurogenesis in the Ventral Spinal Cord’, *Cell*. Cell Press, 97(7), pp. 903–915.

Placzek, M. and Briscoe, J. (2005) *The floor plate: Multiple cells, multiple signals*, *Nature Reviews Neuroscience*. Nature Publishing Group.

du Plessis, L., Skunca, N. and Dessimoz, C. (2011) ‘The what, where, how and why of gene ontology--a primer for bioinformaticians.’, *Briefings in bioinformatics*. 2011/02/17. Oxford University Press, 12(6), pp. 723–35.

R.Milo *et al.* (2002) ‘Network motifs: simple building blocks of complex networks.’, *Science*, 298(5594), pp. 824–827.

Ramírez, F. *et al.* (2014) ‘deepTools: a flexible platform for exploring deep-sequencing data’, *Nucleic acids research*. 2014/05/05. Oxford University Press, 42(Web Server issue), pp. W187–W191.

Reactome (2019) *Version 69 Released - Reactome Pathway Database*. Available at: https://reactome.org/about/news/137-version-69-released (Accessed: 28 May 2019).

Reynolds, C. A. *et al.* (2010) ‘Analysis of lipid pathway genes indicates association of sequence variation near SREBF1/TOM1L2/ATPAF2 with dementia risk.’, *Human molecular genetics*. Oxford University Press, 19(10), pp. 2068–78.

Roberts, R. J., Carneiro, M. O. and Schatz, M. C. (2013) ‘The advantages of SMRT sequencing’, *Genome*

Biology. BioMed Central, 14(6), p. 405.

Roelink, H. *et al.* (1995) ‘Floor plate and motor neuron induction by different concentrations of the amino-terminal cleavage product of sonic hedgehog autoproteolysis’, *Cell*. Cell Press, 81(3), pp. 445–455.

Rohatgi, R., Milenkovic, L. and Scott, M. P. (2007) ‘Patched1 Regulates Hedgehog Signaling at the Primary Cilium’, *Science*, 317(5836), pp. 372–376.

Rowitch, D. H. and Kriegstein, A. R. (2010) ‘Developmental genetics of vertebrate glial–cell specification’, *Nature*. Nature Publishing Group, 468(7321), pp. 214–222.

Ruepp, A. *et al.* (2007) ‘CORUM: the comprehensive resource of mammalian protein complexes’, *Nucleic Acids Research*, 36(Database), pp. D646–D650.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) ‘DNA sequencing with chain-terminating inhibitors.’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 74(12), pp. 5463–7.

Sasai, N., Kutejova, E. and Briscoe, J. (2014) ‘Integration of Signals along Orthogonal Axes of the Vertebrate Neural Tube Controls Progenitor Competence and Increases Cell Diversity’, *PLoS Biology*. Edited by K. G. Storey. Public Library of Science, 12(7), p. e1001907.

Sasaki, H. and Hogan, B. L. (1994) ‘HNF-3 beta as a regulator of floor plate development.’, *Cell*, 76(1), pp. 103–15.

Sasaki, H. and Hogan, B. L. (1996) ‘Enhancer analysis of the mouse HNF-3 beta gene: regulatory elements for node/notochord and floor plate are independent and consist of multiple sub-elements.’, *Genes to cells : devoted to molecular & cellular mechanisms*, 1(1), pp. 59–72.

Schaefer, C. F. *et al.* (2009) ‘PID: the Pathway Interaction Database’, *Nucleic acids research*. Oxford University Press, 37(Database issue), pp. D674-679.

Schmitt, T., Ogris, C. and Sonnhammer, E. L. L. (2014) ‘FunCoup 3.0: database of genome-wide functional coupling networks’, *Nucleic Acids Research*, 42(Database issue), pp. D380-388.

Schroeder, A. *et al.* (2006) ‘The RIN: an RNA integrity number for assigning integrity values to RNA measurements’, *BMC Molecular Biology*, 7(3).

Shannon, P. *et al.* (2003) ‘Cytoscape: a software environment for integrated models of biomolecular interaction networks’, *Genome research*, 13(11), pp. 2498–2504.

Shendure, J. *et al.* (2017) ‘DNA sequencing at 40: past, present and future’, *Nature*. Nature Publishing Group, 550(7676), pp. 345–353.

Shendure, J. and Ji, H. (2008) ‘Next-generation DNA sequencing’, *Nat Biotechnol*. Nature Publishing Group, 26(10), pp. 1135–1145.

Sims, D. *et al.* (2014) ‘Sequencing depth and coverage: key considerations in genomic analyses’, *Nature Reviews Genetics*. Nature Publishing Group, 15(2), pp. 121–132.

Smyth, G. K. *et al.* (2014) ‘limma : Linear Models for Microarray Data User’s Guide (Now Including RNA-Seq Data Analysis)’, (June).

Solomon, M. J., Larsen, P. L. and Varshavsky, A. (1988) ‘Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene.’, *Cell*, 53(6), pp. 937–47.

Soneson, C. and Delorenzi, M. (2013) ‘A comparison of methods for differential expression analysis of RNA-seq data’, *BMC Bioinformatics*. BioMed Central, 14(1), p. 91.

Stark, C. *et al.* (2006) ‘BioGRID: a general repository for interaction datasets.’, *Nucleic acids research*. Oxford University Press, 34(Database issue), pp. D535-9.

Subramanian, A. *et al.* (2005) ‘Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles’, *Proceedings of the National Academy of Sciences*, 102(43), pp. 15545–15550.

Sultan, M. *et al.* (2008) ‘A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.’, *Science (New York, N.Y.)*. American Association for the Advancement of Science, 321(5891), pp. 956–60.

Syed, M. H., Mark, B. and Doe, C. Q. (2017) ‘Playing Well with Others: Extrinsic Cues Regulate Neural

Progenitor Temporal Identity to Generate Neuronal Diversity.', *Trends in genetics : TIG*, 33(12), pp. 933–942.

Szklarczyk, D. *et al.* (2015) 'STRING v10 : protein – protein interaction networks , integrated over the tree of life', *Nucleic acids research*, 43(October 2014), pp. 447–452.

Tarca, A. L. *et al.* (2009) 'A novel signaling pathway impact analysis', *Bioinformatics*. Oxford University Press, 25(1), pp. 75–82.

Tarca, A. L. *et al.* (2012) 'Down-weighting overlapping genes improves gene set analysis', *BMC Bioinformatics*. BioMed Central, 13(1), p. 136.

Tavassoly, I., Goldfarb, J. and Iyengar, R. (2018) 'Systems biology primer: the basic methods and approaches', *Essays in Biochemistry*, p. 20180003.

Tornow, S. and Mewes, H. W. (2003) 'Functional modules by relating protein interaction networks and gene expression', *Nucleic Acids Research*. Narnia, 31(21), pp. 6283–6289.

Trapnell, C. *et al.* (2010) 'Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.', *Nature biotechnology*. Nature Publishing Group, 28(5), pp. 511–5.

Trapnell, C., Pachter, L. and Salzberg, S. L. (2009) 'TopHat : discovering splice junctions with RNA-Seq', *Bioinformatics*, 25(9), pp. 1105–1111.

Ulloa, F. and Martí, E. (2009) 'Wnt won the war: Antagonistic role of Wnt over Shh controls dorso-ventral patterning of the vertebrate neural tube', *Developmental Dynamics*, 239(1), p. NA-NA.

Vallstedt, A., Klos, J. M. and Ericson, J. (2005) 'Multiple dorsoventral origins of oligodendrocyte generation in the spinal cord and hindbrain', *Neuron*.

Velculescu, V. E. *et al.* (1995) 'Serial Analysis of Gene Expression', *Science*, 270(5235), pp. 484–487.

Wang, E. T. *et al.* (2008) 'Alternative isoform regulation in human tissue transcriptomes', *Nature*. Nature Publishing Group, 456(7221), pp. 470–476.

Warde-Farley, D. *et al.* (2010) 'The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function', *Nucleic Acids Research*, 38(suppl_2), pp. W214–W220.

Weinstein, D. C. *et al.* (1994) 'The winged-helix transcription factor HNF-3 beta is required for notochord development in the mouse embryo.', *Cell*, 78(4), pp. 575–88.

Wijgerde, M. *et al.* (2002) 'A direct requirement for Hedgehog signaling for normal specification of all ventral progenitor domains in the presumptive mammalian spinal cord', *Genes & Development*. Cold Spring Harbor Laboratory Press, 16(22), pp. 2849–2864.

Wilhelm, B. T. *et al.* (2008) 'Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution', *Nature*, 453(7199), pp. 1239–1243.

Winter, C. *et al.* (2012) 'Google Goes Cancer: Improving Outcome Prediction for Cancer Patients by Network-Based Ranking of Marker Genes', *PLoS Computational Biology*. Edited by D. K. Slonim, 8(5), p. e1002511.

Wurst, W. and Bally-Cuif, L. (2001) 'Neural plate patterning: Upstream and downstream of the isthmic organizer', *Nature Reviews Neuroscience*. Nature Publishing Group, 2(2), pp. 99–108.

Yu, H. *et al.* (2007) 'The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics.', *PLoS computational biology*. Public Library of Science, 3(4), p. e59.

Zhang, F. and Drabier, R. (2012) 'IPAD: the Integrated Pathway Analysis Database for Systematic Enrichment Analysis', *BMC Bioinformatics*. BioMed Central, 13(S15), p. S7.

Zhang, Y. *et al.* (2008) 'Model-based Analysis of ChIP-Seq (MACS)', *Genome Biology*. BioMed Central, 9(9), p. R137.

Zhu, L. J. *et al.* (2010) 'ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data', *BMC Bioinformatics*. BioMed Central, 11(1), p. 237.