Università degli Studi di Napoli "Federico II"

PhD in Computational Biology and Bioinformatics

Network Inference on RNA-Seq data from Mammalian Retina

Candidate:
Fabiola Curion

Supervisor:
Prof. E. M. Surace

Co-Supervisor:
Prof. D. di Bernardo

2015/2016

To my father,
my Professor.

To Rosa, Filomena, Andrea, Carmela,
and their family.

May they all trust their hearts, and their story.
May they all follow their dreams,
until the very end.

"Not all those who wander are lost"
J. R. R. Tolkien

Abstract

The mammalian retina is an intricate network of cells communicating and cooperating to convey light stimuli to the visual cortex of the brain. Moreover, it is the most accessible part of the Central Nervous System and hence a valuable model to study the CNS.

A hierarchical scheme of transcription factors (TF) that determine each cells' identity is regularly expressed following a precise timeline, since the early stages of development of the embryo. The interplay of those TF controls univocal flows of transcription and genetic programs which direct cells' identities, maintain their specific expression patterns and guarantee the survival of each cell type.

Despite the large interest of the scientific community on retina, and the large variety of databases collecting gene expression profiles from multiple species, very few Next Generation Sequencing experiments on this tissue were collected in public available data. We generated a co-expression network using porcine whole retina RNA-seq data produced in our laboratory to characterize the retina specific Gene Regulatory Networks, which are disrupted in retinal diseases.

Our inferred network shows good performance and reliability of the predicted connections. We characterised retina-specific processes by comparing our dataset with a RNA-seq study on 10 porcine tissues. Furthermore, we

characterized the genome-wide functional effects of a synthetic transcription factor composed of a DNA-binding domain targeted to a 20 bp of *Rhodopsin* (RHO) cis-regulatory sequence, which induced RHO specific transcriptional silencing upon adeno-associated viral (AAV) vector delivery.

Finally, we assessed the rod-specific repression of RHO after FACS-sorting photoreceptors interfered with our construct, and confirmed this results on single cells by qPCR.

Contents

Chapter 1

The Visual System

The visual system is the part of the central nervous system which gives organisms the ability to process visual detail, as well as enabling the formation of several non-image photo response functions [1]. The visual system in animals allows individuals to assimilate information from their surroundings.

In the case of mammals (including humans), the visual system consists of:

- The eye, especially the retina
- The optic nerve
- The optic chiasma
- The optic tract
- The lateral geniculate body
- The optic radiation
- The visual cortex
- The visual association cortex.

The act of seeing starts when the cornea and then the lens of the eye focuses an image of its surroundings onto a light-sensitive membrane in the back of the eye, called the retina. This photo-sensitive tissue serve as a transducer for the conversion of patterns of light into neuronal signals, that travels through the optic nerve to reach the visual cortex, where the perception is built [2].

1.1 Retina

Light entering the eye is refracted by the cornea and the lens, which act together to project an inverted image onto the retina.

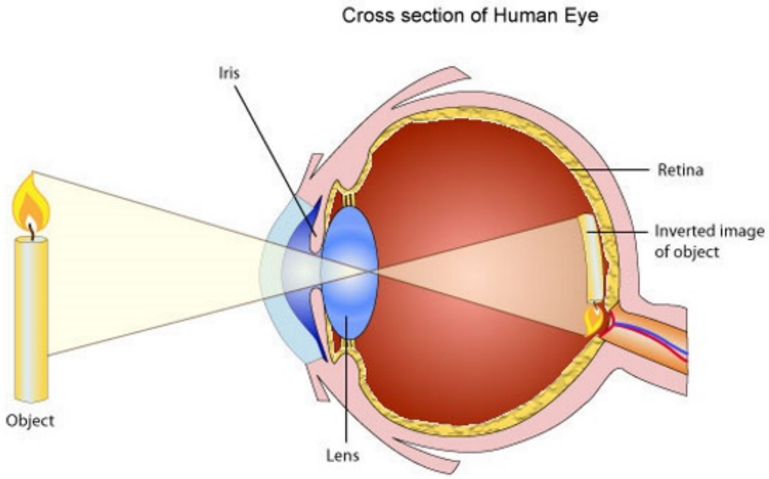


Figure 1.1: The images are inverted when projected on the retina.

Light striking the retina initiates a cascade of chemical and electrical events that ultimately trigger nerve impulses. These are sent to various visual centres of the brain through the fibres of the optic nerve.

The vertebrate retina has ten distinct layers. From closest to farthest from the vitreous body:

1. **Inner limiting membrane** – basement membrane elaborated by Müller cells

2. **Nerve fibre layer** – axons of the ganglion cell nuclei

3. **Ganglion cell layer** – contains nuclei of ganglion cells, the axons of which become the optic nerve fibres for messages and some displaced amacrine cells

4. **Inner plexiform layer** – contains the synapse between the bipolar cell axons and the dendrites of the ganglion and amacrine cells.

5. **Inner nuclear layer** – contains the nuclei and surrounding cell bodies (perikarya) of the amacrine cells, bipolar cells and horizontal cells

6. **Outer plexiform layer** – projections of rods and cones ending in the rod spherule and cone pedicle, respectively. These make synapses with dendrites of bipolar cells.[1] In the macular region, this is known as the Fiber layer of Henle.

7. **Outer nuclear layer** – cell bodies of rods and cones

8. **External limiting membrane** – layer that separates the inner segment portions of the photoreceptors from their cell nucleus

9. **Layer of rods and cones** – layer of rod cells and cone cells

10. **Retinal pigment epithelium** - single layer of cuboidal cells. This is closest to the choroid.

In adult humans, the entire retina is approximately 72% of a sphere about 22 mm in diameter, and is no more than 0.5 mm thick. The optic nerve carries the ganglion cell axons to the brain and the blood vessels that open into the retina. The ganglion cells lie innermost in the retina while the photoreceptive cells lie outermost. Because of this counter-intuitive arrangement, light must first pass through and around the ganglion cells and through the thickness of the retina, before reaching the rods and cones.

1.1.1 Cell types of the mammalian retina

The retina is composed of a repertoire of more that 60 cells types, organized as a miniaturized processor to perform parallel computation of the visual scene. [3, 4] All cells in the retina are derived from the multipotent retinal progenitor cells (RPCs) and can be subdivided in six main classes of cells [5, 6].

The photoreceptors transduce light into an electrical signal. The electrical stimulus is transferred through the synaptic terminals of rods and cones onto bipolar and horizontal cells. Horizontal cells, of which there are between one and three types in mammalian retinae, provide lateral interactions in the outer plexiform layer. One type of rod bipolar cell and at least nine types of cone bipolar cells transfer the light signals into the inner plexiform layer, onto the dendrites of amacrine and ganglion cells. Amacrine cells are inhibitory interneurons, and there are as many as 50 morphological types, while ganglion cell dendrites collect the signals of bipolar and amacrine cells and their axons eventually transmit these signals to the visual centres of the brain [7].

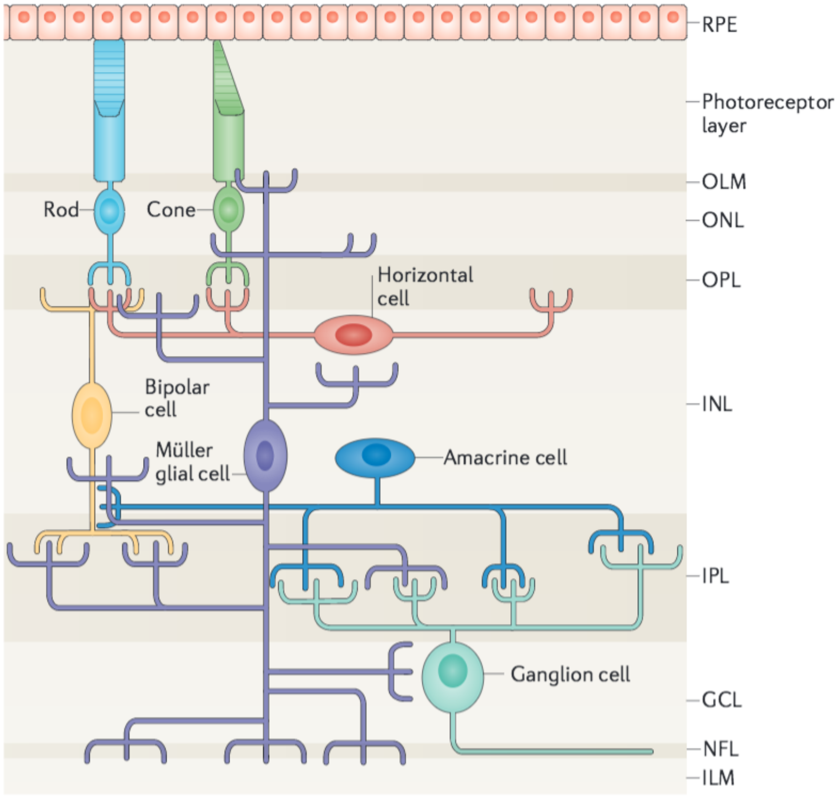


Figure 1.2: Mammalian Retina is structured in stacked interconnected layers, with different cell types. From [8].

Photoreceptors

Figure 1.3: Photoreceptors. Modified from [3].

A photoreceptor is the neuronal cell type responsible for light detection in retina. Photoreceptors convert light (visible electromagnetic radiation) into chemical signals that can stimulate biological processes. There are two major types of photoreceptor cells in mammalian eyes: Rods and Cones. Those cells cooperate to provide information used by the visual system to form a representation of the visual world.[9] A third photoreceptor class of cells was discovered during the 1990s : the photosensitive ganglion cells. These cells do not contribute to sight directly, but are thought to support circadian rhythms and pupillary reflex.[10]

Both rods and cones are characterized by remarkable degree of intracellular compartmentalization, as they are constituted by highly specialized portions: an outer segment, an inner segment, and a synaptic terminal. The outer segment is located toward the outer surface of the retina and is primarily involved in phototransduction. This segment consists of a stack of membranous discs formed by an infolding of the plasma membrane. Those discs contain light-absorbing photo-pigments. In rods, these discs are free floating because they pinch off from the plasma membrane, while in the cones the discs remain attached to the plasma membrane.

The outer segments are constantly being renewed.[11]. The prompt and efficient clearance by receptor-mediated phagocytosis of photoreceptor outer segment fragments (POS) shed daily by photoreceptors in a diurnal rhythm is coordinated by Retina Pigmented Epitelium (RPE) cells and is essential both for long-term viability and functionality of photoreceptors.[12]

The inner segment of each photoreceptor contains the cell's nucleus and organelles such as mitochondria and Golgi bodies, it is connected to the outer segment by a stalk or cilium that contains microtubules. The synaptic terminal makes synaptic contact with the other cells.

Despite their commmon role in photontransduction, rod and cone photoreceptor show some remarkable difference in their structure and function. Rods have a long, cylindrical, outer segment with many discs, while cones have a short, tapering outer segment with relatively few discs.

Rods are extremely sensitive, as a single photon reaching their photosensitive pigment can trigger its signaling cascade. Thus, rods are mainly responsible for dim-light vision.
Cones require significantly brighter light (i.e., a larger numbers of photons) in order to produce a signal. In humans, there are three different types of cone cell, distinguished by their pattern of response to different wavelengths of light. Color experience is calculated from these three distinct signals. The human retina consists of about 120 million rod cells and 6 million cone cells. The number and ratio of rods to cones varies among species, dependent on whether an animal is primarily diurnal or nocturnal.

The distribution of the rods and cones in the retina is not uniform: around the periphery of the retina, rods are far more numerous, whereas in the centre, at the fovea, cones are. The number of photoreceptors connected to a single ganglion cell is also far greater in the peripheral retina.

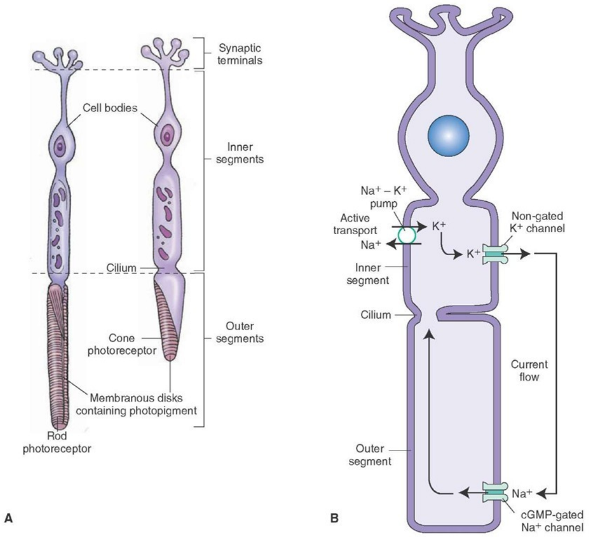


Figure 1.4: a) Rod and Cone Photoreceptors, b) Ion channels regulating the current flow within both cells. From [2]

Bipolar cells

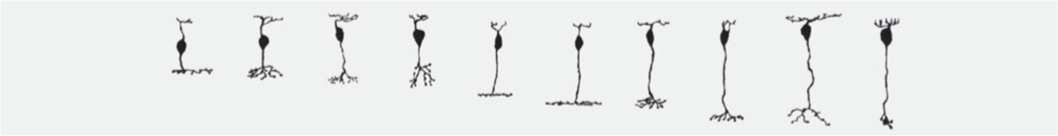


Figure 1.5: Bipolar cells. Modified from [3].

All of the signals coming from the photoreceptor are transmitted to bipolar cells, whose body is located in the inner nuclear layer of the retina together with horizontal cells, müller cells and amacrine cells. Bipolar cells send a single dendrite in the direction of the photoreceptor, and, contacting every other neuron type in the retina, they provide the link between the primary sensory neurons and the highly specialized circuit that communicates to the brain, which is represented by about 20 different types of retinal ganglion cells (RGCs). Types of bipolar cells differentially collect and shape photoreceptor signals for further processing in the inner retina, thereby carrying out the first elementary operations of the visual system [13]. There are more than ten types of bipolar cells in the mammalian retina, divided into ON and OFF, which either synapses with one photoreceptor or split their dendrite into branches that synapse with more cells. The subsection of the Inner Plexiform Layer in which they project their synapses allows to further divide the cells in clusters whitjh distinct specificity for photoreceptors and mechanisms of signal transcduction [13, 14]. Cone bipolar cells are born throughout the period of bipolar cell genesis, and rod bipolar cells are born only in the later part of this period [7].

Horizontal cells

Figure 1.6: Horizontal cells. Modified from [3].

The horizontal cells are inhibitory interneurons positioned within the INL [15]. The expression of the neural bHLH transcription factors neurogenic differentiation 1 (NeuroD1) and mouse atonal homolog 3 (Neurod4) in embryonic RPCs coincides with the birth of amacrine cells and horizontal cells. Committed horizontal cells migrate to the appropriate retinal stratum in their early developmental stages, but they reach their final morphology and place only in late retinal development, with paths still unclear [16]. All rods and cones receive feedback from horizontal cells, but these cells are a numerically small proportion of the retinal interneurons, generally less than 5% of cells of the INL, among the rarest cells in the retina [7, 17]. The rod feedback system is isolated from the one of the cone, because the ranges of brightness covered by rods and cones are enormously different. Horizontal cells adjust the system's response to the overall level of illumination by subtracting the illumination that hit a large region of the retina from the actual signal that reaches the innter retina (basically, they *adjust the contrast* of adjacent region of dark and light of the image perceived). This process reduce redundancy in the signal transmitted, since the mean luminance across a large region of retina is shared by many cones and contains little information [3].

Müller cells

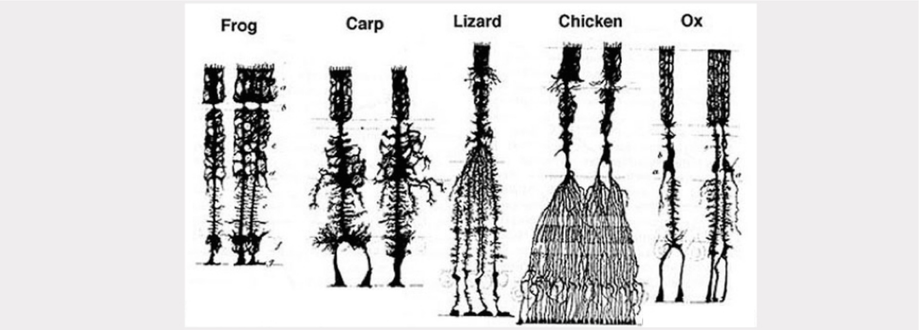


Figure 1.7: Müller cells. From *Cajal, 1892*.

Müller glia are the only cell type to span all retinal layers and have processes that contact neighbouring neurons and form part of the outer and inner limiting membranes. They function as barriers and conduits regulating the homeostasis of a wide range of molecules between different retinal cells and compartments.

They also support the structure of the layers and feed neurons by releasing trophic factors, recycling neurotransmitters and controlling ionic balance in the extracellular space. In addition, Müller glia control physiological processes of cones: they phagocytose cone outer segments, contribute to outer segment assembly and participate in a cone-specific visual cycle that helps to recycle the retinal chromophore for photodetection. Furthermore, those cells contribute directly to photoreceptor stimulation by acting as optic fibers and guiding light to those cells [18].

Müller glia respond to retinal injury and disease by changing their morphology, abundance, biochemistry, an injury response referred to as reactive *gliosis*. The triggers for proliferative gliosis are not well understood. Both proliferative and non-proliferative responses to injury are accompanied by changes in gene expression and are often associated with Müller glial cell

hypertrophy [19, 20]. Those dramatic changes need to be restricted in time and space, since prolongued gliosis can interfere with retinal pathways and thus induce degeneration [8].
Although their proliferative behaviour, müller cells do not function as retinal progenitors *in vivo* [21, 22], but in human cell culture, Müller glia have been observed to generate both neurons and glia and, more importantly, photoreceptors and retinal ganglion cells that have some reparative potential when transplanted into a damaged rodent retina [23, 24].

Amacrine cells

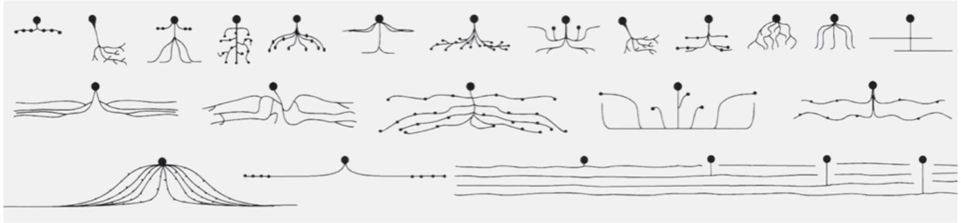


Figure 1.8: Amacrine cells. Modified from [3].

Described by Golgi and Cajal in the first decades of the nineteenth century, Amacrine cells are the most diverse group of neurons in the retina: there are at least 33 different subtypes of amacrine cells based just on their dendrite morphology, stratification, and the type of neurotransmitter that they release [25, 26]. Amacrine cells are interneurons interacting at the second synaptic level of the photoreceptor-bipolar-ganglion cell chain. Those cells provide a feedback synapse onto the bipolar cells and also form numerous synapses throughout the IPL with retinal ganlion cells and other amacrine cells. Those cells contribute to most of the synapses in the inner plexiform layer, thus mediating to vertical communication within the retinal layers. Most of the amacrine cells release inhibitory neurotransmitters, such as GABA and glycine, but an exception is represented by starburst amacrine cells, one of the most famous, which release acetylcholine [27]. Although the reason of acetylcholine release are still unknown, it is thought that through this transmitter, together with dopamine release, amacrine cells can carry out paracrine functions [25].

Through their intricate network of connections and release of neurotransmitters, they contribute to the detection of directional motion, modulate light adaption and circadian rhythm [25], and control high sensitivity in scotopic vision through connections with rod bipolar cells, suggesting a role in converging rod signals from huge areas of retina and in amplifying them at very low light intensities [28].

Retinal Ganglion Cells

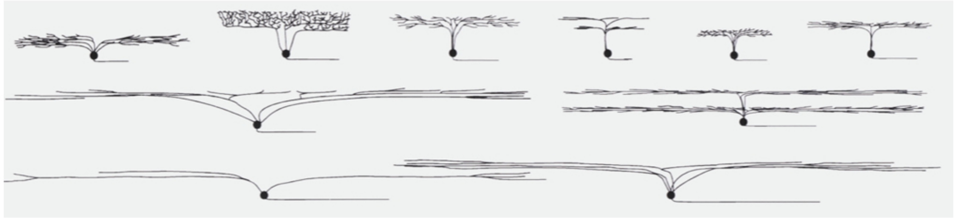


Figure 1.9: Retinal Ganglion Cells. Modified from [3].

Ganglion cells are the final output neurons of the vertebrate retina. They collect all of the preprocessed inputs of the vertical pathway of the retinal cell types and convey all the electrical stimuluses to the brain. There are at least 18 different morphological types of ganglion cell in the human retina, which vary significantly in terms of their size, connections, and responses to visual stimulation but they all share the defining property of having a long axon that extends into the optic nerve. A small percentage of retinal ganglion cells contribute little or nothing to vision, but are themselves photosensitive

[29, 30]. There are about 0.7 to 1.5 million retinal ganglion cells in the human retina, which communicate with the rough 96 million photoreceptors with a ratio of 1:100. However, these numbers vary greatly among individuals and as a function of retinal location: in the fovea, a single ganglion cell can communicate with as few as five photoreceptors, while in the extreme periphery it will receive information from many thousands of photoreceptors [31].

Thus, the density of their receptive fields imposes a fundamental limit on the spatial resolution of human vision. This density varies across the retina, declining rapidly with distance from the fovea.

In most mammals, the axons of retinal ganglion cells are not myelinated where they pass through the retina, since light could be interfered in its path to the photoreceptors by myelinated axons, a phenomenon observed indeed in some eye diseases [32].

1.1.2 How light gets converted into electrical impulses

Most of what is known about the molecular events of phototransduction has been gleaned from experiments in rods.

The photosensitive pigment of rods cells is rhodopsin, a heptahelical transmembrane receptor expressed on rods' outer segment membrane. On the other hand, cone cells contain cone opsins, M-(Medium wave), L-(Long wave) and S-(Short wave) opsins, after their different exciting wavelength.

Rods' and cones' opsins belong to the G-protein Coupled Receptors Protein family (GPCRs), and both are covalently bound to a chromophore, the 11-cis-retinal, a derivate of Vitamin A [33]. Each outer segment disc contains many thousands of visual pigment molecules (the opsin conjugated with the chromophore).

Upon absorption of a photon, the 11-cis-retinal undergoes photoisomerization to all-trans-retinylidene, inducing a change in Rhodopsin from its inactive to its active conformation. The active form, known as meta-Rhodopsin II, then recruits and binds intracellular G proteins, the trasducins, which in turn activate the phosphodiesterase (PDE).

The phosphodiesterase hydrolyzes cGMP, reducing its concentration in the outer segment and leading to the closure of sodium channels in the outer segment membrane. Rods differ from other sensory cells in that light leads to hyperpolarization rather than depolarization. The hyperpolarization of the outer segment leads to the closure of the calcium voltage depending channels, thus decreases or terminates the dark glutamate release at the synaptic terminal. Hence the rod photoresponse is essentially a transient suppression of the circulating current. The signal is further processed by other neurons in the retina before being transmitted in electrical impulses to the visual cortex. Soon after, opsin and the chromophore recombine to regenerate fresh rhodopsin [34, 35].

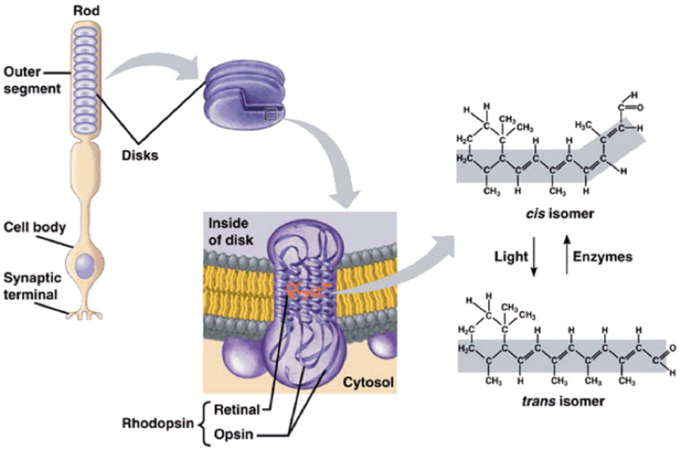


Figure 1.10: Rod outer segment and Rhodopsin.

Rhodopsin

The total amount of rhodopsin per eye is roughly 650 pmoles (3.96×10^{14} Rhodopsin molecules), about 8×10^4 Rhodopsin molecules per disk [36, 35]. Hence, is not surprising that this is the most highly expressed gene in the retina. The balance of a wild-type Rhodopsin protein is essential for proper functioning and survival of photoreceptors and, intuitively, the entire retina [37] [38]. The morphology of the protein, which is of elliptic, cylindrical shape, is due to arrangement of its seven transmembrane helices, which vary in length from 20 to 33 residues [39]. The N-terminal region is located intradiscally (extracellularly) and it's the "plug" of the chromophore [40], while the C-terminal region is cytoplasmic [35, 36]. The aminoacidic sequence of the protein consists of 348 residues, [41, 42] where the most prevalent amino acids are Phe (8.9%), Val (8.9), Ala (8.3), and Leu (8.0), suggesting a major hydrophobic character for this protein.

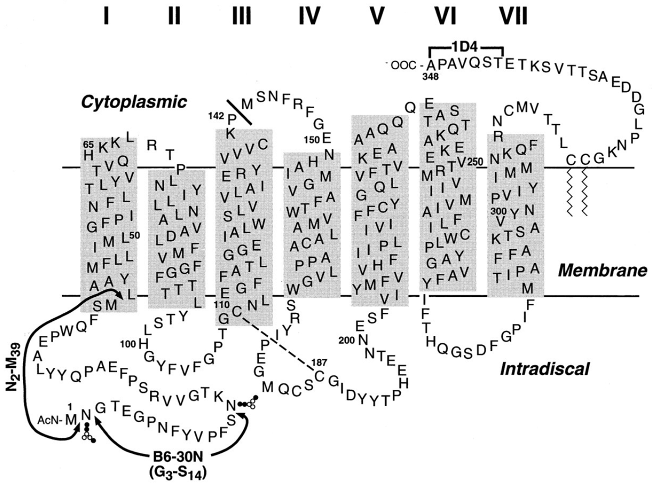


Figure 1.11: Rhodopsin aminoacidic sequence, from[41]

It has been observed by atomic force microscopy and transmission electron microscopy under various conditions that Rhodopsin is able to form rows of dimers containing densely organized higher-order structure [35]. The dimerization of Rhodopsin can explain the autosomal dominant character of rhodopsin mutants, which will be discussed in the next chapter.

Upon activation, one of the fastest photochemical reactions known in biology, rhodopsin undergoes multiple reactions and intermediates that culminate in the formation of the G protein–activating state, termed metarhodopsin II, or Meta II. Being the first component of the visual cascade, Rhodopsin is a substantial hub of the phototransduction [43].

Chapter 2

Vision Impairment and therapy

The World Health Organization classifies as Visual Impairment the decreased ability of a person to see, to a degree that causes problems not fixable by usual means, such as glasses. The term blindness is used for a severe visual impairment, which causes vision loss [44]. Most of the causes of the Vision impairment are cataracts, glaucoma, age- related macular degeneration, diabetic retinopathy, or Mendelian disorders.

Mendelian disorders are a group of genetically diverse conditions affecting all age groups and ethnic background. These retinopathies affect approximately one in 2000 individuals worldwide, whit an early age of onset, severe and topographic pattern of visual loss, involvement of a specific type of photoreceptor, ophthalmoscopic findings. Over 150 loci for inherited retinal degeneration have been identified, and many of the associated genes have been cloned [45].

Retinal disorders are also phenotypically heterogeneous, such that a single mutation may be associated with different phenotypes within a family or between families, and different mutations within the same gene can cause substantially different retinal disorders.

Inherited photoreceptor degenerations (IPDs) can be distinguished by different characteristics: the mode of inheritance, pattern of visual loss, and by the mutant gene involved in the disorder [46]. In the majority of IPDs both the cones and rods die, but the degree to which each cell type is affected differs among the various disorders.
For example, retinitis pigmentosa (RP) is characterized by the initial loss of rods, leading to night blindness and loss of peripheral vision, followed by the loss of cones, leading to a loss of central vision and blindness.

2.1 Retinitis Pigmentosa

Retinitis pigmentosa (RP), is classically characterized by impaired rod function, a progressive degeneration of the retina beginning in the midperiphery, and a characteristic retinal deposit, the "bone spicule" pigmentary deposit [47].
The degeneration at the beginning spares the central retina, which mediates high-acuity vision, until late in the disease. Eventually, most RP patients lose both rod and cone function. In a minority of patients with RP or RP-like diseases, cone dysfunction occurs early in the disease, a condition referred to as cone-rod dystrophy [48].
The worldwide prevalence of retinitis pigmentosa is about 1 in 4000 for a total of more than 1 million affected individuals [49]. RP can have X-linked (XLRP), autosomal recessive (ARRP), or autosomal dominant (ADRP) modes of inheritance, each form showing both locus and allelic heterogeneity; most cases of RP are monogenic [50].

In humans with RP and in mouse models of RP, photoreceptor cell death occurs by apoptosis [51, 52, 53] as determined by analysis of DNA fragmentation and by the absence of an inflammatory response (thus, the term "retinitis" is misleading). Interestingly, in human RP retinas there is patchy loss of both rod and cone photoreceptors, and in mosaic mouse models of RP, diseased or dying photoreceptor cells induce cell death in adjacent genetically normal photoreceptor cells [54, 55, 56]. The deleterious effect of proximity to defective and/or dying cells is presumably responsible for the eventual loss of cones in those RP patients who carry rod-specific gene defects [57].

2.1.1 Diagnosis

Many patients fall into a classic pattern of difficulties with dark adaptation and night blindness in adolescence and loss of mid-peripheral visual field in young adulthood. Clinically, RP patients are diagnosed based on three main abnormalities:

- Atrophy and pigmentary changes or the retina and RPE: Early in the disease, the pigmented posterior pole of the eye, the fundus, develops a granular appearance, due to pigment granules that accumulate in perivascular clusters, known as "bone spicula"[58];

- Abnormal electroretinogram (ERG): this technique provides an objective measure of retinal function. In the procedure, the retina is dark adapted and then stimulated with a brief flash of light. The summed electrical response of the retina is recorded extraocularly with a contact lens electrode. Typically, patients with RP have reduced rod and cone response and delay in their timing;

- Attenuation of the retinal vasculature and changes to the optic nerve head.

2.1.2 Genetic causes

Retinitis pigmentosa is usually confined to the eye, however, some 20–30% of patients have associated non-ocular disease, and such cases fall within more than 30 different syndromes. Usher's syndrome, in which retinitis pigmentosa is associated with hearing impairment, is the most frequent syndromic form, accounting for about 20–40% of individuals with recessive disease (or 10–20% of all cases). Another major form of syndromic retinitis pigmentosa is Bardet-Biedl syndrome, in which the retinal degeneration is associated with obesity, cognitive impairment, polydactily, hypogenitalism, and renal disease [59, 60].

Mutations in genes preferentially expressed in photoreceptors are the most common cause of RP followed by RPE-specific genes. In rare cases RP is caused by mutations in genes expressed in other retinal cell types or outside the eye. Some of the genes identified among the ones responsible of RP are listed in Table 1.

Some encode for phototransduction cascade, thus the resulting mutated proteins interfere with photoreceptor physiology. Subsequent death of rod photoreceptors is probably an outcome of the deranged physiology associated with the defective or absent gene product. For example, without functional rod cGMP phosphodiesterase, arising with recessive defects in $PDE6\alpha$ or $PDE6\beta$, cGMP concentrations in rod photoreceptor outer segment rise and this in turn opes cGMP-gated channels in hte plasma membrane. Thus, it seems that rod's death is caused by the rush of cations resulting from this un-regulated channels [61].

2.1.3 Mutations in Rhodopsin

Rhodopsin was the first protein found to be mutated in RP. In the autosomal dominant form, approximately 30% of families have mutations in that gene. Over 100 mutations have been found in the RHO gene associated with RP thus far, almost all leading to the production of aberrant protein, with one or few amino acids mutated or deleted. Those mutations affect all three domains of Rhodopsin, namely intradiscal, transmembrane, and cytoplasmatic. The resulting mutants are subdivided into 2 categories[62, 63]:

- Class I mutants (15%), with mutations predominantly found in the first transmembrane domain and near the carboxyl terminus of the protein. Those rhodopsin mutants either accumulate in the cell body or in extracellular vesicles. The abnormal function results in faster activation kinetics, which could play a role in RP by altering the stoichiometric balance of the different proteins involved in the phototransduction biochemical reactions;
- Class II mutants (85%), in which is found interference with the folding and/or stability of the protein, that tend to accumulate in the rough endoplasmic reticulum. These mutations are found mainly in the transmembrane and extracellular domains.

Mutations involving the C-terminus of the RHO gene usually lead to a more severe prognosis because of the functions of the C-terminus of Rhodopsin in cellular transport. Indeed, Codon 347 at the C-terminus is a mutational hot spot, with five disease-causing sequence variations identified at this locus and more severe phenotype associated [64].

2.2 Gene Therapy

Gene therapy aims at delivering corrective genetic material to a cell, tissue or target organ in order to prevent or cure a disease. In diseases where the gene defect is known, the wild-type gene could be introduced with a viral vector so that the functional gene product would restore function and/or prevent cell death.

The feasibility of this strategy has some constraint: I) the causative gene has to be known; II) the therapeutic gene has to be cloned into a viral vector; III) the vector safety its ability to transduce the appropriate cells; IV) the underlying diseased tissue having the potential for restoration of function with gene replacement.

The retina represents an ideal target for gene therapy approaches because it is easy to access and manipulate, it is an immunologically privileged tissue, so the immune response against the transgene and the vector is limited; the eye is enclosed in the blood-retinal-barrier of the RPE so this helps avoiding any spread of vectors into other tissues; non invasive technique exist to monitor the treatment (ERG; Optical Coherence Tomography, OCT).

Treatments being explored fo RP included Vitamin A therapy [65] and diltiazem [66], which help to delay the progression of visual loss or the loss of photoreceptors. As specific genes have been identified, there is great interest for using gene therapy to treat RP. Currently, adenovirus, adeno-associated virus, and lentivirus have been used to successfully deliver corrective genes to animal models of RP. Such strategies can be divided in two distinct groups: gene replacement or gene silencing strategies.

Gene replacement

Loss of function mutations are the eligible target for gene replacement [67, 50]. Dominance resulting from inadequate expression level of a gene is a rare condition known as haploinsufficiency. The supply of the missing protein can restore the normal function of the cell. This approach, for example, has been applied to a form of Leber Congenital Amaurosis (LCA), MERTK-associated ARRP [68], and in the Usher Syndrome [69]. These techniques have experimentally been shown to delay and even reverse the course of RP with associated improvement of photoreceptor function in various animal models. ERG response recovery, as well as retinal structural improvement, has been documented in an animal model after gene replacement therapy at an early stage of the disease [70, 71, 50].

Gene silencing

Strategies to treat alterations that lead to gain of function mutations fall in this category. Treatments included mutation-independent approach in which ribozymes, or more recently siRNAs, were designed to target regions of mRNA that are not affected by mutation, so that both wild-type and mutant RNA produced by the disease gene are degraded. [72, 73].

Instead of inhibiting gene expression by degrading mutant mRNAs, a promising alternative is the modulation of gene expression at the transcriptional level by using artificial transcription factors.

Transcriptional regulation of endogenous genes and the precise control of transgene expression are major challenges in gene therapy. Fusion proteins that consist of engineered DNA-binding domains and catalytic effector domains hold great promise for targeted gene regulation under the control of specific cell constitutive promoters.

Artificial Transcription Factors require an effector domain that controls the frequency of transcription initiation at endogenous target genes. These effector domains can be transcriptional activators or repressors, but can also be a chromatin remodeler or epigenetic regulator. Three systems are available for mediating site-specific DNA recognition of artificial TFs: those based on zinc fingers, TALEs, and on the CRISPR/Cas9 technology.

- **ZFPs** are the most frequently used class of transcription factor and account for about 3% of genes in the human genome [74].

  The classical Cys_2His_2 consists of a sequence of about 30 amino acids containing two histidines, two cysteines and three hydrophobic residues, all at conserved positions. It forms a small, independently folded domain stabilized by Zn2+, which can be used in a modular tandem fashion to achieve sequence-specific recognition of DNA.

Each ZF domain interacts with a triplet of consecutive bases on one strand of the DNA through one amino acid residue just before its alpha helix, and two amino acids within its alpha helix. A fourth contact is made with a base on the opposite strand. Changes in the amino acid composition of the alpha helix change the DNA binding specificity [75]. Control of gene expression, with selected combinations of zinc-finger motifs, were already shown to be effective in silencing of target genes. These experiments showed that zinc-finger DNA-binding domains can be engineered de novo to target given promoter DNA sequences, and, by fusing them to an effector domain, to regulate their activity [74].

- Engineered transcription activator-like effectors (**TALEs**) can also be used for targeted gene expression. TALEs are originally produced by the bacterial pathogen *Xanthomonas* and are injected into plant cells, where they bind to the regulatory regions of specific plant genes, activating their transcription [76].

  The core DNA binding domains of TALEs consist of repeats of modules of 34 amino acids that each bind to 1 bp of DNA. However, the recognition of DNA by TALE domains was shown to still be more complex than that simple formula [77].

- The most recently discovered DNA binding domains were found in the **CRISPR/Cas9** system (Clustered Regularly Interspaced Short Palindromic Repeats), which is a defense system employed by a range of bacterial species aimed at the degradation of viral DNA [78]. In this system, specific guide RNAs direct the Cas9 endonuclease protein to their target DNA sequence, leading to subsequent cleavage of that sequence.

The guide RNAs base pair with complementary DNA sequences at their 5' end, and interact with Cas9 through their 3' end [79]. The length of the homology-searching RNA sequence is usually about 20 bases, but shorter sequences have recently been reported to have less off-target effects [80].

The CRISPR/Cas9 system has made an extremely rapid entry into biotechnology, predominantly for making site-specific double strand breaks and thereby targeted mutations within a genome, analogous to zinc finger nuclease and TALEN technology. Derivatives of the Cas9 protein lacking nuclease activity (dCas9) can also be made amendable for generating artificial TFs. Induction of gene expression was achieved via dCas9 fusions to the powerful transcriptional activator VP64 [81, 82]. Specific repression was observed by targeting just a dCas9 protein to potentially regulatory target sites [83].

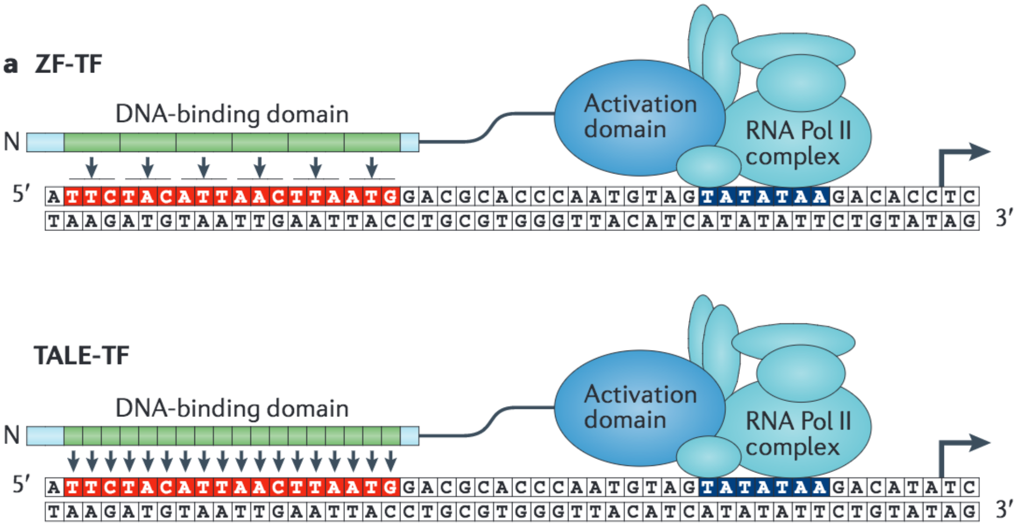

a ZF-TF
DNA-binding domain
N
Activation domain
RNA Pol II complex
5′
ATTCTACATTAACTTAATGGACGCACCCAATGTAGTATATAAGACACCTC
TAAGATGTAATTGAATTACCTGCGTGGGTTACATCATATATTCTGTATAG
3′
TALE-TF
DNA-binding domain
N
Activation domain
RNA Pol II complex
5′
ATTCTACATTAACTTAATGGACGCACCCAATGTAGTATATAAGACATATC
TAAGATGTAATTGAATTACCTGCGTGGGTTACATCATATATTCTGTATAG
3′

Figure 2.1: Artificial transcription Factors based on TALE or ZF structure. Adapted from [84]

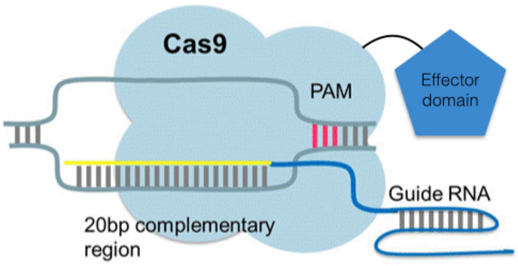


Figure 2.2: Artificial transcription Factors based on CRISPR/Cas9 structure.

The target site of the artificial TFs is usually a region in the promoter of the gene of interest. The eukaryotic core promoter is defined as the region that can be bound by the general transcription factors required for RNA polymerase II-dependent transcription initiation at the transcription start site, and it is the most obvious target sequence to design artificial TFs.

In Prof. E.M. Surace's lab, Mussolino et al. [85] showed that an artificial transcription factor, a Zinc Finger protein with a KRAB regulatory domain, was able to bind *Rhodopsin* promoter and repress its expression in a mouse model of adRP, with improvements in ERG of those mice. In the present work we investigate whether engineered DNA binding proteins without canonic effector domain possess transcriptional repression properties, and assess their functional interference with a genome-wide approach.

Inheritance	Gene	Location	Function
ADRP	CRX	19q13.32	photoreceptor cell transcription factor
ADRP	FSCN2	17q25	morphologic structures of photoreceptor cells
ADRP	HPRP3	1q21.2	no clear functional role
ADRP	IMPDH1	7q32.1	regulate the cell growth
ADRP	NRL	14q11.2	photoreceptor cell transcription factor
ADRP/ARRP	PDC	1q25-32.1	visual transduction cascade
ADRP	PRPF8	17p13.3	No clear functional role
ADRP	PRPF31	19q13.42	pre-mRNA slicing
ADRP	RDS	6p21.2	photoreceptor structure
ADRP/ARRP	RHO	3q22.1	visual transduction cascade
ADRP	ROM1	11q12.3	photoreceptor structure
ADRP	RP1	8q12.1	transcription factor
ADRP	RP9	7p14.3	no clear functional role
ADRP	RP17	17q22	pre-mRNA slicing
ARRP	ABCA4	1p22.1	catabolic function in the retina
ARRP	CNGA1	4p12	visual transduction cascade
ARRP	CNGB1	16q13	visual transduction cascade
ARRP	CRB1	1q31.3	transcription factor
ARRP	LRAT	4q32.1	retinoid metabolism
ARRP	MERTK	2q13	disc shedding
ARRP	NR2E3	15q23	ligand-dependant transcription factor
ARRP	PDE6A	5q33.1	visual transduction cascade
ARRP	PDE6B	4p16.3	visual transduction cascade
ARRP	RGR	10q23.1	retinoid metabolism
ARRP	RLBP1	15q26.1	retinoid metabolism
ARRP	RPE65	1q31.2	retinoid metabolism
ARRP	SAG	2q37.1	visual transduction cascade
ARRP	TULP1	6p21.31	photoreceptor cell transcription factor
ARRP	USH2A	1q41	retinal development
ARRP	RP22	16p12.1p12.1	mapped gene not cloned
ARRP	RP25	6cen-q15	mapped gene not cloned
ARRP	CERKL	2q31-q33	ceramide metabolism
ARRP	RP28	2p16-p11	mapped gene not cloned
ARRP	RP29	4q32-q34	mapped gene not cloned
XLRP	RP2	Xp11.23	protein folding
XLRP	RPGR	Xp11.4	protein transport
XLRP	RP6	Xp21.3-21.2	mapped gene not cloned
XLRP	RP23	Xp22	mapped gene not cloned
XLRP	RP24	Xq26-27	mapped gene not cloned

Table 2.1: List of genes involved in RP. Modified from [49]

Exon	Codon change	Exon	Codon change	Exon	Codon change
1	T4K	2	Y136X	3	M216R
1	N15S	2	V137M	3	M216K
1	T17M	2	C140S	3	F220C
1	P23L	2	A164V	3	C222R
1	P23H	2	A164E	4	4162del3bp
1	Q28H	2	C167R	4	4188del3bp
1	L40R	2	C167W	4	P267L
1	M44T	2	P171E	4	P267R
1	F45L	2	P171S	4	S270R
1	L46R	2	P171L	4	T289P
1	G51R	2	P171Q	4	K296E
1	G51V	2	E150K	4	K296M
1	G51A	2	G174S	4	S297R
1	P53R	3	Y178N	4	Q312X
1	T58R	3	Y178C	4	E249X
1	Q64X	3	P180A	4	G284S
1	496del12bp	3	E181K	5	5168del9bp
1	V87D	3	G182S	5	L328P
1	G89D	3	Q184P	5	5225del17bp
1	G90D	3	S186P	5	998ins4bp
1	G106R	3	S186W	5	5255del24bp
1	G106W	3	C187Y	5	5256delC
1	G109R	3	G188R	5	5258del8bp
1	C110Y	3	G188E	5	T342M
1	C110F	3	D190N	5	Q344X
1	G114D	3	D190Y	5	V345L
1	G114V	3	D190G	5	V345M
2	L125R	3	T193M	5	A346P
2	S127F	3	M207R	5	P347T
2	L131P	3	V209M	5	P347A
2	R135G	3	H211R	5	P347S
2	R135W	3	H211P	5	P347Q
2	R135L	3	P215T	5	P347L
2	R135P			5	P347R

Table 2.2: Most frequent Rhodopsin mutations. Modified from [49]

Chapter 3

Next Generation Sequencing methods for Gene Expression Profiling

Gene expression profiling is the description of the pattern of the genes actively transcribed under specific circumstances, in entire tissues or cells, to give a global picture of the ongoing processes in a context of interest. Techniques used to measure that include "direct" methods like DNA microarrays and sequence based techniques like serial analysis of gene expression (SAGE), which measure the relative activity of previously identified target genes. technologies and offered a limited ability to fully catalog and quantify the diverse RNA molecules that are expressed from genomes over wide ranges of levels [86]. Next generation sequencing (NGS) methods completely revolutionized the gene expression profiling, allowing simultaneous characterization of thousands of sequences, with or without a specific target.

3.1 RNA-Sequencing

The power of sequencing RNA lies in the fact that the twin aspects of discovery and quantification can be combined in a single high-throughput sequencing assay called RNA-sequencing (RNA-seq). Several advantages render RNA-seq preferable over microarray techiques:

- RNA-Seq is not limited to detecting transcripts that correspond to existing genomic sequence.
- It does not have an upper limit for quantification, which correlates with the number of sequences obtained.
- It has a large dynamic range of expression levels over which transcripts can be detected;
- It is highly accurate for quantifying expression levels, as determined using quantitative PCR (qPCR) [87] and spike-in RNA controls of known concentration [88, 89].
- RNA-seq data are highly replicable, comparable, and in some ways superior, compared to existing array-based approaches [90].

Many variations of RNA-seq protocols and analyses have been published, making it challenging to appreciate all of the steps necessary to conduct an RNA-seq study properly [91].

There is no optimal pipeline for the variety of different applications and analysis scenarios in which RNA-seq can be used. Hence, a RNA-seq experiment accurately designed depending on the organism being studied and the research goals is arguably the most important step for the success of the analysis.

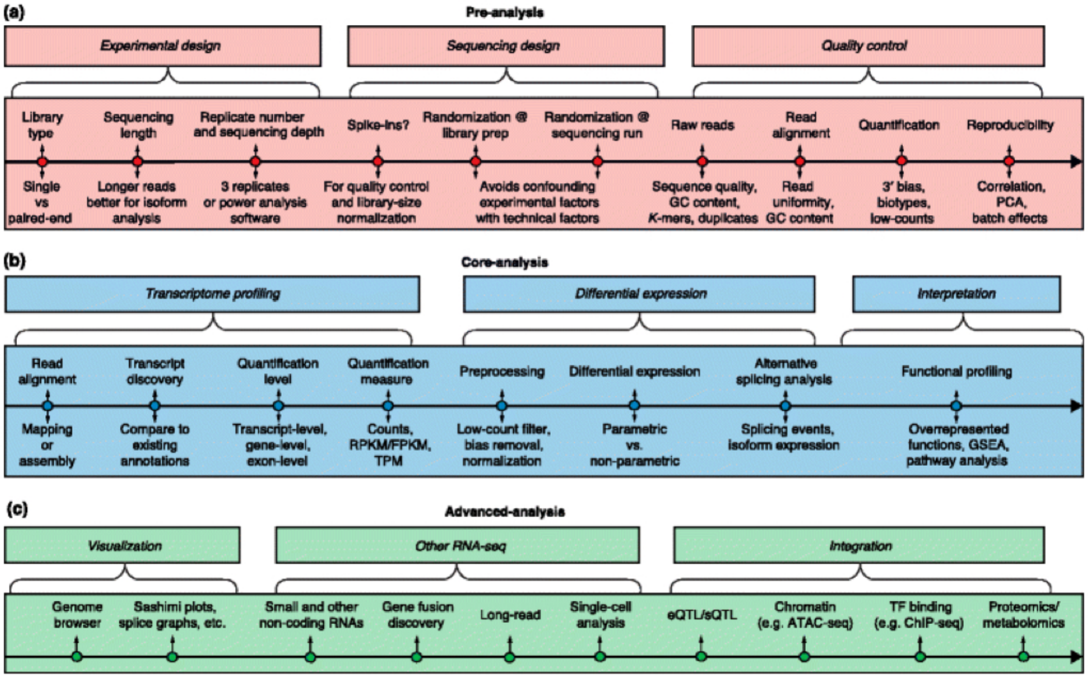

(a)
Pre-analysis
Experimental design
Sequencing design
Quality control
Library type
Sequencing length
Replicate number and sequencing depth
Spike-ins?
Randomization @ library prep
Randomization @ sequencing run
Raw reads
Read alignment
Quantification
Reproducibility
Single vs paired-end
Longer reads better for isoform analysis
3 replicates or power analysis software
For quality control and library-size normalization
Avoids confounding experimental factors with technical factors
Sequence quality, GC content, K-mers, duplicates
Read uniformity, GC content
3' bias, biotypes, low-counts
Correlation, PCA, batch effects
(b)
Core-analysis
Transcriptome profiling
Differential expression
Interpretation
Read alignment
Transcript discovery
Quantification level
Quantification measure
Preprocessing
Differential expression
Alternative splicing analysis
Functional profiling
Mapping or assembly
Compare to existing annotations
Transcript-level, gene-level, exon-level
Counts, RPKM/FPKM, TPM
Low-count filter, bias removal, normalization
Parametric vs. non-parametric
Splicing events, isoform expression
Overrepresented functions, GSEA, pathway analysis
(c)
Advanced-analysis
Visualization
Other RNA-seq
Integration
Genome browser
Sashimi plots, splice graphs, etc.
Small and other non-coding RNAs
Gene fusion discovery
Long-read
Single-cell analysis
eQTL/sQTL
Chromatin (e.g. ATAC-seq)
TF binding (e.g. ChIP-seq)
Proteomics/ metabolomics

Figure 3.1: A roadmap for RNA-seq computational analyses, from[91].

Current RNA-seq methods rely on cDNA synthesis from a population of RNA (total or fractionated, such as poly(A)+). The cDNA fragments have adaptors attached to one or both ends necessary for sequencing. This collection of cDNA (referred to as library) is then amplified and sequenced in high-throughput manner producing millions of short sequence reads typically 30-400 bp long, that correspond to individual cDNA fragments. Following sequencing, the resulting reads are either aligned to a reference genome or transcriptome, or assembled de novo without the genomic sequence. That produces a genome-scale transcription map consisting of both the transcriptional structure and/or level of expression for each gene [92].

RNA extraction and library preparation

The protocol for RNA-extraction and library preparation depends on the underlying biological question. Typically the total RNA is enriched for messenger RNA (mRNA). This can be done by either directly selecting mRNA or by selectively removing ribosomal RNA (rRNA) ; Poly(A) selection requires a relatively high proportion of good quality mRNA, which normally yields a high fraction of reads falling onto known exons [92].

Another important determinant of the quality of the sequencing is the size of the cDNA fragments obtained in this phase: a good fragmentation of the starting RNA is crucial for proper sequencing and subsequent analyses, since larger fragments improve the mappability and de novo transcript identification. Furthermore, single-end (SE) or paired-end (PE) sequencing can be used, the best solution depending on the analysis' needs.

Other techniques which aim at characterizing different RNA populations or aspects of the transcriptome employ specific protocols, for example small RNA profiling (sRNA-seq [93], miRNA-seq [94]), mapping of transcription start sites using CAGE-sequencing [95], strand specific RNA-seq [96], and others [86].

Although few steps are required in the preparation of a RNA-Seq sample, it does involve several manipulation stages during the production of cDNA libraries, which can complicate its use in profiling all types of transcript.

Sequencing depth

Another important factor is the sequencing depth or library size, which is the number of sequenced reads for a given sample. More transcripts will be detected and their quantification will be more precise as the sample is sequenced to a deeper level. Again, there's no unique solution for all types of analyses which can be performed: a higher depth can be required to capture poorly characterized isoforms, while some authors will argue that as few as 5 million mappable reads are sufficient to detect medium to highly expressed genes in most eukaryotic transcriptomes [91]. Therefore, optimal library size depends on the complexity of the targeted transcriptome.

Number of replicates

A crucial design factor is the number of samples required for the analysis. That depends from both the technical variability of an RNA-seq procedure and the biological variability of the system under study. Those considerations are specifically relevant when designing an experiment to detect genes significantly differentially expressed between two conditions (i.e. treated samples vs. controls). It is a good standard to use at least three replicates for biological conditions, but the higher the number of replicates, the better the estimates of within-group variance, which could affect proper identification of genes differentially expressed among different conditions under study [97].

In general, increasing the number of replicate samples significantly improves detection of lowly expressed genes and statistical power over increased sequencing depth [97, 98, 99].

3.2 Analysis of RNA-Seq Data

Quality

Multiple checks need to be performed in each step of the acquisition of RNA-seq data (raw reads, alignment and quantification) in order to assess the quality of the data.

Quality control on raw data include analysis of sequence quality, GC content, duplication rate, over-represented k-mers, which are indicators of sequencing quality and error-rate, PCR artifacts or contaminations.

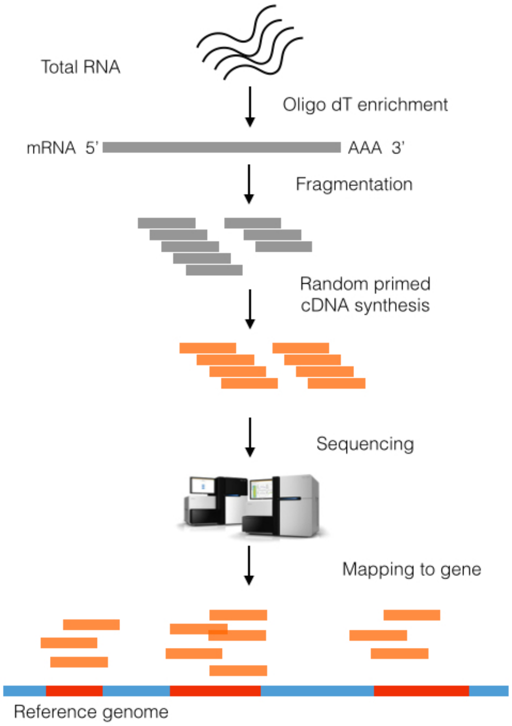


Figure 3.2: A typical RNA-seq experiment.

Those parameters have acceptable ranges within species and experiments, but need to be homogeneous for samples of the same experiment. All of those parameters can be checked using tools such as Picard [100], FASTQC [101]. As a general rule of sequencing, the quality of the reads decreases towards the 3' end. To improve mappability those bases are removed using tools such as Trimmomatic [102] or TrimGalore! [103], which are useful also to trim adaptor sequences from the reads.

Another important checkpoint is the percentage of reads mapping on the reference transcriptome or genome. An high percentage of uniquely mapped reads with low number of multi-mapping reads (those mapping to multiple points of the reference) is an indicator of good quality. Uniformity of coverage is also a parameter to take into account, as non-uniformity could indicate low quality of the starting material.

Finally, it is important to assess the reproducibility among replicates. A high correlation level among technical replicates is desirable, however no clear standard exists for biological samples, which should cluster together in a Principal Component Analysis (PCA) [91].

Alignment

When a reference sequence is available, two alternatives are possible: mapping to the genome or to the transcriptome. Important parameters to consider are the strandedness of the RNA-seq library, the number of mismatches to accept and the length and type of reads. Mapping to a reference genome allows for the identification of novel genes or transcripts, but it's computationally more difficult and requires more time, while mapping to the transcriptome is faster, but doesn't allow discovery of new transcripts. Depending on the type of analysis, many softwares are available: TopHat [104], STAR [105], Bowtie [106] among the most famous. Those algorithms have each one a distinguishing model and optimize memory requirements at their own needs. For example, Bowtie owes its success to its memory-efficient data structure borrowing a method from data-compression, the Burrows–Wheeler transform [107], to index the reference genome, and it allows to scan reads against a mammalian genome using around 2 GB of memory [106].

One of the challenges when searching for novel transcripts is that short reads rarely span across splice junction, making it difficult to infer a full-length transcript. TopHat, one of the most famous mappers, follows a two step strategy in which unspliced reads are first mapped to locate exons, then unmapped reads are split and aligned independently to identify exon junctions [91, 104].

When a reference genome is not available, RNA-seq reads can be assembled de novo. The most famous methods (Trinity[108], SOAPdenovo-Trans [109]) use a step wise approach to assemble contigs and merge them to construct the sequence scaffold. Transcriptome assemblers must recover an unknown number of RNA sequences, typically on the order of tens of thousands, and usually require longer reads and high sequencing depth [91].

Transcript quantification

The most common application of RNA-seq analysis is to quantify gene or transcript expression. Gene-level quantification approaches utilize a gene transfer format (GTF) file containing the coordinates of the genome of interest. The transcript expression is therefore function of the raw reads that map on each transcript sequence.

Raw read counts, however, are not sufficient to compare expression level among samples. Measures like RPKM (Reads per Kilobase of exon model per million reads) FPKM (Fragments per Kilobase of exon model per million mapped reads, PE analogous of RPKM) or TPM (Transcripts per million) are within-sample normalization methods that account for the transcript length and the library size-effect.

Correcting for gene length is not required when comparing changes in expression in the same gene across multiple samples, but is necessary to correctly assess gene expression accounting for the fact that longer transcripts will obtain more reads in sequencing phase.

Algorithms that quantify expression from transcriptome mapping include RSEM [110], Cufflinks [111] both using an expectation maximization approach to obtain the final count estimates for each transcript.

In particular, the RSEM approach consists of using a set of reference transcript sequences, such as one produced by a de novo transcriptome assembler. As it does not rely on the existence of a reference genome, it is particularly useful for de novo transcript quantification. In its default mode as a first step, one can supply RSEM with a FASTA-formatted file of reference transcript sequences, or a GTF and the full genome sequence (in FASTA format). The second step, for the calculation of transcript abundances, relies on Bowtie [106] for read alignment.

However, RSEM can accept other user-provided aligners, whose BAM/SAM output is then submitted to the -rsem-calculate-expression function as in the default pipeline. After the alignment of reads, RSEM computes Maximum Likelihood abundance estimates using the Expectation-Maximization (EM) algorithm, and as result it gives in output the gene/isoform abundance estimates and credibility intervals [110].

3.2.1 Differential Expression Analysis

The expression level of each RNA unit is measured by the number of sequenced fragments that map to a transcript, which is expected to correlate directly with its abundance level. [97].

The primary goal of Differential Expression Analysis is therefore to quantitatively measure differences in the levels of transcripts between two or more treatments or groups. Differential gene expression analysis of RNA-seq data generally consists of three components: normalization of counts, parameter estimation of the statistical model and testing for differential expression. Among the methods available, edgeR [112, 113] and DESeq/DESeq2 [114, 115] are the most widely used.

A common starting point for DEA methods is a count matrix N of n rows (genes) x m columns (samples) where N_{ij} is the number of reads assigned to gene i in sequencing experiment j.

Modeling of gene expression

An RNA-seq experiment consists of a random sampling of reads from a fixed pool of genes. Thus a natural representation of read counts associated to each gene can be the count-based Poisson and Negative Binomial distributions. Initial methods for DE testing used Poisson distribution [90]. It is important to notice that Poisson distribution models both mean and variance using a unique parameter. However, in biology the variance of gene expression across multiple biological replicates is larger than its mean expression values, a problem known as *over-dispersion* [97].

Indeed, both *DESeq, EdgeR* models assume that the number of reads in sample j that are assigned to gene i can be modeled by a negative binomial (NB) distribution.

$$K_{ij} \sim NB(\mu_{ij}, \phi) \tag{3.1}$$

With mean μ_{ij} and variance $\phi = \sigma_{ij}^2$. The relation between the variance ϕ and mean μ is generally defined as $\phi = \mu + \alpha\mu^2$ where α is the dispersion factor.

Estimation of this factor is one of the fundamental differences between edgeR and DESeq. edgeR estimates α as a weighted combination of two components: a gene-specific dispersion effect and a common dispersion effect calculated from all genes. DESeq, on the other hand, breaks the variance estimate into a combination of the Poisson estimate (that is the mean expression of the gene) and a second term that models the biological expression variability [97].

Normalization

Normalization procedures attempt to account for differences among samples such as depth between different sequencing runs or technical biases in library production protocols, to facilitate accurate comparisons between sample groups. One crucial problem is that the proportional representation of each gene is dependent on the expression levels of all other genes. Often a small fraction of highly expressed genes account for large proportions of the sequenced reads, skewing the counts distribution, thus removing power to detect changes in expression of low expressed genes.

Both mean and variance in (3.1) rely on a size factor, s_j, which represents the coverage, or sampling depth, of library j. *DESeq* computes the scaling factor for a given sample by computing the median of the ratio, for each gene, of its read count over its geometric mean across all samples. The purpose of the size factors s_j is to render counts from different samples, which may have been sequenced to different depths, comparable.

The trimmed means of M values (TMM) from Robinson and Oshlack [25], which is implemented in *edgeR*, computes a scaling factor between two experiments by using the weighted average of a subset of genes after excluding genes that exhibit high average read counts and genes that have large differences in expression. When using more than several samples, the scaling factor can be calculated by selecting one sample as a reference and calculating the TMM factor for each non-reference sample [116].

Testing for differential expression

The differential expression analysis aims at testing whether there is a significant difference in expression of a (set of) gene(s) between two conditions. This task can be effectively achieved only if the gene-wise dispersion parameter was accurately estimated.

DESeq2 implements the assumptions of its predecessor method, *DESeq*, and to model the dispersion it assumes that genes of similar average expression strength have similar dispersion. This assumption is crucial to overcome the limitation that in most high-throughput sequencing experiments the number of samples is generally low, affecting the within-group variance estimates. Indeed, noisy estimates of the variance would compromise the accuracy of differential expression testing [114] [115]. *EdgeR* moderates the dispersion estimate for each gene toward a common estimate across all genes, or toward a local estimate from genes with similar expression strength, using a weighted conditional likelihood.

For each gene, a Generalized Linear Model (GLM) [117] is fitted,

$$\mu_{ij} = s_j q_{ij} \tag{3.2}$$

$$\log_2 q_{ij} = x_{j.}\beta_i \tag{3.3}$$

where the fitted mean μ_{ij} is composed of a parameter, q_{ij}, which is proportional to the true (unknown) number of fragments of RNA for sample j. The coefficients β_i give the $\log_2$ fold changes for gene i for each column of the model matrix X.

Data transformation

For analyses on count data other than DEA - visualization or clustering – it might be useful to work with transformed versions of the data, so that they become homoskedastic. To this end, in *DESeq2* two functions are implemented, *rlog* and *Variance-stabilizing transformation.*

Both produce transformed data on the $\log_2$ scale which has been normalized with respect to library size. The point of these two transformations is to remove the dependence of the variance on the mean, particularly the high variance of the logarithm of count data when the mean is low. Both rlog and VST use the experiment-wide trend of variance over mean, in order to transform the data [118].

As suggested, the most obvious choice of transformation for count data is the logarithm. Since count values for a gene can be zero in some conditions, some advocate the use of pseudocounts, i. e. transformations of the form

$$y = log_2(n + n_0) \tag{3.4}$$

where n represents the count values and n_0 is a positive constant.

Regularized log transformation

The function rlog, (*regularized log*), transforms the original count data to the $\log_2$ scale by fitting a model with a term for each sample and a prior distribution on the coefficients which is estimated from the data. This is the same kind of shrinkage (referred in the paper as moderation) of log fold changes used by the DESeq and nbinomWaldTest. The resulting data contains elements defined as:

$$\log_2(q_{ij}) = \beta_{i0} + \beta_{ij} \tag{3.5}$$

where *qij* is a parameter proportional to the expected true concentration of fragments for gene i and sample j, β_{i0} is an intercept which does not undergo shrinkage, and β_{ij} is the sample specific effect which is shrunk toward zero based on the dispersion-mean trend over the entire dataset. The trend typically captures high dispersions for low counts, and therefore these genes exhibit higher shrinkage from the *rlog*.

Variance stabilizing transformation

It's the other function that the DESeq2 package provides alongside with the *rlog* function to deal with count data transformation and break the mean-variance dependence. Compared to *rlog*, *varianceStabilizingTransformation* is more sensitive to size factors, which can be important to consider when those vary widely within the dataset. This function calculates a variance stabilizing transformation (VST) from the fitted dispersion-mean relation(s) and then transforms the count data (normalized by division by the size factors or normalization factors), yielding a matrix of values which are now approximately homoskedastic.

All the functions described here are implemented in stable packages in R/Bioconductor [119].

3.3 Single Cell RNA-Seq

A recent development of RNA-Seq, single cell RNA-Seq is revolutionizing the way in which we look at the expression profile of complex tissues. Generally, profiling of gene expression can be achieved from bulk population of millions of input cells, meaning that the resulting value for each gene is an average of its expression level across the cells composing the specimen. For some studies, bulk approaches can be insufficient [120].

To date, measurements of the expression of a gene at the single-cell level were generated using low throughput approaches, such as RNA-FISH [121], single cell qPCR [122, 123]. Dramatic changes in gene expression can occur within the cells composing a tissue, like the brain [124], the retina [6], the thymus [125], and pathological changes which originate from distinct cells, such as infections, or cancer clonal cells can now be addressed with high throughput methods [126].

Single-cell RNA-seq requires the successful combination of two independent techniques: the isolation of individual cells from culture, tissue or dissociated cell suspensions, and, after the conversion of the extremely low amount of cellular RNA into cDNA, parallel sequencing of cDNA libraries [127].

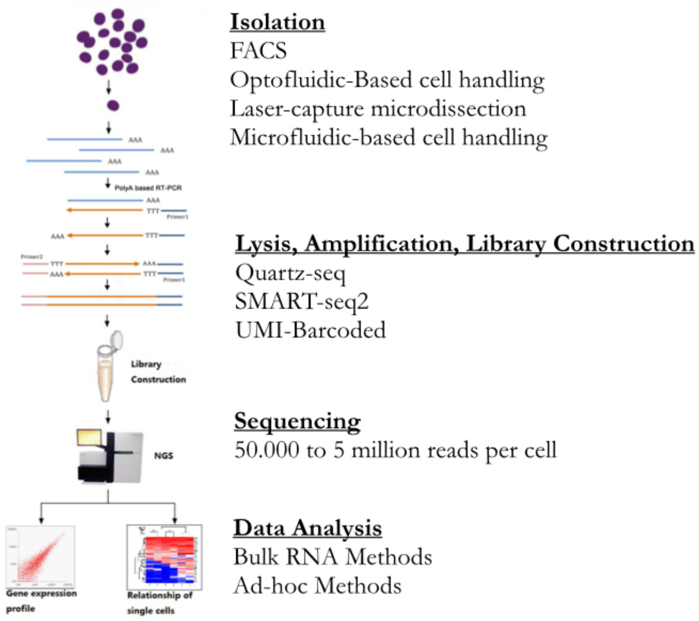

Isolation
FACS
Optofluidic-Based cell handling
Laser-capture microdissection
Microfluidic-based cell handling
AAA
PolyA based RT-PCR
TTT
Primer1
Primer2
Lysis, Amplification, Library Construction
Quartz-seq
SMART-seq2
UMI-Barcoded
Library Construction
NGS
Sequencing
50.000 to 5 million reads per cell
Data Analysis
Bulk RNA Methods
Ad-hoc Methods
Gene expression profile
Relationship of single cells

Figure 3.3: Single-cell RNA-seq pipeline, from cell isolation to analysis methods.

The most common way to isolate cells is the Fluorescence Activated Cell Sorting, in which the cells of interest are marked using a fluorescent antibody against a specific epitope expressed on the cells of interest, or are transfected/transduced with a fluorescent construct. Once sorted, the cells can be lysed and subsequently the RNA is retrotranscribed and amplified in

cDNA libraries, which are then sequenced.
Another common strategy exploits microfluidic devices, which let it possible to perform all the steps from cell culture, single-cell isolation to the biochemical steps of cDNA synthesis and detection, thus can easily be automated. Since eukaryotic cells contain many diverse RNA and the ideal strategy would be to exclude tRNA and rRNA, most methods strategies aim at selectively reverse transcribing poly(A)+ RNAs. Many protocols are available, they differ with respect to the strategy used to amplify cDNA that is obtained by oligo(dT)- or/and random-primed reverse transcription [128, 129, 130, 131, 132, 133].

One of the most used protocols is the Smart-seq2, an improvement of the original Smart-Seq [134, 135]. In this protocol, single cells are lysed in a buffer that contains free dNTPS and tailed oligo(dT) oligonucleotides with a universal 5' anchor sequence. During the reverse transcription phase, 2 to 5 untemplated nucleotides are added to the 3'end of the cDNAs. Then, a template-switching oligo (TSO) is added and after the first strand reaction, the cDNA is amplified. Tagmentation is then used to construct the libraries from the total amplified cDNA. This protocol can work with low amounts of starting RNA ($\sim$ 50pg) [136].

Additionally, although some protocols fragment and then sequence the amplified cDNA fragments, it is also possible to sequence reads derived solely from the 3' or 5' end of the amplified transcript. In this case, **unique molecular identifiers (UMIs)** [137] can be used to barcode individual molecules [123]. As long as the complexity of the library of molecules is maintained, the library can be amplified, normalized or otherwise processed without loss of information about the original molecule count because the number of UMIs is a function of the number of molecules in the starting sample. Upon deep sequencing, each UMI will be observed multiple times, and the number of original DNA molecules can be determined simply by counting each UMI [137].

The most obvious experimental design questions related to scRNA-seq experiments are the number of cells that need to be sequenced and the depth to which each individual cell should be sequenced. As in bulk RNA-Seq, both of these questions depend on the biological problem of interest, as well as on technical and financial constraints. As a general rule, generating data from hundreds to thousands of cells may be necessary to identify and characterize subpopulations of cells (especially if rare) or to study the kinetics of transcription [123].

Given the widespread use of bulk RNA-seq, many tools for data analysis and statistical modeling already exist and have been borrowed for scRNA-seq. However there is plenty of room to develop new analytical strategies to specifically process scRNA-seq data.

Some specific aspects of scRNA-seq data have been already considered in some recent work: for example in scRNA-seq data one has to account for the random noise inherent to such data, and for several hidden factors that might result in gene expression heterogeneity. An approach to detect and account for confounding factors in single-cell RNA-seq studies is represented by the **scLVM** method [138], which estimates and correct the data for the source of confounding variation, such as cell-cycle.

To account for technical variation within the data, it is preferable to add to the samples external control sequences **(ERCC)** [88, 139, 140]. Those “spike-ins”, added at a known concentration, can be used to quantify the degree of technical variability across cells and to examine the relationship between technical variation and gene expression molecules. Additionally, by calculating the ratio between the numbers of reads mapped to the spike-in sequences and to the genes from the organism of interest, the relative amount of mRNA contained in each cell can be estimated. Some procedure exists to perform normalization using those external sequences [141].

In conjunction with spike-ins (which are themselves barcoded before amplification), the use of UMI protocols can improve the normalization procedures and the detection of technical biases affecting the experiments [142].

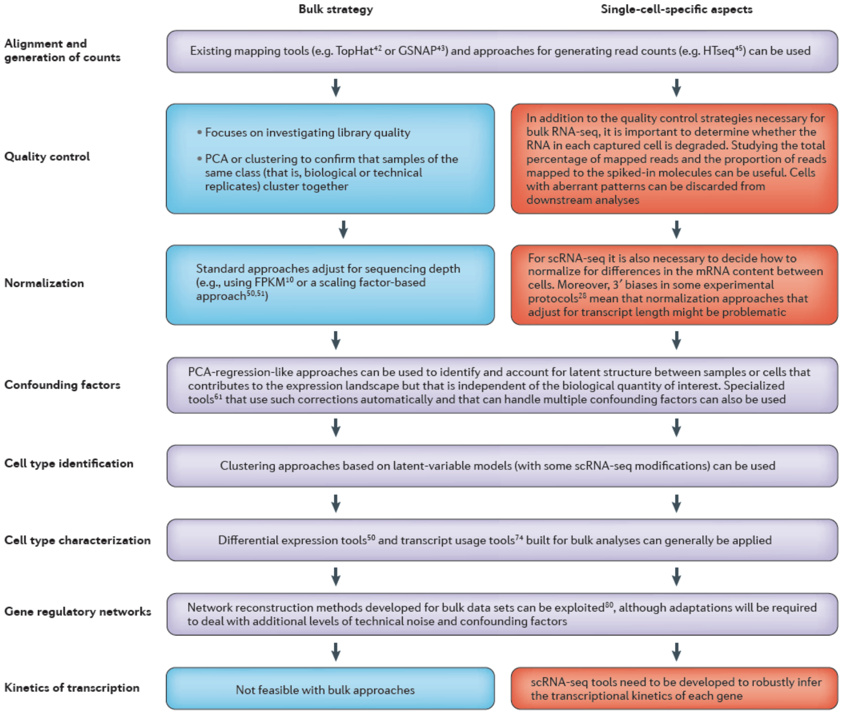

Bulk strategy
Single-cell-specific aspects
Alignment and generation of counts
Existing mapping tools (e.g. TopHat42 or GSNAP43) and approaches for generating read counts (e.g. HTseq45) can be used
Quality control
• Focuses on investigating library quality
• PCA or clustering to confirm that samples of the same class (that is, biological or technical replicates) cluster together
In addition to the quality control strategies necessary for bulk RNA-seq, it is important to determine whether the RNA in each captured cell is degraded. Studying the total percentage of mapped reads and the proportion of reads mapped to the spiked-in molecules can be useful. Cells with aberrant patterns can be discarded from downstream analyses
Normalization
Standard approaches adjust for sequencing depth (e.g., using FPKM10 or a scaling factor-based approach50,51)
For scRNA-seq it is also necessary to decide how to normalize for differences in the mRNA content between cells. Moreover, 3′ biases in some experimental protocols25 mean that normalization approaches that adjust for transcript length might be problematic
Confounding factors
PCA-regression-like approaches can be used to identify and account for latent structure between samples or cells that contributes to the expression landscape but that is independent of the biological quantity of interest. Specialized tools61 that use such corrections automatically and that can handle multiple confounding factors can also be used
Cell type identification
Clustering approaches based on latent-variable models (with some scRNA-seq modifications) can be used
Cell type characterization
Differential expression tools50 and transcript usage tools74 built for bulk analyses can generally be applied
Gene regulatory networks
Network reconstruction methods developed for bulk data sets can be exploited80, although adaptations will be required to deal with additional levels of technical noise and confounding factors
Kinetics of transcription
Not feasible with bulk approaches
scRNA-seq tools need to be developed to robustly infer the transcriptional kinetics of each gene

Figure 3.4: A flowchart of analytical steps in Single-cell and bulk RNA-seq pipeline, from [123]

Nonetheless, some specific aspects of scRNA-seq remain to be fully addressed and will drive next years' developments in that field:

- Normalization of scRNA-seq data must properly account for differences in the total amount of RNA transcribed within a cell and among cells;

- Methods for modeling confounding variables and/or using regression-based analysis to remove them will be required if the biologically relevant signal in scRNA-seq data sets has to be robustly uncovered;

- Accurately modeling technical variability is crucial because without a basic understanding of the underlying noise inherent to scRNA-seq data, downstream interpretation can be misleading or compromised [123].

Chapter 4

Algorithms for Network reconstruction

An inferred gene network is a collection of gene-gene connections captured from expression data. A gene network stores information regarding the relationship among the transcripts and it's helpful to decipher the behaviour of a cell such as the topological organization of its nodes (genes) and the strength of their relationship. A community of genes within a network identifies a group of genes highly connected among them and sparsely connected with genes outside the group. These communities can be used to detect the functional modules in the cell, that is, groups of genes cooperate to accomplish specific functions.

The reconstruction of a Gene Regulatory Network based on experimental data is also called reverse engineering or network inference. Discovering structures and dynamics of GRNs based on large scale data represents a major challenge in systems biology, as the problem itself is of a combinatorial nature (find the right combination of regulators) and available data are often few and inaccurate. Multiple sources of data and network inference methods as well as evaluation metrics for network inference are available.

Even if the model architectures rely on different mathematical structures, all models converge in the representation of a gene regulatory networks as a set of interacting nodes. Nodes are molecular entities such as genes and proteins, or functional modules, whereas edges correspond to regulatory interactions and other relations between those nodes. Due to limitations in the amount and quality of available data and the corresponding computational efforts, network inference methods require simplifications such as linearization, discretization or aggregation of compounds to modules. The usefulness of a GRN inference method mainly depends on both the intended application of identified networks and the data at hand [143].

Computationally inferred interactions therefore offer a useful resource for examining experimental findings into a global context, by finding novel interactions that have yet to be unveiled, unfolding links between pathways under investigation or by identifying the conditions under which a regulator of interest is active, and the state of its interactors [144].

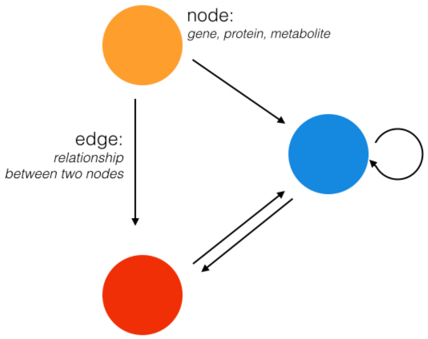


Figure 4.1: A graph, where nodes are genes (or proteins) and an edge is an irreducible relationship between two nodes.

4.1 Associative Networks

Correlation-based methods [145, 146, 147] are the most straightforward way to explore the gene co-expression network. The input data for constructing a gene co-expression network is a $m \times n$ matrix of gene-wise expression measurements of m genes for n samples (conditions).
In first step, the similarity score is calculated between each pair of rows in expression matrix. The resulting, called similarity matrix S, has $m \times m$ dimensions and stores the pairwise correlation coefficients between all the genes. Then either a hard or soft threshold is applied to the similarity matrix to determine the biological meaningfulness of the connections. Associative networks are used to represent pairs of transcripts that coherently change their expression levels across a set of different conditions. These co-expression-based methods have been used in several studies and have shown their usefulness in interpreting biological results and identifying important gene modules.

The Pearson product moment correlation coefficient is a widely used measure of the linear correlation between two variables X and Y, giving a value between -1 and +1 inclusive, where 1 is total positive correlation, 0 is no correlation, and -1 is total negative correlation. It is the covariance of the two variables divided by the product of their standard deviations:

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \tag{4.1}$$

Another measure is the Spearman rank correlation coefficient, it is defined as the Pearson correlation coefficient between the ranked variables. The variables X and Y are converted into ranked scores, *rgX, rgY*:

$$\rho_{rgX,rgY} = \frac{\sigma_{rgX,rgY}}{\sigma_{rgX} \sigma_{rgY}} \tag{4.2}$$

Then, a pruning technique is employed on the correlation matrix to reduce the number of false positive hits. One can choose a hard threshold, for example a value of ρ or a significance level α, which is typically obtained using the t-distribution:

$$t = \rho \times \sqrt{\frac{n-2}{(1-\rho^2)}} \tag{4.3}$$

and the standard error associated:

$$se = \sqrt{\frac{1-\rho^2}{n-2}} \tag{4.4}$$

Two genes are considered linked if their observed correlation level exceeds that corresponding to this significance level.

4.2 Mutual Information

A subcategory of network inference methods are those which infer regulatory interactions between genes based on pairwise mutual information. Those methods make no assumptions about the form of the dependence and in particular they are able to discover also nonlinear relationships among variables. Therefore, those information theory-based methods could outperform correlation based methods if the gene network contains many non-monotonic dependencies. The mutual information (MI) of two random variables is a measure of the mutual dependence between the two variables, it quantifies the "amount of information" obtained about one random variable, through the other random variable. Formally, the mutual information of two discrete random variables X and Y can be defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p_1(x)p_2(y)} \tag{4.5}$$

As a first step, these methods require the computation of the mutual information matrix, a square matrix whose $M_{i,j}$ element is given by the mutual information between X_i and X_j, with X_i and X_j denoting the expression level of genes i and j respectively. Two popular ways of computing MI are: discretizing variables with an equal frequency binning (so that marginal distributions are uniform) [148], and assuming normally distributed variables [149, 150]. Different estimators of MI have been implemented, in the present work we will use some of those implemented in R. A popular algorithm for reverse engineering of gene expression data is ARACNE, which assigns to each pair of interacting nodes a weight equal to their mutual information. If any triplet is found, the algorithm applies a Data Processing Inequality (DPI) step to filter out the weakest edge [151]. Since this approach focuses only on the reconstruction of pairwise interaction networks, a pair of mutually independent genes will never be connected.

4.3 Network validation

Network validation consists of assessing the quality of an inferred model with available knowledge. For quantitative validation of an inferred GRN, it is necessary to employ a scoring methodology that evaluates the model with respect to (a) information already used to generate the model (internal validation) and (b) information independent from the information used to reconstruct the network (external validation) [150]. The assessment of GRN inference algorithms requires benchmark data sets for which the underlying network is known. Benchmarking involves counting the number of links correctly predicted by the algorithm (true positives, TP), the number of incorrectly predicted links (false positives, FP), the number of true links missed in the inferred network (false negatives, FN) and the number of correctly identified non-links (true negatives, TN).

Performance of the algorithm can then be summarized by calculating the true positive rate (TPR= TP/TP + FN) also known as recall, the false positive rate (FPR=FP/FP +TN), and the positive predictive value

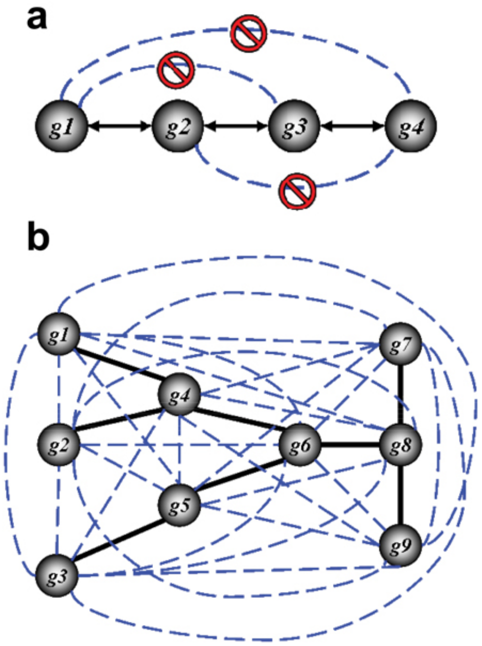


Figure 4.2: DPI processing, from [151]. Although all six gene pairs will likely have enriched mutual information, the DPI will infer the most likely path of information flow. ARACNE will reconstruct the network exactly by removing all false candidate interactions (dashed blue lines) and retaining all true interactions (solid black lines).

(PPV=TP/TP+FP) also known as precision.

The performance can be summarized by: (i) the receiver operating characteristic (ROC) curve, which plots the FPR (equivalent to the specificity) versus the TPR for all thresholds; (ii) the precision-recall (PR) curve, which plots the TPR against the PPV for all thresholds. A well known experimental dataset (gold standard) is the STRING database [152], which stores data on Protein Protein interaction networks on multiple species retrieved from literature and experimental data. For each connection a score from 400 to 900 (lower to higher confidence that the connection is significant) is provided.

However, those gold standard sometimes are hardly available, or contain few link information, especially on non-model species [150, 143].

4.4 Network inference with RNA-Seq data

Co-expression networks use the correlation (or related measures) of gene expression profiles across multiple samples to ascertain common regulation and thus common functions.

Co-expression network analysis with microarrays has been done since the last decades, descending from the first co-expression analyses, there are studies which are highly targeted towards conditions of interest with networks derived from relative small numbers of samples (dozens to the low hundreds) and often focusing on broad network changes in some condition of interest [153]. With increasing amounts of available data, meta-analysis across many data sets became more common, with samples numbering in the many hundreds or thousands.

However, the vast majority of these studies have neglected the growing availability of RNA-seq datasets, which provide several potential advantages over microarrays, as discussed in the previous chapter. The appropriate way to assemble and assess RNA-seq data is still a topic of considerable controversy. An RNA-seq dataset is basically a matrix of gene expression levels expressed in counts, hence data need to be appropriately normalised to minimize inter-libraries differences (depth, batch effect) and intra-library differences (GC content, length of the genes). Then, a measure of co-expression has to be selected, with methods borrowed from existing co-expression analyses [49, 151] and further analysis on network data can be performed, such as clustering, module detection, functional connectivity detection [154, 155, 156, 157]. Typically, the quality of co-expression networks is measured using some variant of the Guilt-By-Association (GBA) principle, or testing the overlap of the network with protein–protein interaction networks.

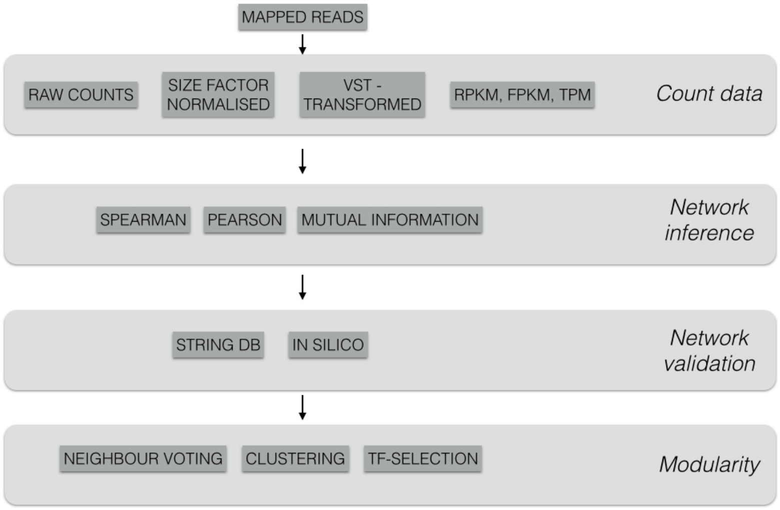


Figure 4.3: A typical pipeline for RNASeq co-expression network

It has been demonstrated in recent studies [158, 159] that basic approaches were among the highest performing, from measuring network connectivity (Spearman correlation) to functional inference (neighbor voting). RNA-seq offers unique virtues but is not a technical panacea. Co-expression analyses still require many samples and a reliable quality reference network as a control and comparative measure. Because of our reliance on pre-existing knowledge (GO) or data (microarray) to provide reference knowledge, we cannot readily assess where RNA-seq application is likely to perform better, for example on novel or more diverse transcript assessment. Unsupervised interpretation of RNA-seq data connectivity is unlikely to be robust to underlying methodological choices without careful filtering, while methods that extract signals using training data (supervised or semi-supervised methods) appear to safely recover known information [158].

In particular, networks inferred from VST-normalized data possess microarray-like behavior with regards to correlation coefficient distribution and topological network properties, perhaps owing to the heteroskedasticity of these types of data [159, 114].

Chapter 5

SusNet: a database for gene co-expression in Porcine Retina

5.1 The Swine Model

Nowadays, at least 90% of genetically modified pigs are generated for biomedical studies. Sequencing and annotation of the pig genome are important milestones to accelerate the generation of transgenic models. Since physiology, anatomy, pathology, genome organization, body weight, and life span of pigs and minipigs are more similar to those of humans, the domesticated pig represents an alternative biomedical model to rodents for specific human diseases.

The swine model has been successfullly employed in a large series of studies, including: cardiovascular diseases, cancer, immunology, cystic fibrosis, diabetes, neurodegenerative diseases, xenotransplantation and ophtalmology [160].

Unlike laboratory rodents or dogs, swine is an appropriate choice as a RP model, since it has morphological and functional similarities, including similar medium- and short-wavelength cone photoreceptors and spatial distribution of rod photoreceptors. The porcine and human eye share many similarities, including a nontapetal fundus with a holangioticvascular pattern and retinal layers of similar thickness. With the exception of the lack of a fovea, the porcine retina shares some significant similarities within the photoreceptor mosaic of humans and other primates [161]. Transgenic pigs expressing P347L mutation in the RHO gene created by pronuclear microinjection showed similar progressive cone photoreceptor degeneration and loss of function in cone-mediated vision as humans. Expression of the most common human P23H mutation in the *Rhodopsin* gene was developed in a minipig model via Somatic cell nuclear transfer (SCNT) and resulted in successful mirroring of the human phenotype of RP. Phenotyping of swine RP models showed photoreceptor loss, disorganized inner and outer segments, and diminished electroretinography responses similar to the human counterpart, suggesting that the transgenic pigs mirror macular degeneration and provide an unique model for therapeutic interventions.

Given the advantages of using the porcine model over the rodent, feline or canine for our purposes, we collected retinae from multiple animal studies conducted in our laboratory and sequenced them by RNA-Seq.

5.2 RNAseq - Data analysis

RNASeq data were generated using Illumina protocols and run on an HiSeq 1500 sequencer (details in the Appendix). Sequence Reads were trimmed using Trim Galore! software (v.0.3.3), that trims low-quality ends and removes adapter from reads, using a Default Phred score of 20. To obtain a precise estimation of this yet uncharacterized tissue, the libraries were aligned on the full transcriptome for Sus scrofa (Pig) as provided by ENSEMBL (SusScrofa 10.2.73). The GTF included the sequences for the 20 canonical chromosomes plus 4563 scaffolds, and counted 30.567 transcripts plus the sequences for the exogenes used in the animal studies. Alignment was performed with RSEM (v.1.2.11) [110] with default parameters. The resulting expected counts (the sum of the posterior probability of each read coming from a specific transcript over all reads) were used for subsequent analysis.

5.3 Network Reconstruction

5.3.1 Dataset: sample selection

We collected 52 samples of pig retinae from different animal studies and runs of RNA-seq. The initial dataset comprised: 11 wild-type retinae (Non-injected, eGFP-injected, Non-transduced areas of treated eyes), 3 samples of sorted Rod photoreceptors, and 38 samples from interference studies. The initial matrix of expression data consisted of 52 samples and 25326 genes. We selected high quality samples based on their concordance within the dataset (Spearman/Pearson correlation score between samples), and Principal Component Analysis.

We removed genes whose raw count estimates were less than 1 Count Per Million, obtaining a final dataset of 52 samples per 16652 genes. To render the libraries comparable despite the unavoidable differences in depth, we normalised the samples using *DESeq2* R-package, using a GLM with one factor and 7 levels (one for each condition considered in the experimental design).

The dataset homogeneity in terms of coverage and depth of sequencing was assessed on raw counts, count estimates normalised by size factor as in DESeq2 default pipeline, *variancestabilizingtransformation* transformed data. No difference was observed among the three count measures, for further analyses we used VST-normalised count data as suggested in previous data [159].

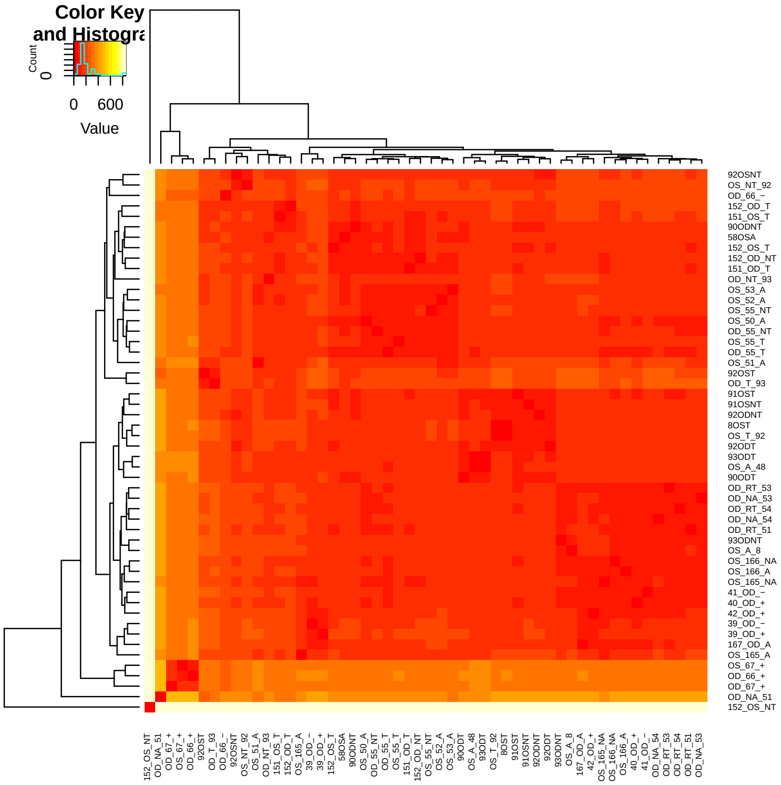


Figure 5.1: Correlation values among 52 samples. A clear cluster of homogeneous samples is present, with 5 samples excluded.

We removed from the original dataset the samples that showed clear distance and low correlation with the biggest cluster comprising 47 samples. Among the removed samples, we found 2 samples which we knew had bad quality after sequencing due to problems in loading phase. Interestingly, the other 3 samples with a clear separation were Rod Photoreceptors sequenced after sorting. This further confirms the uniqueness of photoreceptors' transcriptome if compared to whole retina expression profile. We further filtered the dataset retaining again genes with showed at least 1 CPM in each exper-

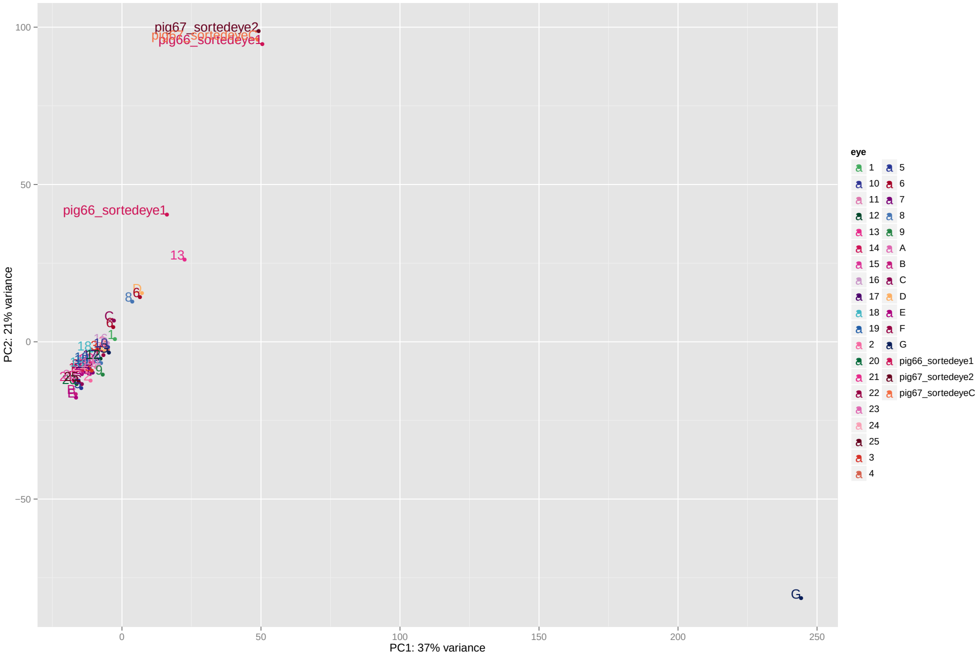


Figure 5.2: PCA of the 52 samples. Aggregation of homogeneous samples on the left, 5 samples fall out of the cluster as previoulsy seen by correlation coefficient.

imental group.
The final dataset was hence composed of 47 samples and 16385 genes.

5.3.2 Co-expression estimates

We estimated co-expression using Spearman, Pearson and Mutual Information.

SCC and PCC - corr.test function To estimate the Pearson/Spearman correlation coefficients (hereafter denoted as PCC and SCC) we used the *corr.test()* from *psych* R package. The function takes in input the matrix of gene expression, uses the *cor()* (*stats* R package) function to find the correla-

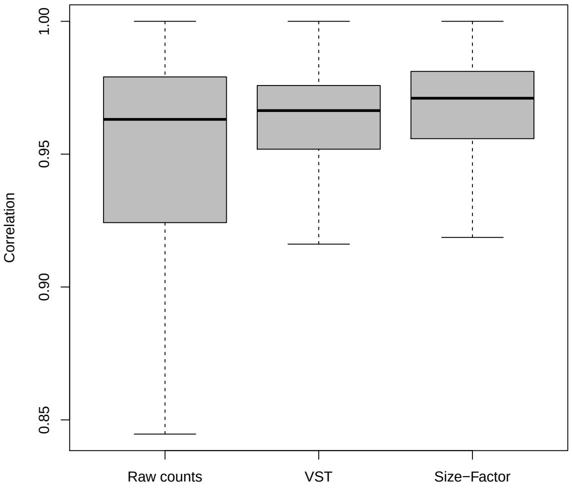


Figure 5.3: Correlation coefficients between samples with the three gene expression measures show no difference.

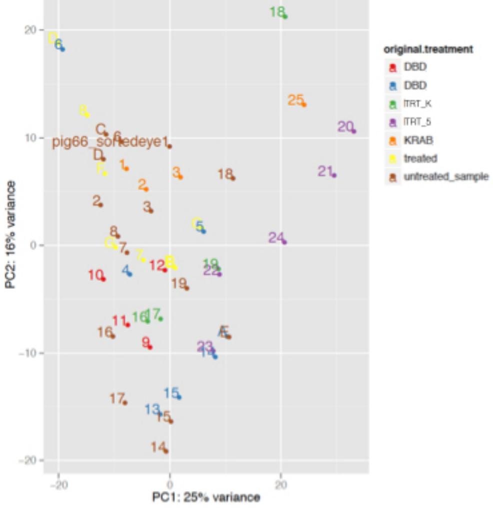


Figure 5.4: PCA of the 47 samples.

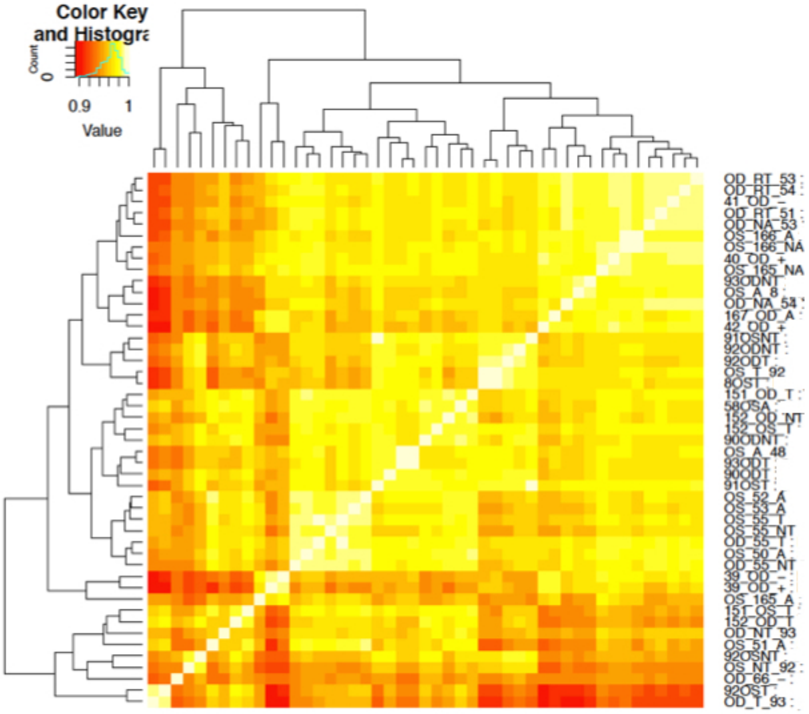


Figure 5.5: Hetamap of the final dataset, all the samples show high level of correlation.

tions, and then applies a t-test to the individual correlations. We computed the PCC and SCC separately, together with corrected pvalues (Benjamini-Hochberg [162]).

We obtained the correlation coefficients of 134 225 920 edges. The absolute value of the correlation values span from 0.34 to 1 in both SCC and PCC. We filtered out correlations whose FDR was higher that 0.05, retaining around 35 % of the correlations in both SCC/PCC.

We then imposed a threshold on the correlation value, retaining correlations above 0.6 (PCC/SCC). This generated two dataset with 7 154 938 edges for PCC and 6 466 521 edges for SCC. Those numbers are consistent with the common procedure of retaining the highest ranking 5% connections.

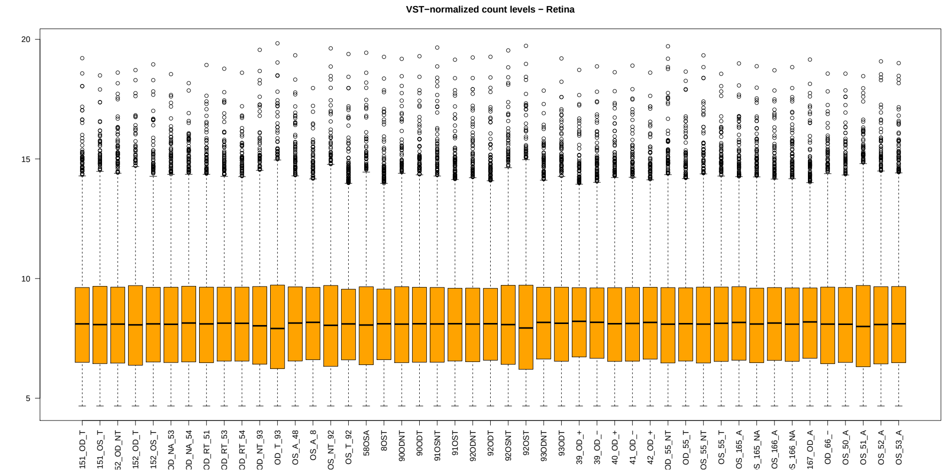


Figure 5.6: Normalised samples

MI: the Parmigene R package We estimated the Mutual Information with the *knnmi()* function of the *parmigene* R package. This function computes the mutual information between all pairs of rows of the gene expression matrix mat using entropy estimates from K-nearest neighbor distances. We then applied the *aracne()* function on MI estimates. We also used an empirical filtering strategy, imposing a threshold to retain the highest 5% correlation, resulting in 6 711 296 connections with minimum MI 0.3.

5.3.3 Performance

For all the network inference methods described, we used the receiver operating characteristic (ROC) curves to study the sensitivity and specificity of each algorithm to minimize the influence of any default thresholds or cutoff values, and the area under the curve (AUC) was used to quantify the performance of each method. As gold standard we used the STRING porcine protein-protein interaction database with cumulative score .900. Our comparative set consisted of 2 519 968 edges and 10305 nodes

SCC method performed better compared to PCC and MI estimates with the empirical threshold, but their AUC values were just slightly different one from another. However, and consistent with previous data [159], the overlap between protein-protein intereaction network and coexpression networks was really low (AUC around 0.56). ARACNE multiplicative, additive or CLR inferred networks had the worst performance (AUC= 0.52 or 0.5) We reintroduced negative PC/SC correlations retaining the absolute value of the correlation above 0.6, but we didn't see any improvements of the performances, suggesting that the positive correlations were driving most the network inference performance. Hence we decided to use only positive correlations.

To evaluate whether we were retaining known strong biological connections, we evaluated the connections among a set ribosomal genes, which are usually strongly correlated. Ribosomal genes showed high level of correlations, confirming the ability of the methods to capture most significant correlations in the data.

We didn't observe any substantial difference among the 3 inferred networks in network topology, degree distribution and overall performance, hence we decided to use the SCC inferred network with 0.6 correlation threshold, a strategy successfully used in previous works [158, 159].

5.3.4 Modularity

We used different methods to test the modularity of the inferred SCC network: the *igraph* R package offers several algorithms, among these we tested the *fast.greedy* which has a hierarchical approach and the *spinglass.community* based on the statistical physics Pott model. We also tested the Affinity Propagation clustering method [163] and a hierarchical clustering computing the jaccard distance on the correlation matrix. The algorithms produced ap-

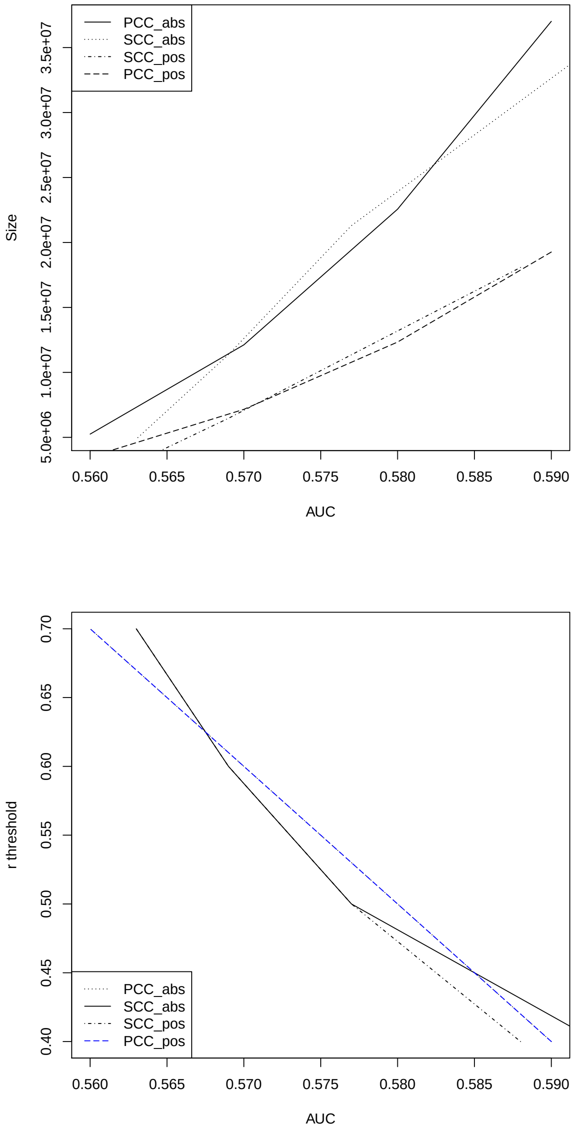


Figure 5.7: Number of nodes compared to AUC and Correlation threshold compared to AUC for SCC and PCC inferred networks. “abs”: absolute value of correlation, “pos”: positive value of correlation.

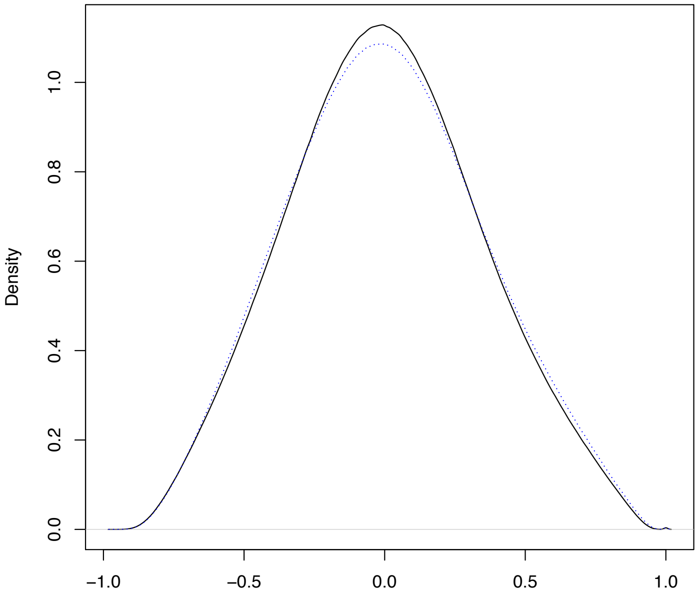


Figure 5.8: Density distributions of SCC (solid black line) and PCC (dotted blue line) inferred correlations.

proximately the same number of clusters (around 300), with 3-5 the clusters containing 90% of the connections. This result suggests that connected component of the network was really strong and we discarded those clustering methods because large clusters ($\geq$ 1000 edges) were non informative in terms of specific biological processes.

We then tried another approach using a master regulator analysis. We searched within the genes of the network those which were known Transcription Factors, using the GO term “transcription factor activity” we retrieved 784 genes. Among these we found the known retina specific TF *Crx, Nrl* [164]. Then we retrieved the nodes connected to each of this TF (hereafter called ‘regulons’).

To assess whether the newtork was capturing specific retinal processes and in particular photoreceptorial ones, we selected the regulons which where ex-

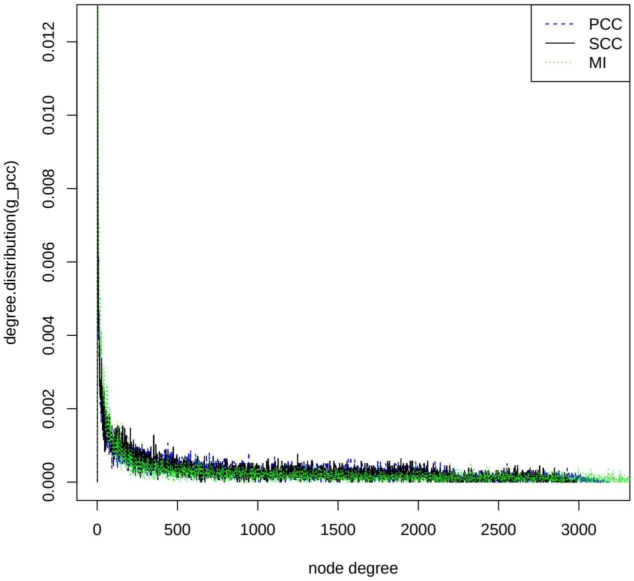


Figure 5.9: Degree distributions of SCC, PCC, MI inferred networks.

pressing a known rod photoreceptor marker, the *Gnat1* gene.

With this strategy we found 7 regulons, all of them regulated by a TF previously reported in literature to have retina specific function, except for one gene, *ENSSSCG00000009560*, a predicted porcine gene whose human ortholog, *Tfdp1* wasn't previously characterised in retina and in photoreceptors. We applied a hypergeometric function [165] to assess the GO functions enriched within those 7 clusters, and we found enriched the categories "phototransduction", "photoreceptor outer segment" confirming that the predicted connections were specific.

Rhodopsin gene was found in the photoreceptorial regulon regulated by *Esrrb* gene, a known transcriptional regulator of energy metabolism that protects rod photoreceptors from dystrophy in retinal adult phases [166].

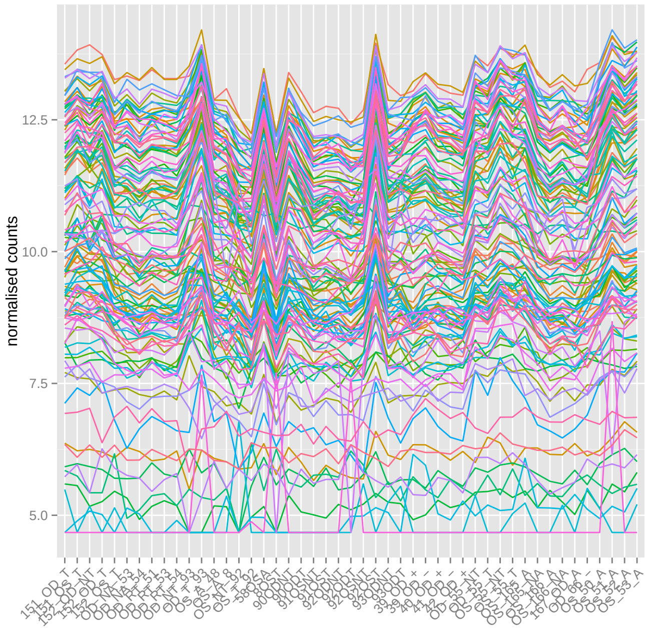


Figure 5.10: Ribosomal genes' expression levels show high concordance within the dataset.

We cross-compared Esrrb regulon with genes downregulated in Esrrb KO published in [167]. Interestingly, 23 genes were in common between our regulon (445) and those down-regluated in that work (425), with a significant pvalue (0.0007, fisher exact test). To asses whether the inferred network was sensitive to Rhodopsin connections, we extracted the genes connected to Rhodopsin and applied the hypergeomtric function to this set, again founding photoreceptorial processes among the most significant. Within this set we found, among others, the *Gnat1* gene, confirming previous data on the coexpression of this two markers.

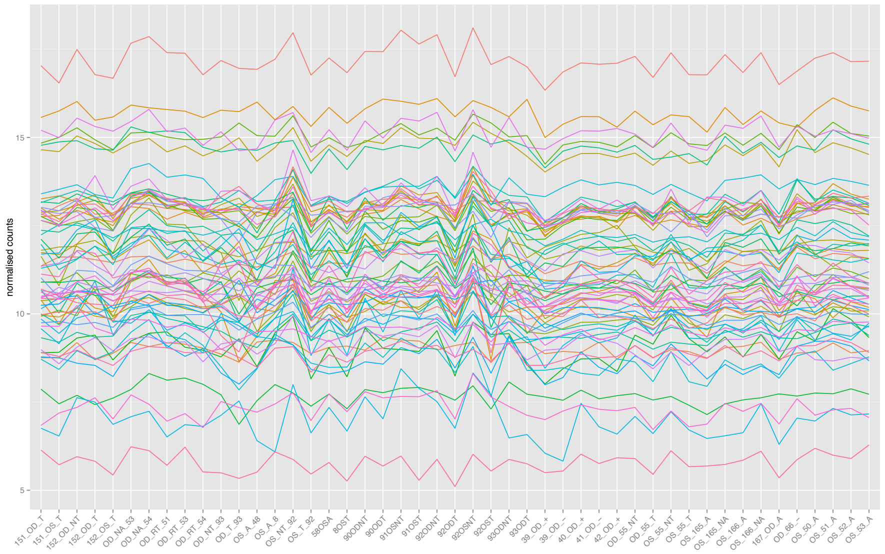
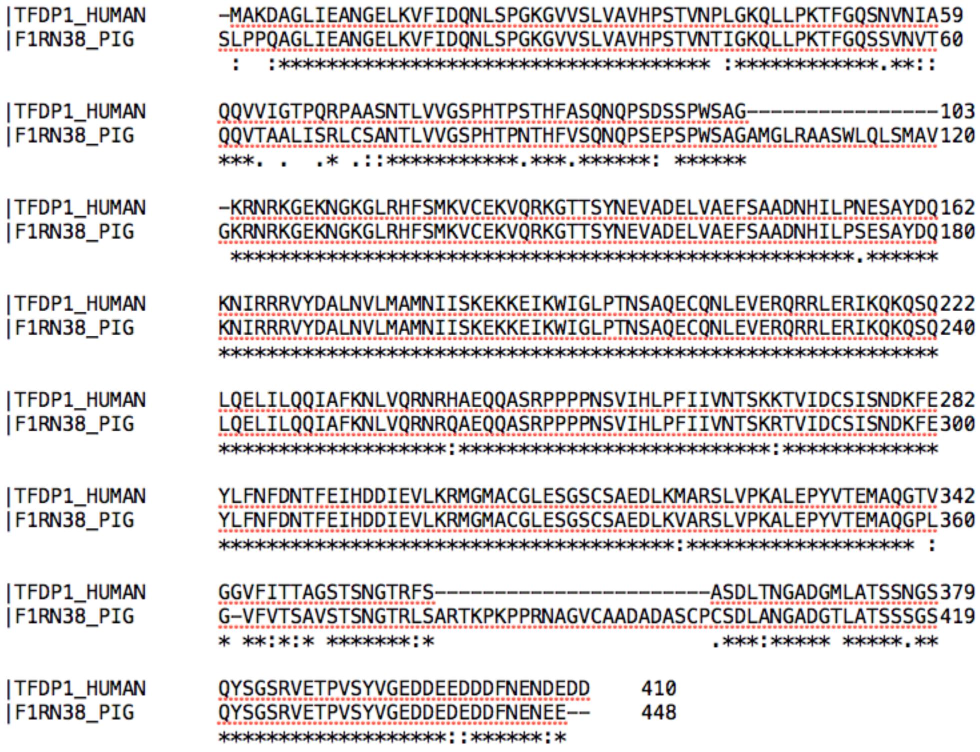


Figure 5.11: Rhodopsin regulon: expression levels of the 65 genes connected to Rhodopsin show high concordance within the dataset.

5.3.5 Precedently uncharacterised porcine ortholog of TFDP1 is expressed in Rods

ENSSSCG00000009560 is an uncharacterised protein coding gene on forward strand of Chromosome 11 (86495533 to 86510641) which in pig produces one transcript and a 448 aminoacid long protein. It belongs to the E2F/DP family, a family of Transcription Factors involved through the dimerization with E2F proteins in the cell cycle regulation and synthesis of DNA in mammalian cells. Porcine TFDP1 has 90% sequence homology with the human TFDP1, a TF reported to bind DNA cooperatively with E2F family members through the E2 recognition site, 5'-TTTC[CG]CGC-3', found in the promoter region of a number of genes whose products are involved in cell cycle regulation or in DNA replication. The E2F1:DP complex appears to mediate both cell proliferation and apoptosis. Its expression has been assessed in muscle, brain, placenta, liver and kidney. Lower levels were detected in lung and pancreas, but not in retina.



```
|TFDP1_HUMAN      -MAKDAGLIEANGELKVFIDQNLSPGKGVVSLVAVHPSTVNPLGKQLLPKTFGQSNVNIA59
|F1RN38_PIG       SLPPQAGLIEANGELKVFIDQNLSPGKGVVSLVAVHPSTVNTIGKQLLPKTFGQSSVNVT60
                   :  :************************************ :************.**::

|TFDP1_HUMAN      QQVVIGTPQRPAASNTLVVGSPHTPSTHFASQNQPSDSSPWSAG----------------103
|F1RN38_PIG       QQVTAALISRLCSANTLVVGSPHTPNTHFVSQNQPSEPSPWSAGAMGLRAASWLQLSMAV120
                  ***. .  .* .::***********.***.******: ******

|TFDP1_HUMAN      -KRNRKGEKNGKGLRHFSMKVCEKVQRKGTTSYNEVADELVAEFSAADNHILPNESAYDQ162
|F1RN38_PIG       GKRNRKGEKNGKGLRHFSMKVCEKVQRKGTTSYNEVADELVAEFSAADNHILPSESAYDQ180
                   ****************************************************.******

|TFDP1_HUMAN      KNIRRRVYDALNVLMAMNIISKEKKEIKWIGLPTNSAQECQNLEVERQRRLERIKQKQSQ222
|F1RN38_PIG       KNIRRRVYDALNVLMAMNIISKEKKEIKWIGLPTNSAQECQNLEVERQRRLERIKQKQSQ240
                  ************************************************************

|TFDP1_HUMAN      LQELILQQIAFKNLVQRNRHAEQQASRPPPPNSVIHLPFIIVNTSKKTVIDCSISNDKFE282
|F1RN38_PIG       LQELILQQIAFKNLVQRNRQAEQQASRPPPPNSVIHLPFIIVNTSKRTVIDCSISNDKFE300
                  *******************:**************************:*************

|TFDP1_HUMAN      YLFNFDNTFEIHDDIEVLKRMGMACGLESGSCSAEDLKMARSLVPKALEPYVTEMAQGTV342
|F1RN38_PIG       YLFNFDNTFEIHDDIEVLKRMGMACGLESGSCSAEDLKVARSLVPKALEPYVTEMAQGPL360
                  **************************************:******************* :

|TFDP1_HUMAN      GGVFITTAGSTSNGTRFS-----------------------ASDLTNGADGMLATSSNGS379
|F1RN38_PIG       G-VFVTSAVSTSNGTRLSARTKPKPPRNAGVCAADADASCPCSDLANGADGTLATSSSGS419
                  * **:*:* *******:*                       .***:***** *****.**

|TFDP1_HUMAN      QYSGSRVETPVSYVGEDDEEDDDFNENDEDD    410
|F1RN38_PIG       QYSGSRVETPVSYVGEDDEDEDDFNENEE--    448
                  *******************::******:*
```

Figure 5.12: Alignment of ENSSSCG00000009560 and *Tfdp1* shows high conservation between the sequences.

We assessed swine TFPD1 expression in retinal samples and in particular in sorted rods photoreceptors by RT-PCR. Among the most enriched functions in the regulon we found photoreceptor-related processes and methyltransferase activity, a process that has been already reported for its ortholog, in complexes with other histone regulating proteins. Interestingly and confirming its photoreceptorial expression, we found the term "Cajal body", which are spherical sub-organelles of 0.3-1.0 μm in diameter found in the nucleus of proliferative cells like embryonic cells and tumor cells, or metabolically active cells like neurons [168] .

Chapter 6

Retina Specific Processes

6.1 Network inference on 100 samples from 10 porcine tissues

To further test the strategy used, we downloaded the raw RNA-seq data of 10 porcine tissues from [169]. The data comprised 100 samples, with 10 tissues sequenced: heart, spleen, liver, kidney, lung, musculus longissimusdorsi, occipital cortex, hypothalamus, frontal cortex, and cerebellum. To obtain the count estimates from the .bam files downloaded by Array Express, we used a custom R script exploiting functions from different R/Bioconductor packages for raw sequences manipulation (*AnnotationDbi, Rsamtools, GenomicAlignments, BSgenome.Sscrofa.UCSC.susScr3, GenomicFeatures*).
We then removed genes with less than 1 CPM on average, and proceeded to the normalization of count data using DESeq2, with a GLM with 1 factor and 10 levels (one for each tissue), we finally applied the VST transformation. To estimate the correlations, we used again the *corr.test* function with BH correction. We then filtered the correlations retaining those greater or equal to 0.60 with 5% FDR.

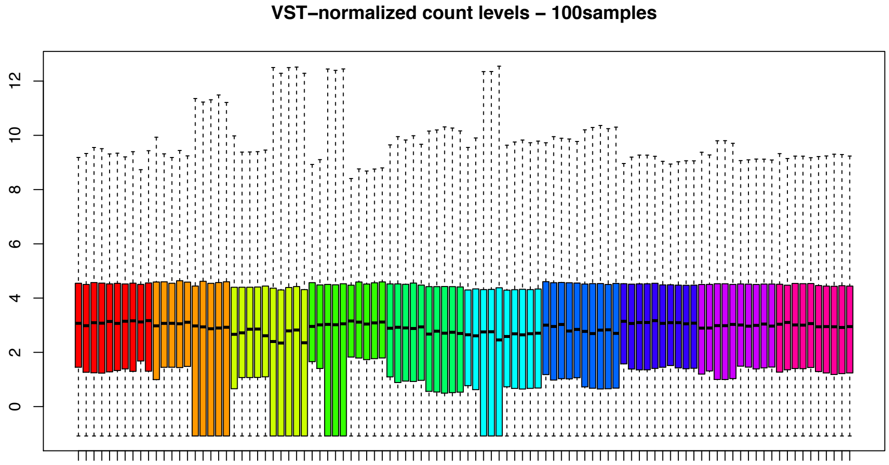


Figure 6.1: VST- Normalised counts for the 100-samples dataset.

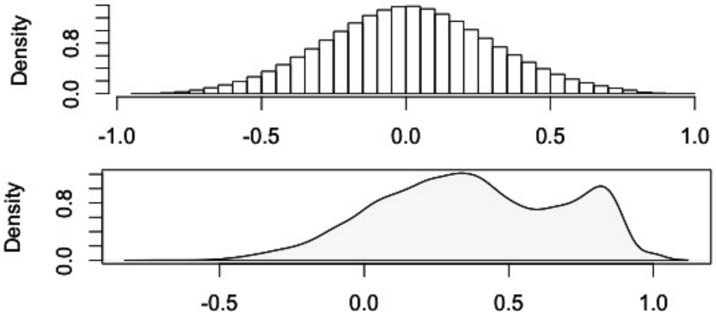


Figure 6.2: Density distribution of all the correlations (upper panel) and of the Ribosomal genes (lower panel) from the 100-tissues inferred network

6.2 Modularity

We used the same strategy used for SusNet using the genes which had "transcription factor activity" retrieved in SusNet, we obtained a total of 469 TFs. Coherently with our speculations, many of the Retina specific factors weren't expressed in this dataset. As expected, Rhodopsin expression was undetectable.

Four of the 7 “photoreceptor”-TF were present in this dataset, *Mef2d, Sp2, Fdz4 and Esrrb*. Interestingly, the TF which had *Rhodopsin* in its regulon, *Esrrb*, had only 2 connections in the 100-samples inferred network, confirming the specificity of the connections found in the retina dataset.

As expected no photoreceptorial processes were found by hypergeometric enrichment in those modules.

We selected a Transcription Factor, *Pax6*, known to be a regulator of brain development and involved in retinoblastoma via interaction with p53, p21, p27. In this dataset, Pax6 module is composed of 1108 genes. Coherently with reported data [170], Gene ontology analysis confirmed significant enrichment of neuronal processes.

6.3 Retina processes: Differential expression analysis

We constructed a new dataset adding to the 100 samples our 47 retinal samples, and normalized them following *DESeq2* pipeline. Count estimates were further processed to obtain the differential expression pattern in Retina vs Other tissues. We applied a GLM with 2 factors, one for the 11 tissues in the cumulative dataset an the second with 2 levels for “retina” and “other”. We were interested in retina-specific processes so we used a 2 level contrast (the second factor in the design matrix). Notably, upon the most highly differentially expressed genes we found *Rhodopsin* and genes belonging to mitochondria, including genes crucial in respiratory chain. It has been already demonstrated indeed that retina is an high energy-requiring tissue, hence the total RNA of this tissue is enriched in mitochondrial genes [171]. Upon the most down-regulated genes we found CYP genes, liver-specific, TNNC1, muscle-specific, HPD, whose expression has been reported in pituitary gland.

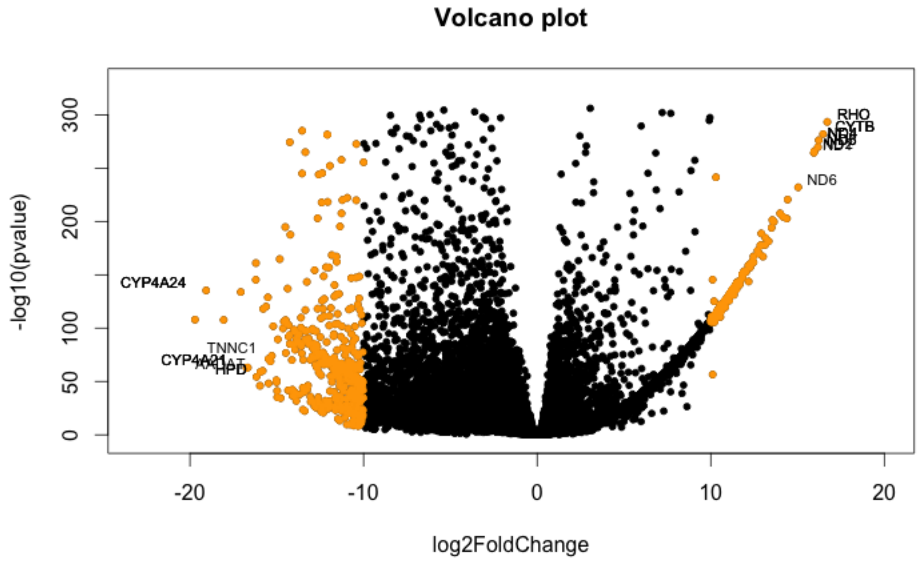


Figure 6.3: Top Differentially expressed genes in Retina compared to all the other tissues

Genes	GO terms	FDR
6 of 4	glutamate receptor activity	1e-05
4 of 4	protein localization to juxtaparanode region of axon	1e-05
3 of 3	α-1,6-mannosylglycoprotein 6-beta-N-acetylglucosaminyltransferase activity	1e-05
4 of 4	regulation of dendritic spine morphogenesis	1e-05
4 of 4	benzodiazepine receptor activity	1e-05
4 of 4	inhibitory extracellular ligand-gated ion channel activity	1e-05
332 of 2436	integral component of membrane	1e-05
412 of 3245	membrane	1e-05
30 of 76	axon	1e-05
31 of 82	synaptic transmission	1e-05
18 of 36	postsynaptic density	1e-05
41 of 154	cell adhesion	1e-05
140 of 915	plasma membrane	1e-05
9 of 11	regulation of synaptic transmission, glutamatergic	1e-05
12 of 19	synapse organization	1e-05
31 of 108	neuron projection	1e-05
34 of 127	ribosome	1e-05
22 of 64	nervous system development	1e-05
14 of 29	membrane depolarization during action potential	1e-05
35 of 139	structural constituent of ribosome	1e-05
25 of 83	neuronal cell body	1e-05
15 of 36	cytosolic small ribosomal subunit	1e-05
21 of 65	regulation of membrane potential	1e-05
10 of 18	small ribosomal subunit	1e-05
30 of 119	synapse	1e-05
8 of 12	GABA-A receptor complex	1e-05
8 of 12	GABA-A receptor activity	1e-05
9 of 15	AMPA glutamate receptor complex	1e-05

Table 6.1: Some of the significantly enriched processes Pax6 regulon

Category	GO terms	FDR
2/2	positive regulation of rhodopsin gene expression	0
8/19	respiratory chain	0
7/34	photoreceptor outer segment	0
6/24	NADH dehydrogenase (ubiquinone) activity	0
3/4	respiratory chain complex IV	0
5/19	photoreceptor inner segment	0
6/40	mitochondrial respiratory chain complex I	0
11/194	mitochondrial inner membrane	0
6/45	mitochondrial membrane	0
26/1109	mitochondrion	0
3/8	ATP synthesis coupled electron transport	0
6/59	visual perception	1e-05
2/3	negative regulation of cholesterol efflux	1e-05
2/3	photoreceptor cell development	1e-05
5/43	retina development in camera-type eye	2e-05
4/25	ATP synthesis coupled proton transport	2e-05
3/12	phototransduction	3e-05
3/12	positive regulation of mRNA splicing, via spliceosome	3e-05
2/4	opsin binding	3e-05
3/14	respiratory electron transport chain	5e-05
4/31	hydrogen ion transmembrane transporter activity	6e-05
2/6	leucine zipper domain binding	0.00014
2/6	group III metabotropic glutamate receptor activity	0.00014
5/64	proton transport	0.00017
2/7	protein-chromophore linkage	0.00021
2/7	photoreceptor activity	0.00021
2/7	sensory perception of light stimulus	0.00021
2/7	dopamine receptor signaling pathway	0.00021
2/7	proton-transporting ATP synthase complex, coupling factor F(o)	0.00021
2/7	poly(U) RNA binding	0.00021
3/22	tricarboxylic acid cycle	0.00031
3/22	cytochrome-c oxidase activity	0.00031
2/8	striatum development	0.00032
2/8	ATP biosynthetic process	0.00032

Table 6.2: List of significantly enriched processes in up-regulated genes in Retina vs other tissues

Chapter 7

Rhodopsin-targeted silencing by DNA-binding

Transcription factors operate by the combined activity of their DNA-binding domains (DBDs) and effector domains (EDs) enabling the coordination of gene expression on a genomic scale. TF operate by recruiting co-activator or co-repressor complexes [172, 173] resulting in either transcriptional activation or repression of specific genes. We investigated the hypothesis that engineered DNA-binding proteins without canonical ED activity possess transcriptional repression properties.

To interfere with Rhodopsin gain-of-function mutations we engineered a DNA-binding protein that targets 20 base pairs (bp) of a RHO cis-regulatory element and demonstrate Rho specific transcriptional silencing upon adeno-associated viral (AAV) vector-mediated expression in photoreceptors.

7.1 AAV generation and delivery

We generated a DNA-binding protein targeted to a cis-regulatory element (CRE) of the human proximal RHO promoter region by deconstructing a synthetic TF composed of a DBD (ZF6-DNA-binding protein, ZF6-DB) and the ED (Kruppel-associated box, KRAB repressor domain, KRAB), which we have shown to be effective in repressing specifically the human RHO transgene carried in an adRP mouse model [85]. The deletion of the ED resulted in a protein, ZF6-DB, targeting 20 bp of genomic CRE, here named ZF6-cis, found at -84 bp to -65 bp from the transcription start site (TSS) of the human RHO gene [174]. Genomic ZF6-cis is without apparent photoreceptor-specific endogenous transcription factor-binding sites. To evaluate whether ZF6-DB represses transcription of the RHO gene in a physiological genomic context, we used the porcine retina [85], which shares 19 out of 20 DNA bp with the human genomic ZF6-cis sequence. We generated a construct carrying the sequence of the ZF6 protein under an ubiquitous promoter, and subretinal delivery of a low AAV8 vector dose (1×10^{10} genome copies; gc) of ZF6-DB (AAV8-CMV-ZF6-DB) resulted in a 45% decrease of porcine Rho transcript levels at 15 days post-injection.

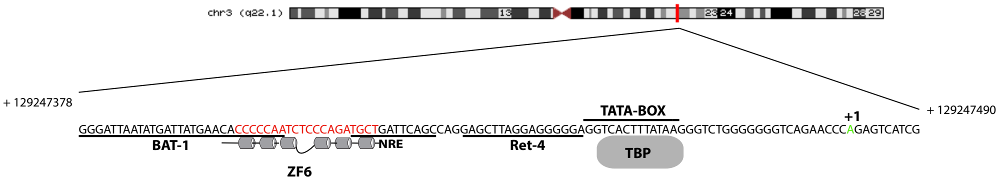

chr3 (q22.1)
13
23
24
28
29
+ 129247378
+ 129247490
TATA-BOX
+1
GGGATTAATATGATTATGAACACCCCCAATCTCCCAGATGCTGATTCAGCCAGGAGCTTAGGAGGGGGAGGTCACTTTATAAGGGTCTGGGGGGGTCAGAACCCAGAGTCATCG
BAT-1
NRE
Ret-4
TBP
ZF6

Figure 7.1: Schematic representation of the chromosomal location of the RHO locus and its proximal promoter elements indicating the transcription start site (in green, +1) and the location of ZF6-DB binding site (in red, ZF6-Cis) and ZF6-DB

7.2 Rhodopsin downregulation is Rod-specific

Rhodopsin is expressed only in rods. To evaluate whether ZF6-DB was effectively repressing its expression and the result was not due to the averaging effect of the pooled retinal cells which doesn't express *Rhodopsin* at all, we performed FACS analysis on eGFP-labelled rod cells.

Rod cells were isolated from porcine retina that had received a subretinal injection of an AAV vector encoding eGFP under the control of a rod-specific promoter (human Guanine Nucleotide Binding Protein1, GNAT1 promoter elements [175]; AAV8-GNAT1-eGFP; dose 1×10^{12} gc) with or without the vector encoding ZF6-DB (5×10^{10} gc). Fifteen days after injection, the retinae were disaggregated and FACS-sorted. Cells co-transduced with eGFP and ZF6-DB vector showed virtually a 'somatic knock-out' of Rho expression with a 85% decrease of Rho transcript levels.

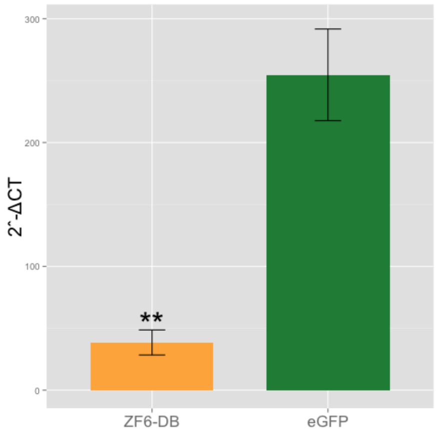


Figure 7.2: qReal Time PCR on sorted rods shows Rhodopsin downregulation in rods ZF6-DB interfered compared to controls.

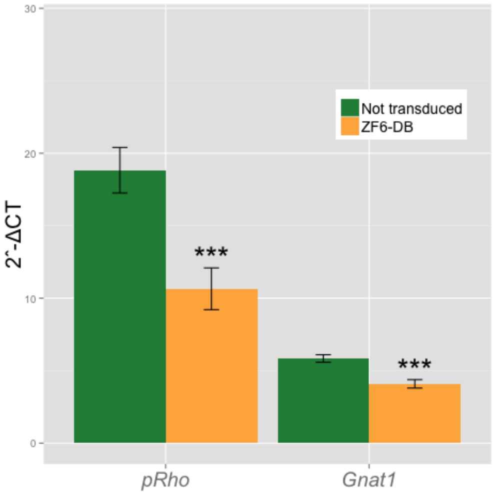


Figure 7.3: qReal Time PCR of mRNA levels $(2^{-}DCT)$ of adult porcine retina injected subretinally with AAV8-CMV-ZF6-DB at a vector dose of 1×10^{12} gc compared with non-transduced area of the same eye 15 days after vector delivery, resulted in robust transcriptional repression of the *Rhodopsin* transcript and downregulation of *Gnat1*.

7.3 RNA-seq

To evaluate genome-wide transcriptional specificity, we analyzed the porcine retinal transcriptome by RNA-Seq from retina harvested 15 days after subretinal injection of the AAV8 vector encoding ZF6-DB. For comparison we used the engineered TF with the ED, KRAB (AAV8-CMV-ZF6-KRAB).

The low vector doses delivered to the porcine retina (1×10^{10}) resulted in about twenty-fold lower expression levels of the ZF6-DB and ZF6-KRAB transgenes compared to *Crx* and *Nrl*, two retina-specific TFs [164]. Interestingly, despite the construct's low expression levels, we observed robust Rho transcriptional repression.

We analyzed the transcriptional perturbation in response to the AAV retinal gene transfer of ZF6-DB by determining the differentially expressed genes (DEGs). To this end we evaluated two algorithms performance, *edgeR* and *DESeq2*, obtaining overall concordance on the results. Interestingly, *DESeq2* showed a better control on the outlier genes, so we decided to proceed with that pipeline for further analyses. The dataset was composed of 17 samples and 25.325 genes, divided in 3 experimental groups: 7 Controls, 4 ZF6-KRAB-treated, 6 ZF6-DB-treated.

We analyzed the data following the standard Differential Expression Analysis Pipeline with DESeq2 R/Bioconductor package (v.1.8.1) [115], filtering and normalizing all libraries together. We filtered low tag counts retaining those which had 1 CPM in at least 3 samples.

We fitted a unique Generalized Linear Model (GLM) with 1 factor and 3 levels (Control, ZF6-KRAB-treated, ZF6-DB-treated). Differentially expressed genes were obtained out of the 2 contrasts (each treatment compared to the controls), an adjusted pvalue (FDR) of less than or equal to 0.1 was considered significant. We observed the expected upregulation of the exogenous genes used for the treatment (ZF6-KRAB, ZF6-DB, eGFP) and for further evaluations we didn't take into account their differential expression.

Remarkably, in vivo the ZF6-DB protein generated about ten-fold less transcriptional perturbation compared with the ZF6-KRAB protein (19 vs. 222 DEGs). Notably, this magnitude of perturbation is twenty five-fold lower than that induced by the ablation of an endogenous rod-specific TF (NRL, 500 DEGs vs 19 DEGs, ZF6-DBD;[176]). Retinal-specific pathway analysis of DEGs showed that ZF6-DB induced down-regulation is restricted to the Rho biochemical interactor *Gnat1* [177], and the up-regulation of 2 genes associated with acute phase inflammatory response, alpha-2-macroglobulin (*A2m*) and glial fibrillary acidic protein (*Gfap*). Interestingly the Gnat1 gene was among the genes connected to Rhodopsin.

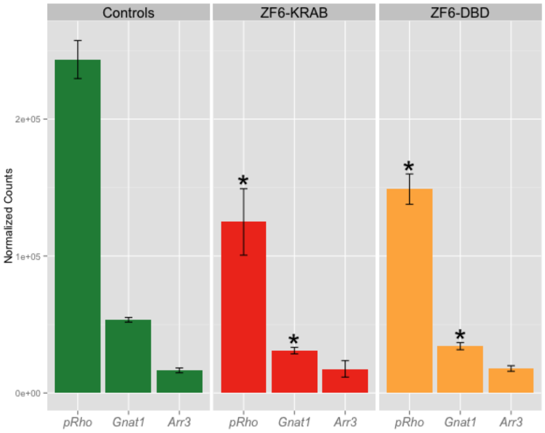


Figure 7.4: Rhodopsin gene expression in size-factor normalised counts in RNA-Seq data

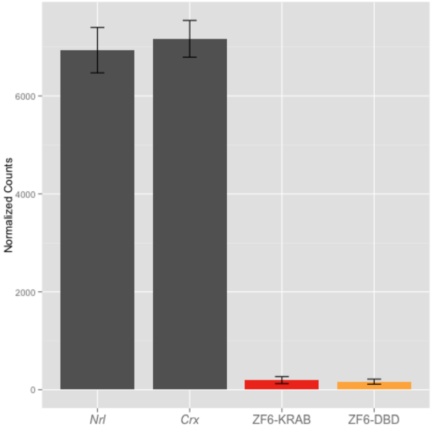


Figure 7.5: ZF6-DB expression compared to Nrl and Crx in size-factor normalised counts in RNA-Seq data

These results suggest that both ZF6-DB and the consequent Rho down-regulation marginally interferes with photoreceptor specific pathways, apart from Gnat1 repression, and that the up-regulation of the inflammatory response genes may be due to the collapse of the retinal scaffold caused by Rho depletion.

The intersection of DEGs between ZF6-DB and ZF6-KRAB showed that both drive similar perturbation in the expression of 16 genes, which represent 84% of the entire pool of ZF6-DB DEGs. Consistently, both ZF6-DB and ZF6-KRAB generated similar functional effects, i.e. concordant up- or down- differential expression of these 16 shared genes. These results suggest that both ZF6-DB and ZF6-KRAB bind to similar genomic targets.

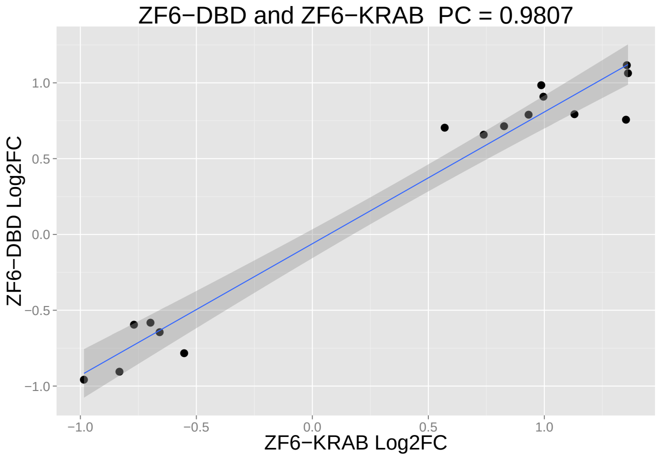


Figure 7.6: Genes in common between ZF6-KRAB and ZF6-DB show high functional concordance (PCC on log2 Fold Change levels)

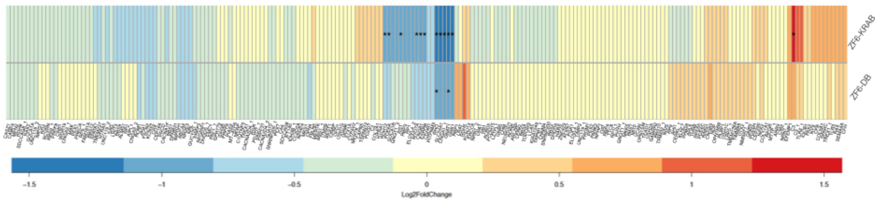


Figure 7.7: Heatmap of genes associated to retinal patologies. Genes tagged with an asterisk are significantly DE (FDR 10%).

A manually curated list of Human Gene IDs including representative Retinal Markes and a subset of Retina Disease Genes [45] was used to show the interference power of the 2 TRs with the overall retinal regulatory circuitry. The human IDs were used to retrieve their homolog Porcine genes, if present. We observed no significant up- or down-regulation of genes involved in retinal pathologies in the functional interference induced by ZF6-DB protein.

7.4 DEGs mapping on SusNet

To assess the interference of our TF on the retina Gene Regulatory Network, we mapped the 19 genes differentially expressed in ZF6-DB on SusNet. We intersected the 19 DEGs with the 784 regulons of the network and assessed any significant enrichment by hypergeometric function. We found a significant enrichment on modules regulated by the known retinal transcription factors Nrl and Crx. One of the most significant hits was the regulon of CEBP, a transcription factor implicated in various aspects of cell survival, apoptosis and inflammation. This gene is activated in the initial phases of retinal degeneration, its role is thought to be key in the preservation of photoreceptors inducing a delay of their phagocytosis from macrophages [178]. Interestingly, DE genes showed a significant intersection with the regulon of swine *Tfdp1* (*ENSSSCG00000009560*), confirming previous speculations.

Two genes differentially expressed out of 19 were found in Rhodopsin regulon itself, *Gnat1* and *Lrrc8e* (p value $\ll 1\times10^{-5}$), both downregulated by the ZF6-DB interference, further confirming the significance of the predicted connection.

Intersection	Regulon size	Gene Name	FDR
9	224	CEBPD	9,80E+02
9	107	BCL3	1,54E+00
7	76	TCF7	8,50E+01
5	1173	STAT3	0.0441
5	97	STK3	3,13E+07
4	100	EGR2	1,69E+09
4	876	NRL	0.0533
4	85	TEAD3	8,94E+07
4	225	ENSSSCG00000024816	0.0007
4	450	ELK3	0.010
4	272	IRF1	0.001
4	261	SOX9	0.001
3	326	ENSSSCG00000009560	0.0301
3	133	CLU	0.001
3	306	RIPK3	0.027
2	278	SHH	0.06
2	189	BCL6	0.0359
2	183	CRX	0.0346
2	22	JUN	0.0002
2	249	TP53	0.0522
2	147	TEAD4	0.0301
2	65	FZD4	0.004
1	47	LMO2	0.0356
1	12	ATF3	0.006
1	89	NDP	0.066

Table 7.1: TFs relevant for ZF6-DB interference, obtained from the intersection of DE genes with SuSnet regulons.

Chapter 8

Single cell RNA-seq of Rod Photoreceptors

We have carried out profiling of gene expression levels from bulk populations of millions of retinal input cells by RNA-seq.
However, in this way gene expression levels are an average of the input cells, and we could lose resolution on the specific dynamics occurring in photoreceptors. Any single functional rod with the proper balance of wild type Rhodopsin is essential to maintain retinal functionality, hence to profile rod cells and measure *Rhodopsin* expression avoiding any confounding averaging effect, we are exploiting single cell RNA-seq.

We want to study the transcriptional and molecular changes induced when interfering with Rhodopsin transcription in Mammalian Rod Photoreceptors; in the same time, we want to profile wild type Photoreceptors' Gene Expression, in order to asses potential cell-to-cell heterogeneity in this terminally differentiated neuronal cells assumed identical. We have designed an artificial transcription factor targeted to a sequence on Rhodopsin prossimal promoter and demonstrated that this ATF can bind and repress Rhodopsin

Transcription. Thus, we want to study Photoreceptors' transcriptome at varying amounts of ATF and, consequently, of its target gene, Rhodopsin in a dose-response context, at single cell resolution, to further clarify Rhodopsin balance in Rods cells and its role in Retinal pathologies.

8.1 Experimental Strategy

To sort single photoreceptors with FACS (Fluorescence Activated Cell Sorting), instead of using antibodies which couldn't ensure the cell specific fluorescence at the desired resolution, we decided to generate an AAV vector carrying the eGPF under the control of a rod-specific promoter, GNAT1 (AAV8-GNAT1-eGFP). To artificially interfere with Rhodopsin transcription, we designed another construct with the sequence of the ATF together with the fluorescent reporter eGFP, both under the control of a Rod specific promoter (AAV8-GNAT1-ATF-2A-eGFP).

8.2 FACS sorting and qPCR

Retinae were explanted after 15 days of treatment. The injection of the two constructs in pig retina (5×10^{10} AAV8-GNAT1-ATF-2A-eGFP and 1×10^{12} AAV8-GNAT1-eGFP) resulted in high level of transduction and high number of egfp-positive cells.

We disaggregated retinae and sorted fluorescent rods in bulk and in 96 well plates containing the lysis buffer as suggested by SMART-seq2 protocol [135]. We evaluated the expression of the ZF6-DB and the repression of Rhodopsin transcription in treated cells in bulk by RT-PCR. Then we profiled 288 single photoreceptors by qPCR: after retrotranscription, cDNA libraries coming from control and ZF6-DB-interfered single cells were used to evaluate the expression of Actin, Rhodopsin, ZF6-DB and ERCC.

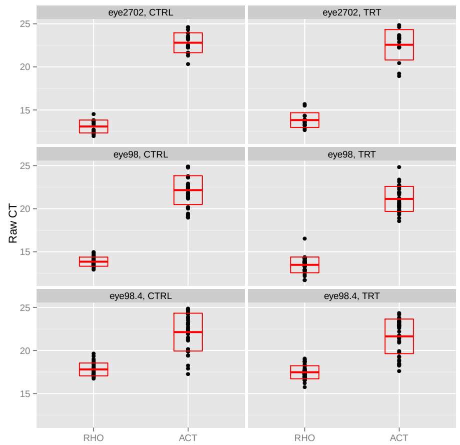


Figure 8.1: Raw CT levels of Rhodopsin and Actin in the cells analysed dy qPCR in three experiments evaluated. Red boxes show mean and standard deviation.

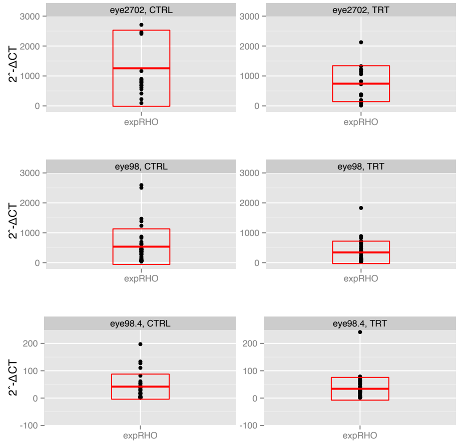


Figure 8.2: Rhodopsin normalized on Actin in the cells analysed dy qPCR in the three experiments evaluated. Red boxes show mean and standard deviation.

Despite the expected variability within the cells analyzed, which is well

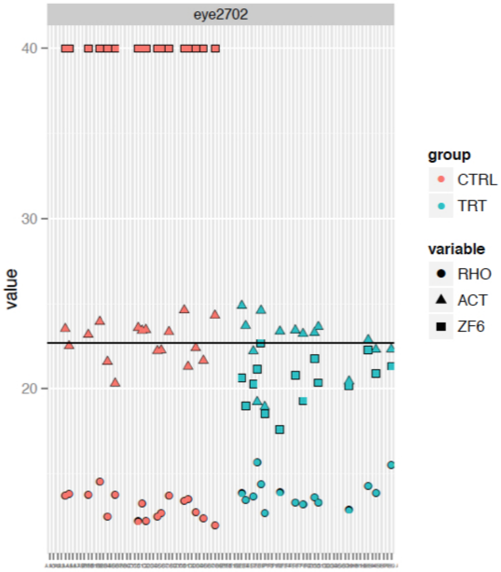


Figure 8.3: Raw CT levels of Rhodopsin, ZF6-DB and Actin in the cells analysed dy qPCR.

known in single cell quantitative biology, we could assess *Rhodopsin* downregulation and the expression of ZF6 specific only treated cells. Those cells represent the purest model to analyze Rhodopsin balance in wild type Rod cells, the specificity of the ZF6-DB for Rhodopsin, thus the cell-specific transcriptional mechanisms activated in response to Rhodopsin-interference, hence we are going to profile them by scRNA-seq.

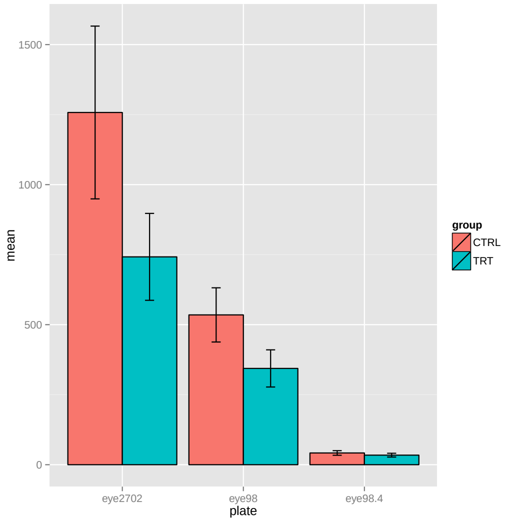


Figure 8.4: Downregulation of Rhodopsin in treated cells compared to controls in three experiments evaluated.

Chapter 9

Discussion

9.1 SusNet: a database for gene co-expression in Porcine Retina

Co-expression network represent a valid tool for predicting gene function, characterize co-expression patters of genes without common processes, and discovery of novel protein-protein interactions and functions. Notably, retina transcriptome is still poorly characterized by RNA-seq even in higher-order species, leaving a gap in inferential analysis of Gene Regulatory Networks specific to this tissue. To our knowledge, we produced the first retina-specific porcine RNA-Seq dataset and co-expression network. Despite the fact that in inferential network analysis better performance and predictions are achieved with high number of samples, SusNet performance was good, coherent with what reported in literature on RNA-seq co-expression network, and supported by experimental data.

It is important to remember that it is yet difficult to have a complete Gold Standard containing all of the real interactions for a transcription factor or gene-gene interaction of interest because of the partial knowledge inherent in biological data. More importantly still little is known about non-canonical animal models such as the Pig, due to lack of experimental data.

SusNet was able to detect links among known retinal TFs and their regulon, and known *Rhodopsin* interactions such as that with *Gnat1* and *Esrrb* genes. *Gnat1* encodes for the rod-specific transducin which stimulates the coupling of Rhodopsin and cGMP-phoshodiesterase during visual impulses, mutations in this gene result in vision impairment diseases. *Esrrb* is a Transcription Factor expressed downstream *Nrl* required for long-term survival of rods in adult mice [166]. Mutations in the human Esrrb gene are associated with autosomal-recessive deafness.

Interestingly, SusNet predicted the photoreceptorial expression of a precedently uncharacterized porcine protein, *ENSSSCG00000009560*, that we call swine *Tfdp1*. All that is known about swine *Tfdp1* protein is retrieved by its highly conserved human ortholog, whose expression pattern excludes retina. This protein is responsible of cell cycle regulation via dimerization with factors of the E2F family, which are negatively regulated by cell-cycle suppressor Rb1 (retinoblastoma protein).

We found that swine *Tfdp1* is consistently expressed in retina and in photoreceptors in particular, where it interacts with *Gnat1*, *Nrl* and *Crx*. Notably, we found that swine *Tfdp1* interacts with Wrap53, a gene producing a transcript stabilizing p53 mRNA and a second transcript, *Wrap53β*, which coordinates the formation and stabilization of Cajal bodies, a function that we found significantly enriched in *Tfdp1* regulon (see 6.2). This protein is involved in numerous cell functions such as telomere elongation, DNA double strand break repair and ribonucleoproteins biogenesis.

9.2 Retina specific processes

We tested our network inference strategy on a set of porcine RNA-Seq samples to asses the sensitivity of the method and the specificity of the links found.

We built a second co-expression network and evaluated the presence of conserved links within that and SusNet. None of the regulons retina-specific were found in this dataset, and no photoreceptorial processes were found enriched indeed.

We randomly selected a TF, *Pax6*, which is a key factor in the development of neural tissues, particularly the eye. Overexpression of *Pax6* in human retinoblastoma cells resulted in increased tumor cell proliferation in vitro paralleled by a downregulation of the p53, p21, and p27 proteins and an upregulation of the cdc2 protein [179]. In brain, *Pax6* expression coordinates the development of different areas already at embryonic stages, with stage and region-specific pattern [170]. Confirming this data, we found significantly enriched neuronal processes in Pax6 regulon.

Differential Expression Analysis between Retina and other tissues demonstrated a high enrichment of genes involved in energy production, mitochondrial genes, components of mitochondria and photoreceptor activities. Retina is an extremely specialized tissue whose energy requirement are massive, moreover it has been demonstrated a direct effect of mitochondria dysregulation induced by drugs can impair the glutamate production inducing selective death of photoreceptors, while previous studies found that depletion of glutamate correlates with cell death in the retina [180, 181, 182].

9.3 Rhodopsin-targeted silencing by DNA-binding

In this study we demonstrated that photoreceptor delivery of an AAV8 vector containing an artificial zinc finger-based transcriptional repressor without its effector domain (ZF6-DB) is capable of binding on a 20bp region on Rhodopsin promoter, pivotal to its expression. From a therapeutic prospective, a relevant property of the orthogonal ZF6-DB interference is the high rate of transcriptional silencing observed after in vivo gene transfer, which is consistent with canonical TFs mode of action. DNA binding interference via ZF6-DB in transduced retina generated 45% Rho transcriptional repression, which reached 85% when rods were sorted, supporting its use for diseases requiring correction of a large number of affected cells, such as adRP and other Mendelian disorders due to gain-of-function mutations.

The transcriptional repression mechanism of ZF6-DB binding likely relies on the interference occurring between TFs and local DNA sequence features within the RHO proximal promoter region, which we showed here to be necessary to control Rho expression at the genomic level. The absence of an effector domain which is the classical co-factors recruiter and the lack of known TFBSs in the 20-bp region of ZF6-DB binding suggest that the molecular determinant of silencing may not be due to other repressor recruitment or the simple displacement of key RHO TFs.

9.4 Single Cell RNA-seq of Rod Photoreceptors

Studies of genome and gene expression heterogeneity and plasticity aim at resolving the relationship between DNA and phenotype, which in some cases has pathological outcome. Identifying the factors that drive cell-to-cell heterogeneity and the variability in the response to therapy can help uncover the functional variations that drive specific biological outcome. For this, single cell studies are paramount. Our strategy to mark specifically rods via AAV expressing a fluorescent reporter under the control of the GNAT1 rod-specific promoter was indeed successful and allowed us to obtain an high number of cells to be profiled by scRNA-seq.

qPCR on those cells confirmed the presence of the ZF6-DB construct in the interfered rods and showed the reduction of *Rhodopsin* expression in the same cells.

scRNA-Seq of those cells will help us to unravel the cell-specific mechanisms that in rods govern physiological response to light, and to characterize this cells assumed identical, showing an unexpected complexity in their finely regulated processes.

More importantly, this study will let us characterize the quantitative aspects of *Rhodopsin* function and the mechanisms that are activated in *Rhodopsin*-depleted rods, providing new insights in physiology and pathological features of mammalian phototransduction.

9.5 Conclusions and Future Work

The retina is the region of the brain at the boundaries between the outside world and our perception. This finely organized network of co-regulated cells works like a processor for parallel computation of the stimuli that are conveyed to the brain to be elaborated as images. This extremely specialized tissue has evolved to optimize the way in which we interpret the visual information, with specific subset of cells deputed to light perception and others which integrate the signals from multiple sources.
People with vision-threatening retinopathy are likely to experience enhanced

social and emotional strain, with clear physical implications and worsening of life quality. Research on those diseases has provided crucial insights on the physiological mechanisms that are altered to contribute in generating targeted treatments. Rarest retina diseases, such those originating by Mendelian disorders, are currently in study for gene therapy approaches aimed at correcting the mutated genes that are responsible of the phenotype.

To better understand the Gene Regulatory Networks of the mammalian tetina we produced an exhaustive RNA-seq dataset from 47 porcine retinae and generated a co-expression network with a simple still powerful approach. Our strategy captured correctly known interactions and led us to understand strong relationships among known players of phototransduction processes. Moreover, we characterized a novel porcine protein whose involvement in retina and photoreceptors was still unknown.

Network inference from RNA-Seq data is an important methodological challenge. It appears that having more biological replicates instead of increased depth per sample is crucial to correctly infer relationships among genes. However, for tissue-targeted studies, such as ours on retina, it's necessary to construct dedicated datasets that can specifically capture tissue-specific gene behaviors, as their performance evaluation is commonly carried out relying on Gene Ontology as a reference. The creation of novel approaches to properly normalize count data and estimate gene-gene correlations from RNA-seq data will pose a future fundamental challenge for co-expression investigators.

We generated a therapy strategy, published on eLife [183], to treat an autosomal dominant form of Retinitis Pigmentosa involving a mutation in *Rhodopsin* gene. Our study support the use of a similar strategy for other Mendelian disorders due to gain-of-function mutations.

We demonstrated that an engineered non-canonical transcription factor composed of a DNA binding domain alone is able to repress *Rhodopsin* expression. Our data suggest the presence of a complex architecture of regulatory interactions between the DNA and the proteins that interpret and convey the genomic information. Understanding this code is key to correctly interpret the biological problem and design therapies using factors to regulate gene expression with a targeted strategy.

Moreover, single cell approaches are opening important discussions to quantitative biology and therapy, in this context, our scRNA-seq on ZF6-DB interfered rods will help to re-interpret the way in which we design therapeutical strategies aiming at rescuing selectively million of cells from the degeneration.

In conclusion, we believe that the data produced in our work will provide a useful resource for studies aimed at hijacking retina regulatory systems.

Appendix A

Materials and Methods

A.1 Plasmid construction

The ZF6-DNA-binding domain (Nδ96 deletion mutant, ZF6-DB) was amplified by PCR from AAV2.1 CMV-ZF6-KRAB [85] using primers ZF6-DBfw (TTGCGGCCGCATGATCGATC TGGAACCTGGCG) and ZF6-DBrv (AAGCTTTCAAGATGCATAGTCT). The PCR product was digested using NotI and HindIII restriction enzymes and cloned in pAAV2.1. The hGNAT1 promoter was synthetized by Eurofins MWG based on [175] adding the 5?UTR. The fragment was cloned in pAAV2.1 using NheI and NotI restriction enzymes. The human Rhodopsin CDS was ampli- fied by PCR from human retina cDNA using the hRHOfw (GCGGCCGCATGAATGGCACA- GAAGGCCC) e hRHOrv (AAGCTTTTAGGCCGGGGCCACCTG) primers and the PCR fragment was digested using NotI and HindIII restriction enzymes and cloned in pAAV2.1 plasmid under the con- trol of hGNAT1 promoter. For the generation of DBR-R plasmid the Eurofins MWG synthetized the expression cassette RHOD-ZF6-DB-bGHpolyA (bovine growth hormone polyA) that we cloned in pAAV2.1 hGNAT1-hRHO using NheI restriction enzyme.

A.2 AAV vector preparations

AAV vectors were produced by the TIGEM AAV Vector Core, by triple transfection of HEK293 cells followed by two rounds of CsCl2 purification (Auricchio et al., 2001). For each viral preparation, physical titers [genome copies (GC)/mL] were determined by averaging the titer achieved by dot- blot analysis [184] and by PCR quantification using TaqMan (Applied Biosystems, Carlsbad, CA, USA).

A.3 Vector administration and animal models

All procedures were performed in accordance with institutional guidelines for animal research and all of the animal studies were approved by the authors. P347S+/+ animals [85, 185] were bred in the animal facility of the Biotechnology Centre of the Cardarelli Hospital (Naples, Italy) with C57Bl/6 mice (Charles Rivers Laboratories, Calco, Italy), to obtain the P347S+/- mice.

Pigs

Eleven-week-old Large White (LW) female piglets were utilized. Pigs were fasted overnight leaving water ad libitum. The anesthetic and surgical procedures for pigs were previously described [186]. AAV vectors were inoculated sub-retinally in the avascular nasal area of the posterior pole between the two main vascular arches, as performed in Mussolino et al. [186]. This retinal region is crossed by a streak-like region that extends from the nasal to the temporal edge parallel to the horizontal meridian, where cone density is high, reaching 20,000 to 35,000 cone cells per mm2. Each viral vector was injected in a total volume of 100 ml, resulting in the formation of a subretinal bleb with a typical ‘dome-shaped’ retinal detachment, with a size corresponding to 5 optical discs.

A.4 Cloning and Purification of the proteins

DNA fragments encoding the sequence of the engineered transcription factors and ZF6-KRAB, to be expressed as maltose-binding protein (MBP) fusion were generated by PCR using the plasmids pAAV2.1 CMV-ZF6-KRAB and pAAV2.1 CMV-ZF6-DB as a DNA template.
The following oligonucleotides were used as primers: primer 1, (GGAATTC-CATATGGAATTCCCCATGGATGC) and primer 2, (CGGGATCCCTATC-TAGAAGTCTTTTTACCGGTATG), for ZF6-KRAB primer 3, (GGAATTC-CATA TGCTGGAACCTGGCGAAAAACCG) and primer 4,(CGGGATC-CCTATCTAGAAGTCTTTTTACCGG TATG) for ZF6-DB. All the PCR products were digested with the restriction enzymes NdeI and BamH1 and cloned into NdeI BamH1-digested pMal C5G (New England Biolabs, Ipswich, MA) bacterial expression vector. All the plasmids obtained were sequenced to confirm that there were no mutations in the coding sequences. The fusion proteins were expressed in the Escherichia coli BL21DE3 host strain. The transformed cells were grown in rich medium plus 0.2% glucose (according to protocol from New England Biolabs) at 37 °C until the absorbance at 600 nm was 0.6-0.8, at which time the medium was supplemented with 200 mM ZnSO4, and protein expression was induced with 0.3 mM isopropyl 1-thio-b-D-galactopyranoside and was allowed to proceed for 2 hr. The cells were then harvested, resuspended in 1X PBS (pH 7.4), 1 mM phenylmethylsulfonyl fluoride, 1 mM leupep- tin, 1 mM aprotinin, and 10 mg/ml lysozyme, sonicated, and centrifuged for 30 min at 27,500 relative centrifugal force. The supernatant was then loaded on amylose resin (New England Biolabs) accord- ing to the manufacturer?s protocol. To remove the MBP from the

proteins, bound fusion proteins as cleaved in situ on the amylose resin with Factor Xa (1 unit/20 mg of MBP fusion protein) in FXa buffer (20 mM Tris, pH 8.0, 100 mM NaC1, 2 mM CaC12) for 24?48 hr at 4 °C and collected in the same buffer after centrifugation at 500 relative centrifugal force for 5 min. The supernatant containing the protein without the MBP tag was then recovered.

A.5 qReal-time PCR

RNAs from tissues were isolated using RNAeasy Mini Kit (Qiagen, Germany), according to the manu- facturer protocol. cDNA was amplified from 1 mg isolated RNA using QuantiTect Reverse Transcrip- tion Kit (Qiagen), as indicated in the manufacturer instructions.
The PCRs with cDNA were carried out in a total volume of 20 ml, using 10 ml LightCycler 480 SYBR Green I Master Mix (Roche, Switzerland) and 400 nM primers under the following conditions: pre-Incubation, 50°C for 5 min, cycling: 45 cycles of 95°C for 10 s, 60°C for 20 s and 72°C for 20 s. Each sample was analysed in duplicate in two-independent experiments.
Transcript levels of pig retinae were measured by quantitative Real Time PCR using the LightCycler 480 (Roche) and the following primers: pRho-forward (ATCAACTTCCTCACGCTCTAC) and pRho-reverse (ATGAAGAGGTCAGCCACTGCC), pGnat1-forward (TGTGGAAGGACTCGGGTATC) and pGnat1-reverse (GTCTTGACACGTGAGCGTA), pArr3-forward (TGACAACTGCGAGAAACAGG) and pArr3-reverse (CACAGGACACCATCAGGTTG).
humanRho-forward (TCATGGTCCTAGGTGGCTTC), humanRho-reverse (ggaagttgctcatgggctta) and eGFP-forward (ACGTAAACGGCCACAAGTTC) and eGFP-reverse (AAGTCGTGCTGCTTCATGTG). All of the reactions were standardized against porcine Actb using the following primers: Act-

Forward (ACGGCATCGTCACCAACTG) and Act-reverse (CTGGGTCATCTTCTCACGG).

A.6 FACS rods sorting

Co-injected porcine retina with AAV8-CMV-ZF6-DB (dose 5×10^{10} gc)and AAV8-GNAT1-eGFP (dose 1×10^{12} gc) were disaggregated using Papain Dissociation System (Worthington biochemical corpora- tion) following the manufacturers protocol. Dissociated retinal cells were analysed using BD FACS Aria at IGB (Institute of Genetic and Biophysic"A. Buzzati-Traverso") FACS Facility and sorted, dividing eGFP positive cells (rods) from eGFP negative fraction.

A.7 RNA-Seq library preparation, sequencing and alignment

The 17 libraries were prepared using the TruSeq RNA v2 Kit (Illumina, San Diego, CA) according to manufacturer's protocol. Libraries were sequenced on the Illumina HiSeq 1000 platform and in 100- nt paired-end format to obtain approximately 30 million read pairs per sample. Sequence Reads were trimmed using Trim Galore! software (v.0.3.3) [187], that trims low-quality ends and removes adapter from reads, using a Default Phred score of 20. The 17 libraries were aligned on the full transcriptome for Sus scrofa (Pig) as provided by ENSEMBL (SusScrofa 10.2.73). The GTF included the sequences for the 20 canonical chromosomes plus 4563 scaffolds, and counted 30.567 transcripts plus the sequences for the 3 exogenes used in the analysis (the 2 TRs and eGFP). Alignment was performed with RSEM (v.1.2.11) with default parameters. The resulting expected counts (the sum of the posterior probability of each read coming from a specific transcript over all reads) were used for subsequent analysis.

Appendix B

Codes

```
packenv<-c("edgeR","DESeq2","NOISeq","gplots","ggplot2","sva","ROCR","ap.cluster")
lapply(packenv,require, character.only=T)

#------
#load data
#------

load("~/Documents/suScrofaNetwork/CurionSuracePigNetwork/corPigs.Rdata")
colnames(PigCounts)
legenda_all<-read.table('~/Documents/RNAseq_curion2015/
RNAseq_all_projects.txt',header=T,sep="\t")

#------
#create the design matrix for differential expression analysis
#------

colData<-legenda_all[match(colnames(PigCounts),legenda_all
$sample),c("name","eye","original.treatment")]
rownames(colData)<-colnames(PigCounts)
countData<-round(countData[rowSums(cpm(countData)>1)>=1,])
dim(countData)
sample=colData$sample
eye=colData$eye
ddsPig<-DESeqDataSetFromMatrix(countData=countData,
                              colData=colData,
                              design= ~sample)
ddsPig<-DESeq(ddsPig)
dim(corPigs)

#------
#Vst transform count data normalised by sizefactor and explore the results
#------

vst<-varianceStabilizingTransformation(ddsPig)
distVST<-dist(t(assay(vst)))
mat<-as.matrix(distVST)
rownames(mat)<-colnames(mat)<-with(colData(ddsPig),paste(rownames(colData), name, sep=" : "))
hc<-hclust(distVST)

#------
#heatmap of the original dataset
#------

heatmap.2(mat,Rowv=as.dendrogram(hc),
        trace='none',lwid = c(0.6,3),lhei=c(0.7,3),margins=c(8,8),
        cexRow=0.5,cexCol=0.5)

colScale <- scale_colour_manual(name = "eye",values = rmcol)

#------
#PCA of the original dataset
#------

plotPCA(vst,intgroup=c("name", "eye"))
data <- plotPCA(vst, intgroup=c("eye", "name","original.treatment"), returnData=TRUE)
percentVar <- round(100 * attr(data, "percentVar"))
pdf("PCA56vstNorm.pdf",15,10)
ggplot(data, aes(PC1, PC2, color=original.treatment)) +geom_point(size=2) +
  geom_text(aes(label=eye),hjust=1,vjust=0) +
  xlab(paste0("PC1: ",percentVar[1],"% variance")) +
  ylab(paste0("PC2: ",percentVar[2],"% variance")) +
  guides(col= guide_legend(nrow=20)) +
  scale_colour_brewer(palette="Set1")

#------
#Now remove samples that don't correlate with the others
#------

remove<-as.character(legenda_all[legenda_all$name%in
%c("pig_retina_11","CTRL_6","T*_ROD_1","C_ROD_1","T_ROD_2"),'sample'])

rPigCounts<-PigCounts[,-which(colnames(PigCounts)%in%remove)]

colData<-legenda_all[match(colnames(rPigCounts),legenda_all
$sample),c("name","eye","original.treatment")]
rownames(colData)<-colnames(rPigCounts)
countData<-rPigCounts
countData<-round(countData[rowSums(cpm(countData)>1)>=1,])

sample=colData$sample
eye=colData$eye
ddsPig<-DESeqDataSetFromMatrix(countData=countData,
                              colData=colData,
                              design= ~sample)
#------
#final dataset heatmap and PCA
#------

vst<-varianceStabilizingTransformation(ddsPig)
distVST<-dist(t(assay(vst)))
mat<-as.matrix(distVST)
rownames(mat)<-colnames(mat)<-with(colData(ddsPig),paste(rownames(colData), name, sep=" : "))
hc<-hclust(distVST)
heatmap.2(mat,Rowv=as.dendrogram(hc),
        trace='none',lwid = c(0.6,3),lhei=c(0.7,3),margins=c(8,8),
        cexRow=0.5,cexCol=0.5)

data <- plotPCA(vst, intgroup=c("eye", "name","original.treatment"), returnData=TRUE)
percentVar <- round(100 * attr(data, "percentVar"))


ggplot(data, aes(PC1, PC2, color=original.treatment)) +geom_point(size=2) +
  geom_text(aes(label=eye),hjust=1,vjust=0) +
  xlab(paste0("PC1: ",percentVar[1],"% variance")) +
  ylab(paste0("PC2: ",percentVar[2],"% variance")) +
  guides(col= guide_legend(nrow=20)) +
  scale_colour_brewer(palette="Set1")



#------
#package sva, to test whether any other source of variation is affecting data
#------

dat <- counts(ddsPig, normalized=TRUE)
idx <- rowMeans(dat) > 1
dat <- dat[idx,]
mod <- model.matrix(~ eye, colData(ddsPig))

mod0 <- model.matrix(~ 1, colData(ddsPig))
```

```
svseq <- svaseq(dat, mod, mod0, n.sv=2)
svseq$sv
par(mfrow=c(2,1),mar=c(3,5,3,1))
stripchart(svseq$sv[,1] ~ ddsPig$eye,vertical=TRUE,main="SV1")
abline(h=0)
stripchart(svseq$sv[,2] ~ ddsPig$eye,vertical=TRUE,main="SV2")
abline(h=0)
#####
ddssva <- ddsPig
ddssva$SV1 <- svseq$sv[,1]
ddssva$SV2 <- svseq$sv[,2]
design(ddssva) <- ~ SV1 + SV2 + eye
ddssva <- DESeq(ddssva)

vst_sva<-varianceStabilizingTransformation(ddssva)
data <- plotPCA(vst_sva, intgroup=c("SV1","SV2","original.treatment"), returnData=TRUE)
percentVar <- round(100 * attr(data, "percentVar"))

ggplot(data, aes(PC1, PC2, color=original.treatment)) +geom_point(size=2) +
  geom_text(aes(label=eye),hjust=1,vjust=0) +
  xlab(paste0("PC1: ",percentVar[1],"% variance")) +
  ylab(paste0("PC2: ",percentVar[2],"% variance")) +
  guides(col= guide_legend(nrow=20)) +
  scale_colour_brewer(palette="Set1")

#------
#the variance is well explained by the single level factor that we already use in the design
#------

boxplot(cbind(as.vector(cor(PigCounts[,colnames(RetinaVST)])),
             as.vector(cor(RetinaVST)),
             as.vector((cor(counts(ddsPig[,colnames(RetinaVST)], normalized=T)))))
       ,outline=F,names=c("Raw counts","VST","Size-Factor"), col="grey", ylab= "Correlation")

#----
#Mutual Information Estimation with parmigene R package
#----

RetinaVST<-vst
library(parmigene)
MI.data<-knnmi.all(RetinaVST)
mat<-MI.data
mat[lower.tri(mat, diag = TRUE)] <- NA
MI_retina<-na.omit(data.frame(as.table(mat)))

#-----
#Gold standard: STRING
#----

stringDB<-read.table("pig9823.protein.links.v10.txt",header=T, sep=" ")
colnames(stringDB)<-c('protein1','protein2','score')
proteinlevels<-strsplit(as.vector(stringDB$protein1),split="[.]")
proteinlevels<-sapply(proteinlevels, "[[", 2)
head(proteinlevels)
stringDB$protein1<-proteinlevels
proteinlevels<-strsplit(as.vector(stringDB$protein2),split="[.]")
proteinlevels<-sapply(proteinlevels, "[[", 2)
stringDB$protein2<-proteinlevels
#STRING threshold: select edges with .9 combined score
```

```
stringDB.900<-as.numeric(stringDB$score>=900)
stringDB.900<-data.frame(stringDB$protein1,stringDB$protein2,stringDB.900)
colnames(stringDB.900)<-c('node1','node2','value')

ensGene<-union(as.character(stringDB.900$node1), as.character(stringDB.900$node2))
ensembl<-useMart("ENSEMBL_MART_ENSEMBL",dataset="sscrofa_gene_ensembl",
host="www.ensembl.org")
attributes = listAttributes(ensembl)
filters = listFilters(ensembl)
filters="ensembl_peptide_id"
mart=ensembl
attributes=c("ensembl_gene_id","ensembl_peptide_id")#,#"hsapiens_homolog_ensembl_gene",
"hsapiens_homolog_ensembl_peptide",
        #"hsapiens_homolog_orthology_type")
s<-getBM(attributes=attributes,filters=filters,values=ensGene,mart=mart)
stringGeneId<-stringDB.900
stringGeneId[['node1']] <- s[ match(stringGeneId[['node1']], s[['ensembl_peptide_id']] ) ,
'ensembl_gene_id']
stringGeneId[['node2']] <- s[ match(stringGeneId[['node2']], s[['ensembl_peptide_id']] ) ,
'ensembl_gene_id']
if(sum(stringGeneId$node1=="")>0 | sum(stringGeneId$node2=="")>0) {
  stringGeneId<-stringGeneId[-which(stringGeneId$node1==""),]
  stringGeneId<-stringGeneId[-which(stringGeneId$node2==""),]

}else{
  stringGeneId<-na.omit(stringGeneId)
}

ids=length(union(stringGeneId$node1,stringGeneId$node2))
StringBM<-matrix(0,nrow=ids,ncol=ids)
rownames(StringBM)<-colnames(StringBM)<-union(stringGeneId$node1,stringGeneId$node2)
StringBM[as.matrix(stringGeneId[,1:2])]<-stringGeneId$value
length(rownames(StringBM))
dim(StringBM)

auc<-c()
thr<-seq(0.3,1.47, 0.01)

for(n in 1:length(thr)){
print(n)
GeneId<-MI_retina[MI_retina$Freq>=thr[n],]
colnames(GeneId)<-c("node1","node2","value")

ids=length(union(GeneId$node1,GeneId$node2))
CorrBM<-matrix(0,nrow=ids,ncol=ids)
rownames(CorrBM)<-colnames(CorrBM)<-union(GeneId$node1,GeneId$node2)
length(rownames(CorrBM))
CorrBM[as.matrix(GeneId[,1:2])]<-GeneId$value

int<-intersect(rownames(CorrBM),rownames(StringBM))
length(int)
corr<-CorrBM[int,int]
corr<-corr[upper.tri(corr)]
actual<-StringBM[int,int]
actual<-actual[upper.tri(actual)]
pred<-prediction(corr,actual)
auc<-c(auc,performance(pred,"auc")@y.values[[1]]

}
```

```
df<-cbind(thr,auc)

#-----
#estimate the Spearman\Pearson correlations
#-----

corSpearman<-corr.test(t(RetinaVST),method="spearman",adjust="BH",alpha=.05,ci=FALSE)
#corPearson<-corr.test(t(RetinaVST),method="pearson",adjust="BH",alpha=.05,ci=FALSE)
#the process is the same for Pearson correlation, I will discuss only the Spearman data

mat<-corSpearman$r
mat[lower.tri(mat, diag = TRUE)] <- NA
SCC.corr<-na.omit(data.frame(as.table(mat)))
mat<-corSpearman$p #correlation values
mat[lower.tri(mat, diag = TRUE)] <- NA
SCC.FDR<-na.omit(data.frame(as.table(mat)))
SCC.corr<-SCC.corr[SCC.FDR$value<=0.05,]

#test on STRING gold standard
stringDB<-read.table("pig9823.protein.links.v10.txt",header=T, sep=" ")
colnames(stringDB)<-c('protein1','protein2','score')
proteinlevels<-strsplit(as.vector(stringDB$protein1),split="[.]")
proteinlevels<-sapply(proteinlevels, "[[", 2)
head(proteinlevels)
stringDB$protein1<-proteinlevels
proteinlevels<-strsplit(as.vector(stringDB$protein2),split="[.]")
proteinlevels<-sapply(proteinlevels, "[[", 2)
stringDB$protein2<-proteinlevels

#STRING threshold: select edges with .9 combined score
stringDB.900<-as.numeric(stringDB$score>=900)
stringDB.900<-data.frame(stringDB$protein1,stringDB$protein2,stringDB.900)
colnames(stringDB.900)<-c('node1','node2','value')

ensGene<-union(as.character(stringDB.900$node1), as.character(stringDB.900$node2))
length(ensGene)
library(biomaRt)
ensembl<-useMart("ENSEMBL_MART_ENSEMBL",dataset="sscrofa_gene_ensembl",
host="www.ensembl.org")
attributes = listAttributes(ensembl)
filters = listFilters(ensembl)
filters="ensembl_peptide_id"
mart=ensembl
attributes=c("ensembl_gene_id","ensembl_peptide_id")#,#"hsapiens_homolog_ensembl_gene",
"hsapiens_homolog_ensembl_peptide",
                                  #"hsapiens_homolog_orthology_type")
s<-getBM(attributes=attributes,filters=filters,values=ensGene,mart=mart)

stringGeneId<-stringDB.900
stringGeneId[['node1']] <- s[ match(stringGeneId[['node1']], s[['ensembl_peptide_id']] ) ,
'ensembl_gene_id']
stringGeneId[['node2']] <- s[ match(stringGeneId[['node2']], s[['ensembl_peptide_id']] ) ,
'ensembl_gene_id']
if(sum(stringGeneId$node1=="")>0 I sum(stringGeneId$node2=="")>0) {
        stringGeneId<-stringGeneId[-which(stringGeneId$node1==""),]
        stringGeneId<-stringGeneId[-which(stringGeneId$node2==""),]

}else{
        stringGeneId<-na.omit(stringGeneId)
}
```

```
ids=length(union(stringGeneId$node1,stringGeneId$node2))
StringBM<-matrix(0,nrow=ids,ncol=ids)
rownames(StringBM)<-colnames(StringBM)<-union(stringGeneId$node1,stringGeneId$node2)
StringBM[as.matrix(stringGeneId[,1:2])]<-stringGeneId$value
length(rownames(StringBM))
dim(StringBM)

thr<-seq(0.5,0.8,0.1)
auc<-c()
for(n in 1:length(thr)){
print(n)

GeneId<-SCC.corr[SCC.corr$value>=thr[n],]
ids=length(union(GeneId$node1,GeneId$node2))
CorrBM<-matrix(0,nrow=ids,ncol=ids)
rownames(CorrBM)<-colnames(CorrBM)<-union(GeneId$node1,GeneId$node2)
length(rownames(CorrBM))
CorrBM[as.matrix(GeneId[,1:2])]<-GeneId$value

int<-intersect(rownames(CorrBM),rownames(StringBM))
length(int)
corr<-CorrBM[int,int]
corr<-corr[upper.tri(corr)]
actual<-StringBM[int,int]
actual<-actual[upper.tri(actual)]
auc<-c(auc,performance(pred,"auc")@y.values[[1]]
}




#------
#select the threshold
#------

scc<-SCC.corr[SCC.corr$value>=.6,]
g<-graph.data.frame(scc,directed=F)
E(g)$weight<-scc$value
ensGene=V(g)$name

#------
#communities detection:
#------

scc.comm<-fastgreedy.community(g)
V(g)$size =1
E(g)$count=1

comm.graph<-contract.vertices(g,scc.comm$membership,vertex.attr.comb=list(size="sum","ignore"))
comm.graph<-simplify(comm.graph,remove.loops=TRUE,edge.attr.com=list(count="sum","ignore"))


plot.igraph(comm.graph,
layout=layout.fruchterman.reingold,
vertex.size=x.vec+5,
vertex.label.cex=.5,
edge.arrow.size=.5,
edge.color="black",
edge.width=1.5)
dev.off()

dg<-decompose.graph(g,method="weak")
sping.comm<-spinglass.community(dg[[1]])

#------
#ap.cluster
#------

mat<-corPigSpearman$r
ap.results<-apcluster(mat,includeSim=T,details=TRUE)
datalist<-list()
for(n in 1:length(ap.results)){
  datalist[n]<-ap.results[n]
}
datalist<-lapply(datalist,names)

#Second level APCluster
dx<-mat
APRes_1<-ap.results
  dx_L2<-
matrix(data=max(dx),nrow=length(APRes_1@exemplars),ncol=length(APRes_1@exemplars))
  diag(dx_L2)<-0
  for(kk in 1:(length(APRes_1@exemplars)-1)){
    for(ll in (kk+1):length(APRes_1@exemplars))
      dx_L2[kk,ll]=dx_L2[ll,kk]=mean(dx[APRes_1@clusters[[kk]],APRes_1@clusters[[ll]]])
  }
  APRes_2<-apcluster(-as.matrix(dx_L2),maxits=5000,seed=1)#Cluster items based on negative
distances


  APRes_3<-apcluster(as.matrix(dx_L2),maxits=5000,seed=1)#Cluster items based on positive
distances

sum(unlist(lapply(APRes_1@clusters[APRes_2@clusters[[1]]], length)))
#sapply(gonames,"[[")

#------
#discard the clustering methods, use Transcription Factors- Master Regulator Analysis
#instead
#find the transcription factors
#------

ensembl<-useMart("ENSEMBL_MART_ENSEMBL",dataset="sscrofa_gene_ensembl",
host="www.ensembl.org")
attributes = listAttributes(ensembl)
filters = listFilters(ensembl)
filters="ensembl_gene_id"
mart=ensembl
attributes=c("ensembl_gene_id","external_gene_name","name_1006")
ensGene=V(g)$name
s<-getBM(attributes=attributes,filters=filters,values=ensGene,mart=mart)


tfCheck<-unique(s[grep("transcription factor activity",s$name_1006),"ensembl_gene_id"])
ltarget<-list()
for(n in 1:length(tfCheck)){
ltarget[[n]]<-attributes(neighbors(g,v=tfCheck[n]))$names
print(n)
```

```
}
names(ltarget)<-tfCheck

#------
#find the "photoreceptorial regulons" by searching gnat1 : "ENSSSCG00000024609"
#------

test<-which(!is.na(sapply(lapply(ltarget,function(x) grep("ENSSSCG00000024609" ,x)),"[",1)))
ensGene<-tfCheck[test]
attributes=c("ensembl_gene_id","external_gene_name")
rho_regulators<-getBM(attributes=attributes,filters=filters,values=ensGene,mart=mart) #(they are
indeed!)

#------
#rhodopsin regulon
#------

Reg.RHO<-attributes(neighbors(g,v="ENSSSCG00000011590"))$names
attributes=c("ensembl_gene_id","name_1006")
regulon.GO<-list()
for(n in 1:length(test)){
  print(n)
ensGene<-ltarget[[test[n]]]
regulon.GO[[n]]<-getBM(attributes=attributes,filters=filters,values=ensGene,mart=mart)
}
#prepare data for hypergeometric test
#from "phyper" - Hypergeometric {stats}
#phyper(x, m, n, k, lower.tail = FALSE, log.p = FALSE)

#x: vector of quantiles representing the number of white balls drawn without replacement
#from an urn which contains both black and white balls. --GENES IN THE REGULON BELONGING
TO "GO:TERM A"
#m: the number of white balls in the urn. --no of GENES IN THE REGULON
#n: the number of black balls in the urn. --no of GENES IN UNIVERSE minus GENES IN THE
REGULON
#k: the number of balls drawn from the urn. --no of GENES IN THE UNIVERSE BELONGING TO
"GO:TERM A"

attributes=c("ensembl_gene_id","external_gene_name","name_1006")
ensGene=V(g)$name
universe<-getBM(attributes=attributes,filters=filters,values=ensGene,mart=mart)

#------
#functions of the "rhodopsin regulon"
#------

goRHO<-getBM(attributes=attributes,filters=filters,values=Reg.RHO,mart=mart)
terms<-unique(goRHO$name_1006)
  log10p<-NULL
  catlength<-NULL
  Go_Universe<-c()
  for (i in 1:length(terms)){
    print(i)
    x=length(unique(goRHO[goRHO$name_1006==terms[i],"ensembl_gene_id"]))
    m=length(unique(goRHO$ensembl_gene_id))
    n=length(unique(universe_all$ensembl_gene_id))-length(unique(goRHO$ensembl_gene_id))
    k=length(unique(universe_all[universe_all$name_1006==terms[[i]],"ensembl_gene_id"]))
      log10p<-c(log10p,phyper(x,m,n,k,lower.tail=F))
      catlength<-c(catlength,x)
      Go_Universe<-c(Go_Universe,k)
  }
  rho<-data.frame(catlength,terms,Go_Universe,log10p,BH_corrected= p.adjust(log10p, "BH"))
}

#------
#functions of the photoreceptorial regulons
#------

colnames(universe)<-c('gene','go_term')
Enrichment<-list()
for(t in 1:length(regulon.GO)){
  print(paste("regulon.GO: ",t))
  terms<-unique(regulon.GO[[t]]$name_1006)
  log10p<-NULL
  catlength<-NULL
  Go_Universe<-c()
  for (i in 1:length(terms)){

    x=length(unique(regulon.GO[[t]][regulon.GO[[t]]$name_1006==terms[[i]],"ensembl_gene_id"]))
    m=length(unique(regulon.GO[[t]]$ensembl_gene_id))
    n=length(unique(universe$gene))-length(unique(regulon.GO[[t]]$ensembl_gene_id))
    k=length(unique(universe[universe$go_term==terms[[i]],"gene"]))
      log10p<-c(log10p,phyper(x,m,n,k,lower.tail=F))
      catlength<-c(catlength,x)
      Go_Universe<-c(Go_Universe,k)
  }
  Enrichment[[t]]<-data.frame(catlength,terms,Go_Universe,log10p,BH_corrected= p.adjust(log10p,
"BH"))

#------
#Load data from Botta et al. 2016
#Map Differentially expressed genes on Regulons:
#------

load("results_botta_etall.Rdata")
log10p<-c()
catlength<-c()
Go_Universe<-c()
for (i in 1:length(ltarget)){
    print(i)
    x=sum(ltarget[[i]]%in%rownames(r1))
    m=19
    n=14982-m
    k=length(ltarget[[i]])
      log10p<-c(log10p,phyper(x,m,n,k,lower.tail=F))
      catlength<-c(catlength,x)
      Go_Universe<-c(Go_Universe,k)
  }

attributes=c("ensembl_gene_id","external_gene_name")
nms<-getBM(attributes=attributes,filters=filters,values=ensGene,mart=mart)
nams<-nms[match(names(ltarget),nms$ensembl_gene_id),]

mapDegsDBD<-data.frame(catlength,"Regulon size"=sapply(lapply(ltarget,length),"["),
  names(ltarget),nams,FDR=p.adjust(log10p,"BH"))
mapDegsDBD<-mapDegsDBD[order(mapDegsDBD$catlength,decreasing=T),]
head(mapDegsDBD)
subset(mapDegsDBD,mapDegsDBD$names.ltarget.%in%names(ltarget[test]))
```

```
#------
#Generate count data from raw RNASeq data from Farajzadeh et al. 2013
#------

library(AnnotationDbi)
library(Rsamtools)
library(GenomicAlignments)
library(BSgenome.Sscrofa.UCSC.susScr3)
library(GenomicFeatures)
library(parallel)
library(rtracklayer)



#-------------
#Directory (locate BAM files)
#-------------

setwd("/Users/curion/Documents/suScrofaNetwork/")
bamfiles<-list.files(pattern=".bam")



#------
#scanBamParam= explore bam file
#------

scanBamHeader(bamfiles[1])
param<-ScanBamParam(tag="NH")
nhs <- scanBam(bamfiles[[1]], param=param)[[1]]$tag$NH

param <- ScanBamParam(what="cigar")

cigars <- scanBam(bamfiles[[1]], param=param)[[1]]$cigar
cigar.matrix <- cigarOpTable(cigars)
intron.size <- cigar.matrix[,"N"]
intron.size[intron.size>0]
plot(density(intron.size[intron.size>0]))

param<-ScanBamParam(what="mapq")
nameread<-scanBam(bamfiles[[1]],param=param)[[1]]

#----
#Create the Annotation variable
#histogram(log10(intron.size[intron.size>0]),xlab="intron size (log10 bp)")
#----

supportedUCSCtables()
#homos<-makeTranscriptDbFromUCSC(genome="hg19",tablename="ensGene")
#annot<-exons(homos,columns=cols)

SusScr3.tx <- makeTranscriptDbFromUCSC(genome="susScr3",
tablename="ensGene")
save(SusScr3.tx,file="SusScr3.rda")
cols<-c("tx_name","tx_id","gene_id")
geneAnot<-genes(SusScr3.tx,columns=cols)
eAnnot <- exons(SusScr3.tx)
gAnnot <- genes(SusScr3.tx)
tAnnot <- transcripts(SusScr3.tx)
```

```
colnames(elementMetadata(gAnnot))

GRList<-transcriptsBy(SusScr3.tx, by="gene")

seqlevels(tAnnot)<-
  c(c(1:18),"X","Y",seqlevels(gAnnot)[21:length(seqlevels(gAnnot))])





#---------
#function to import BAM files in R
#---------
gAlns <- mclapply(bamFileList,function(bamFile){
  open(bamFile)
  gAln <- GAlignments()
  while(length(chunk <- readGAlignmentsFromBam(bamFile))){
  gAln <- c(gAln,chunk)
  }
  close(bamFile)
  return(gAln)
  })

bamFileList <- BamFileList(bamfiles)


#---
#Prelude: change seqLevels in "Chrx" In the annotation DB!!!
#---

gAlns<-mclapply(bamfiles[1:3],readGAlignmentPairsFromBam)
chrnames<-levels(seqnames(gAlns[[1]]))

names<-sapply(chrnames,function(x) paste("chr",x,sep=""))
names(names)=NULL
length(gAlns)
for(n in 1:length(gAlns)){
  levels(seqnames(gAlns[[n]]))<-names
}

seqlevels(gAnnot)<-
  c(c(1:18),"X","Y",seqlevels(gAnnot)[21:length(seqlevels(gAnnot))])

#---
#create a count table from Bam files
#---


#testgAlns <- mclapply(bamFileList,function(bamFile){
#open(bamFile)
#gAln <- GAlignments()
#while(length(chunk <- readGAlignmentsFromBam(bamFile))){
#gAln <- c(gAln,chunk)
#}
#close(bamFile)
#return(gAln)
#})

count.list <- mclapply(bamFileList,function(bamFile){
  open(bamFile)
  counts <- vector(mode="integer",length=length(gAnnot))
  while(length(chunk <- readGAlignmentsFromBam(bamFile))){
  counts <- counts + assays(summarizeOverlaps(gAnnot,chunk,mode="Union"))$counts
  }
  close(bamFile)
  return(counts)
  })

count.table <- do.call("cbind",count.list)
head(count.table)

colnames(count.table)<-names(count.list)
rownames(count.table)<-tAnnot$tx_name
save(transcripts.count.table,count.table, file="100samplesTranscript_GeneCounts.Rdata")


#---
#Prelude: filter out genes with low counts values
#---


countCPM<-cpm(count.table)
v<-apply(countCPM,1,mean)
countCPM<-count.table[which(v>1),]

#---
#Normalization: Deseq2's VST
#---


countData=countCPM
colData=data.frame(samples=factor(colnames(count.table)),group=group)
dds <- DESeqDataSetFromMatrix( countData = countData,
                               colData = colData,
                               design = ~ group)

dds<-estimateSizeFactors(dds)
dds<-estimateDispersions(dds)
dds<-nbinomWaldTest(dds)

dds <- varianceStabilizingTransformation(dds, blind=TRUE)
gene_vstMat <- assay(dds)
colnames(gene_vstMat)<-colnames(countData)

#---
#The same Network inference Analysis is done on 100 tissues
#---

myData<-gene_vstMat

corPig100SCC<-corr.test(t(myData),method="spearman", adjust="BH",alpha=.05,ci=FALSE)

mat<-cor100Spearman$r
mat[lower.tri(mat, diag = TRUE)] <- NA
SCC.corr<-na.omit(data.frame(as.table(mat)))
mat<-cor100Spearman$p
mat[lower.tri(mat, diag = TRUE)] <- NA
SCC.Pval<-na.omit(data.frame(as.table(mat)))
SCC.corr<-SCC.corr[SCC.Pval$Freq<=0.05,]
```

```
SCC100.06<-SCC.cor[SCC.corr$Freq)>=0.6,]

g<-graph.data.frame(SCC100.06)
#---
#take transcritiption factors form data on retina to use exactly the same
#---

tf100<-tfCheck[tfCheck%in%union(SCC100.06$node1,SCC100.06$node2)]

ltarget100<-list()
for(n in 1:length(tf100)){
ltarget100[[n]]<-attributes(neighbors(g,v=tf100[n]))$names
print(n)
}
names(ltarget100)<-tf100
lapply(ltarget100,length)
rho_regulators[rho_regulators$ensembl_gene_id%in%tf100,]

#---
##Go analysis on regulons
#---

library(biomaRt)
ensembl<-useMart("ENSEMBL_MART_ENSEMBL",dataset="sscrofa_gene_ensembl",
host="www.ensembl.org")
attributes = listAttributes(ensembl)
filters = listFilters(ensembl)
filters="ensembl_gene_id"
mart=ensembl
attributes=c("ensembl_gene_id","external_gene_name","name_1006")
ensGene=V(g)$name

universe<-
getBM(attributes=attributes,filters=filters,values=union(unique(unlist(ltarget)),tf100),mart=mart)
colnames(universe)<-c('gene','name','go_term')

ltarget100<-ltarget100[-which(lapply(ltarget100,length)==0)]


regulon.GO<-list()
for(n in 1:length(ltarget100)){
  print(n)
ensGene<-ltarget100[[n]]
regulon.GO[[n]]<-getBM(attributes=attributes,filters=filters,values=ensGene,mart=mart)
}


Enrichment<-list()
for(t in 1:length(regulon.GO)){
  print(paste("regulon.GO: ",t))
  terms<-unique(regulon.GO[[t]]$name_1006)
  log10p<-NULL
  catlength<-NULL
  Go_Universe<-c()
  for (i in 1:length(terms)){

    x=length(unique(regulon.GO[[t]][regulon.GO[[t]]$name_1006==terms[[i]],"ensembl_gene_id"]))
    m=length(unique(regulon.GO[[t]]$ensembl_gene_id))
    n=length(unique(universe$gene))-length(unique(regulon.GO[[t]]$ensembl_gene_id))
    k=length(unique(universe[universe$go_term==terms[[i]],"gene"]))
```

```
      log10p<-c(log10p,phyper(x,m,n,k,lower.tail=F))
      catlength<-c(catlength,x)
      Go_Universe<-c(Go_Universe,k)
  }
  Enrichment[[t]]<-data.frame(catlength,terms,Go_Universe,log10p,BH_corrected= p.adjust(log10p,
"BH"))
}
names(Enrichment)<-tf100

#------
#Retina Specific Processes: Differential Expression Analysis
#------

colData<-rbind(colnames(countCPM),colnames(RetinaVST))
rownames(colData)<-rbind(colnames(countCPM),colnames(RetinaVST))
colData$tissue<-factor(c(rep("Ret",47),sapply(strsplit(colnames(countCPM,"-"),"[[",1))))
colData$binary<-factor(c(rep("retina",47),rep("other",100)))
int<-intersect(rownames(PigCount),rownnames(count.table))
countData<-cbind(PigCount[int,],count.table[int,])

dds_binaryRet<-DESeqDataSetFromMatrix(countData = round(countData),
      colData = colData,
      design = ~ binary)


dds_binaryRet<-DESeq(dds_binaryRet)

res<-results(dds)dds_binaryRet
res<-na.omit(res)

#----
#explore DE genes, downregulated in Retina vs all
#----

x<-subset(res, padj<.00001 & (log2FoldChange)<(-10))

rownames(x)
attributes=c("ensembl_gene_id","name_1006")
ensGene<-rownames(x)
down_group<-getBM(attributes=attributes,filters=filters,values=ensGene,mart=mart)
colnames(down_group)<-c('gene','go_term')

terms<-unique(down_group$go_term)
log10p<-NULL
catlength<-NULL
Go_Universe<-c()
for (i in 1:length(terms)){
  print(i)
  x=length(unique(down_group[down_group$go_term==terms[i],"gene"]))
  #x=length(unique(regulon.GO[[t]][regulon.GO[[t]]$name_1006==terms[[i]],"ensembl_gene_id"]))
  m=length(unique(down_group$gene))
  n=length(unique(universe$gene))-length(unique(down_group$gene))
  k=length(unique(universe[universe$go_term==terms[[i]],"gene"]))
  log10p<-c(log10p,phyper(x,m,n,k,lower.tail=F))
  catlength<-c(catlength,x)
  Go_Universe<-c(Go_Universe,k)
}
down_enrichment<-data.frame(catlength,terms,Go_Universe,log10p,BH_corrected= p.adjust(log10p,
"BH"))
down_enrichment<-down_enrichment[order(down_enrichment$BH_corrected,decreasing=F),]
```

```
#-------
#upregulated genes
#------

x<-subset(res, padj<.00001 & (log2FoldChange)>(10))

rownames(x)
attributes=c("ensembl_gene_id","name_1006")
ensGene<-rownames(x)
up_group<-getBM(attributes=attributes,filters=filters,values=ensGene,mart=mart)
colnames(up_group)<-c('gene','go_term')

terms<-unique(up_group$go_term)
log10p<-NULL
catlength<-NULL
Go_Universe<-c()
for (i in 1:length(terms)){
  print(i)
  x=length(unique(up_group[up_group$go_term==terms[i],"gene"]))
  #x=length(unique(regulon.GO[[t]][regulon.GO[[t]]$name_1006==terms[[i]],"ensembl_gene_id"]))
  m=length(unique(up_group$gene))
  n=length(unique(universe$gene))-length(unique(up_group$gene))
  k=length(unique(universe[universe$go_term==terms[[i]],"gene"]))
  log10p<-c(log10p,phyper(x,m,n,k,lower.tail=F))
  catlength<-c(catlength,x)
  Go_Universe<-c(Go_Universe,k)
}
up_enrichment<-data.frame(catlength,terms,Go_Universe,log10p,BH_corrected= p.adjust(log10p,
"BH"))
up_enrichment<-up_enrichment[order(up_enrichment$BH_corrected,decreasing=F),]

#---
#volcano plot
#---

with(res, plot(log2FoldChange, -log10(pvalue), pch=20, main="Volcano plot", xlim=c(-23,20),
ylim=c(-4,330)))
# Add colored points: orange if padj<thr & log2FC>thr2)
with(subset(res, padj<.000001 & abs(log2FoldChange)>10), points(log2FoldChange, -log10(pvalue),
pch=20, col="orange"))
```

Bibliography

[1] Anat London, Inbal Benhar, and Michal Schwartz. The retina as a window to the brain—from eye research to cns disorders. *Nature Reviews Neurology*, 9(1):44–53, 2013.

[2] Dale Purves, George J Augustine, David Fitzpatrick, Lawrence C Katz, Anthony-Samuel LaMantia, James O McNamara, and S Mark Williams. Central visual pathways. 2001.

[3] Richard H Masland. The fundamental plan of the retina. *Nature neuroscience*, 4(9):877–886, 2001.

[4] Botond Roska and Frank Werblin. Vertical interactions across ten parallel, stacked representations in the mammalian retina. *Nature*, 410 (6828):583–587, 2001.

[5] Sandra Siegert, Brigitte Gross Scherf, Karina Del Punta, Nick Didkovsky, Nathaniel Heintz, and Botond Roska. Genetic address book for retinal cell types. *Nature neuroscience*, 12(9):1197–1204, 2009.

[6] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5): 1202–1214, 2015.

[7] Connie Cepko. Intrinsically different retinal progenitor cells produce specific types of progeny. *Nature Reviews Neuroscience*, 2014.

[8] Daniel Goldman. Muller glial cell reprogramming and retina regeneration. *Nature Reviews Neuroscience*, 15(7):431–442, 2014.

[9] Heinz Wässle. Parallel processing in the mammalian retina. *Nature Reviews Neuroscience*, 5(10):747–757, 2004.

[10] RG Foster, I Provencio, D Hudson, S Fiske, W De Grip, and M Menaker. Circadian photoreception in the retinally degenerate mouse (rd/rd). *Journal of Comparative Physiology A*, 169(1):39–50, 1991.

[11] Richard W Young. The renewal of photoreceptor cell outer segments. *The Journal of cell biology*, 33(1):61–72, 1967.

[12] Yingyu Mao and Silvia C Finnemann. Analysis of photoreceptor outer segment phagocytosis by rpe cells in culture. In *Retinal Degeneration*, pages 285–295. Springer, 2013.

[13] Thomas Euler, Silke Haverkamp, Timm Schubert, and Tom Baden. Retinal bipolar cells: elementary building blocks of vision. *Nature Reviews Neuroscience*, 15(8):507–519, 2014.

[14] Moritz Helmstaedter, Kevin L Briggman, Srinivas C Turaga, Viren Jain, H Sebastian Seung, and Winfried Denk. Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature*, 500 (7461):168–174, 2013.

[15] Ross A Poche and Benjamin E Reese. Retinal horizontal cells: challenging paradigms of neural development and cancer biology. *Development*, 136(13):2141–2151, 2009.

[16] Evelyne Sernagor, Stephen Eglen, Bill Harris, and Rachel Wong. *Retinal development*. Cambridge University Press, 2012.

[17] Chang-Jin Jeon, Enrica Strettoi, and Richard H Masland. The major cell populations of the mouse retina. *The Journal of neuroscience*, 18 (21):8936–8946, 1998.

[18] Kristian Franze, Jens Grosche, Serguei N Skatchkov, Stefan Schinkinger, Christian Foja, Detlev Schild, Ortrud Uckermann, Kort Travis, Andreas Reichenbach, and Jochen Guck. Müller cells are living optical fibers in the vertebrate retina. *Proceedings of the National Academy of Sciences*, 104(20):8287–8292, 2007.

[19] Andreas Bringmann, Thomas Pannicke, Bernd Biedermann, Mike Francke, Ianors Iandiev, Jens Grosche, Peter Wiedemann, Jan Albrecht, and Andreas Reichenbach. Role of retinal glial cells in neurotransmitter uptake and metabolism. *Neurochemistry international*, 54(3):143–160, 2009.

[20] Andreas Bringmann and Peter Wiedemann. Müller glial cells in retinal disease. *Ophthalmologica*, 227(1):1–19, 2011.

[21] Ani V Das, Kavita B Mallya, Xing Zhao, Faraz Ahmad, Sumitra Bhattacharya, Wallace B Thoreson, Ganapati V Hegde, and Iqbal Ahmad. Neural stem cell properties of müller glia in the mammalian retina: regulation by notch and wnt signaling. *Developmental biology*, 299(1): 283–302, 2006.

[22] Jean M Lawrence, Shweta Singhal, Bhairavi Bhatia, David J Keegan, Thomas A Reh, Philip J Luthert, Peng T Khaw, and Gloria Astrid Limb. Mio-m1 cells and similar müller glial cell lines derived from

adult human retina exhibit neural stem cell characteristics. *Stem cells*, 25(8):2033–2043, 2007.

[23] Shweta Singhal, Bhairavi Bhatia, Hari Jayaram, Silke Becker, Megan F Jones, Phillippa B Cottrill, Peng T Khaw, Thomas E Salt, and G Astrid Limb. Human müller glia with stem cell characteristics differentiate into retinal ganglion cell (rgc) precursors in vitro and partially restore rgc function in vivo following transplantation. *Stem cells translational medicine*, 1(3):188–199, 2012.

[24] Serena G Giannelli, Gian Carlo Demontis, Grazia Pertile, Paolo Rama, and Vania Broccoli. Adult human müller glia cells are a highly efficient source of rod photoreceptors. *Stem Cells*, 29(2):344–356, 2011.

[25] Richard H Masland. Neuronal diversity in the retina. *Current opinion in neurobiology*, 11(4):431–436, 2001.

[26] Revathi Balasubramanian and Lin Gan. Development of retinal amacrine cells and their dendritic stratification. *Current ophthalmology reports*, 2(3):100–106, 2014.

[27] WR Taylor and RG Smith. The role of starburst amacrine cells in visual signal processing. *Visual neuroscience*, 29(01):73–81, 2012.

[28] Helga Kolb, Eduardo Fernandez, Ralph Nelson, and Helga Kolb. Roles of amacrine cells. 2007.

[29] Ralph Nelson. Visual responses of ganglion cells. 2007.

[30] David M Berson, Felice A Dunn, and Motoharu Takao. Phototransduction by retinal ganglion cells that set the circadian clock. *Science*, 295(5557):1070–1073, 2002.

[31] Andrew B Watson. A formula for human retinal ganglion cell receptive field density as a function of visual field location. *Journal of vision*, 14 (7):15, 2014.

[32] Julie B Shelton, Kathleen B Digre, James Gilman, Judith EA Warner, and Bradley J Katz. Characteristics of myelinated retinal nerve fiber layer in ophthalmic imaging: Findings on autofluorescence, fluorescein angiographic, infrared, optical coherence tomographic, and red-free images. *JAMA ophthalmology*, 131(1):107–109, 2013.

[33] Yoshinori Shichida and Take Matsuyama. Evolution of opsins and phototransduction. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1531):2881–2895, 2009.

[34] Dale Purves, George J Augustine, David Fitzpatrick, Lawrence C Katz, Anthony-Samuel LaMantia, James O McNamara, and S Mark Williams. Phototransduction. 2001.

[35] Krzysztof Palczewski. G protein–coupled receptor rhodopsin. *Annual review of biochemistry*, 75:743, 2006.

[36] Anne R Murray, Linda Vuong, Daniel Brobst, Steven J Fliesler, Neal S Peachey, Marina S Gorbatyuk, Muna I Naash, and Muayyad R Al-Ubaidi. Glycosylation of rhodopsin is necessary for its stability and incorporation into photoreceptor outer segment discs. *Human molecular genetics*, page ddv031, 2015.

[37] Liya Yuan, Mariko Kawada, Necat Havlioglu, Hao Tang, and Jane Y Wu. Mutations in prpf31 inhibit pre-mrna splicing of rhodopsin gene and cause apoptosis of retinal cells. *The Journal of neuroscience*, 25 (3):748–757, 2005.

[38] Edwin S Lee and John G Flannery. Transport of truncated rhodopsin and its effects on rod function and degeneration. *Investigative ophthalmology & visual science*, 48(6):2868, 2007.

[39] Sławomir Filipek, Ronald E Stenkamp, David C Teller, and Krzysztof Palczewski. G protein-coupled receptor rhodopsin: a prospectus. *Annual review of physiology*, 65:851, 2003.

[40] Elaine C Meng and Henry R Bourne. Receptor activation: what does the rhodopsin structure tell us? *Trends in pharmacological sciences*, 22(11):587–593, 2001.

[41] Kiweon Cha, Philip J Reeves, and H Gobind Khorana. Structure and function in rhodopsin: Destabilization of rhodopsin by the binding of an antibody at the n-terminal segment provides support for involvement of the latter in an intradiscal tertiary structure. *Proceedings of the National Academy of Sciences*, 97(7):3016–3021, 2000.

[42] Paul A Hargrave. Rhodopsin structure, function, and topography the friedenwald lecture. *Investigative ophthalmology & visual science*, 42 (1):3–9, 2001.

[43] Thaddeus P Dryja. Rhodopsin and autosomal dominant retinitis pigmentosa. *Eye*, 6(1):1–10, 1992.

[44] Centers for Disease Control et al. Gateway to health communication and social marketing practice. *Availab le www. cdc. gov/healthcommunication/HealthBasics/WhatIsHC. html (Accessed: August 8, 2012)*, 2011.

[45] Retnet - retinal information network. URL `https://sph.uth.edu/RetNet/`.

[46] Laura R Pacione, Michael J Szego, Sakae Ikeda, Patsy M Nishina, and Roderick R McInnes. Progress toward understanding the genetic and biochemical mechanisms of inherited photoreceptor degenerations. *Annual review of neuroscience*, 26(1):657–700, 2003.

[47] DA Newsome. Retinitis pigmentosa, usher's syndrome, and other pigmentary retinopathies. In *Retinal dystrophies and degenerations*, pages 161–194. Raven Press New York, 1988.

[48] Christian P Hamel. Cone rod dystrophies. *Orphanet J Rare Dis*, 2(7): 1750–72, 2007.

[49] DY Wang, WM Chan, POS Tam, L Baum, DSC Lam, KKL Chong, BJ Fan, and CP Pang. Gene mutations in retinitis pigmentosa and their clinical implications. *Clinica chimica acta*, 351(1):5–16, 2005.

[50] Kelly Shintani, Diana L Shechtman, and Andrew S Gurwood. Review and update: current treatment trends for patients with retinitis pigmentosa. *Optometry-Journal of the American Optometric Association*, 80(7):384–401, 2009.

[51] Ann H Milam, Zong-Yi Li, and Robert N Fariss. Histopathology of the human retina in retinitis pigmentosa. *Progress in retinal and eye research*, 17(2):175–206, 1998.

[52] C Portera-Cailliau, CH Sung, J Nathans, and R Adler. Apoptotic photoreceptor cell death in mouse models of retinitis pigmentosa. *Proceedings of the National Academy of Sciences*, 91(3):974–978, 1994.

[53] Richard N Lolley, Hongmei Rong, and CM Craft. Linkage of photoreceptor degeneration by apoptosis with inherited defect in phototransduction. *Investigative ophthalmology & visual science*, 35(2):358–362, 1994.

[54] Patti C Huang, Alicia E Gaitan, Ying Hao, Robert M Petters, and Fulton Wong. Cellular interactions implicated in the mechanism of photoreceptor degeneration in transgenic mice expressing a mutant rhodopsin gene. *Proceedings of the National Academy of Sciences*, 90 (18):8484–8488, 1993.

[55] Wojciech Kedzierski, Dean Bok, and Gabriel H Travis. Non-cell-autonomous photoreceptor degeneration in rdsmutant mice mosaic for expression of a rescue transgene. *The Journal of neuroscience*, 18(11): 4076–4082, 1998.

[56] Gabriel H Travis. Mechanisms of cell death in the inherited retinal degenerations. *The American Journal of Human Genetics*, 62(3):503–508, 1998.

[57] Amir Rattner, Hui Sun, and Jeremy Nathans. Molecular genetics of human retinal disease. *Annual review of genetics*, 33(1):89–131, 1999.

[58] Gesine B Jaissle, Christian Albrecht May, Serge A van de Pavert, Andreas Wenzel, Ellen Claes-May, Andreas Gießl, Peter Szurman, Uwe Wolfrum, Jan Wijnholds, MD Fisher, et al. Bone spicule pigment formation in retinitis pigmentosa: insights from a mouse model. *Graefe's archive for clinical and experimental ophthalmology*, 248(8):1063–1070, 2010.

[59] James K Phelan and Dean Bok. A brief review of retinitis pigmentosa and the identified retinitis pigmentosa genes. *Mol Vis*, 6:116–124, 2000.

[60] Dyonne T Hartong, Eliot L Berson, and Thaddeus P Dryja. Retinitis pigmentosa. *The Lancet*, 368(9549):1795–1809, 2006.

[61] Debora B Farber and Richard N Lolley. Cyclic guanosine monophosphate: elevation in degenerating photoreceptor cells of the c3h mouse retina. *Science*, 186(4162):449–451, 1974.

[62] Anna Andrés, Pere Garriga, and Joan Manyosa. Altered functionality in rhodopsin point mutants associated with retinitis pigmentosa. *Biochemical and biophysical research communications*, 303(1):294–301, 2003.

[63] CH Sung, CM Davenport, and J Nathans. Rhodopsin mutations responsible for autosomal dominant retinitis pigmentosa. clustering of functional classes along the polypeptide chain. *Journal of Biological Chemistry*, 268(35):26645–26649, 1993.

[64] Kean T Oh, Reid Longmuir, Dawn M Oh, Edwin M Stone, Kelly Kopp, Jeremiah Brown, Gerald A Fishman, Peter Sonkin, Karen M Gehrs, and Richard G Weleber. Comparison of the clinical expression of retinitis pigmentosa associated with rhodopsin mutations at codon 347 and codon 23. *American journal of ophthalmology*, 136(2):306–313, 2003.

[65] Eliot L Berson, Bernard Rosner, Michael A Sandberg, KC Hayes, Britain W Nicholson, Carol Weigel-DiFranco, and Walter Willett. A randomized trial of vitamin a and vitamin e supplementation for retinitis pigmentosa. *Archives of ophthalmology*, 111(6):761–772, 1993.

[66] Maria Frasson, Jose A Sahel, Michel Fabre, Manuel Simonutti, Henri Dreyfus, and Serge Picaud. Retinitis pigmentosa: rod photoreceptor rescue by a calcium-channel blocker in the rd mouse. *Nature medicine*, 5(10):1183–1187, 1999.

[67] Prateek K Buch, Robert E MacLaren, Yanaí Durán, Kamaljit S Balaggan, Angus MacNeil, Frank C Schlichtenbrede, Alexander J Smith, and

Robin R Ali. In contrast to aav-mediated cntf expression, aav-mediated gdnf expression enhances gene replacement therapy in rodent models of retinal degeneration. *Molecular therapy*, 14(5):700–709, 2006.

[68] M Tschernutter, FC Schlichtenbrede, S Howe, KS Balaggan, PM Munro, JWB Bainbridge, AJ Thrasher, AJ Smith, and RR Ali. Long-term preservation of retinal function in the rcs rat model of retinitis pigmentosa following lentivirus-mediated gene therapy. *Gene therapy*, 12(8):694–701, 2005.

[69] T Hashimoto, D Gibbs, C Lillo, SM Azarian, E Legacki, XM Zhang, XJ Yang, and DS Williams. Lentiviral gene replacement therapy of retinas in a mouse model for usher syndrome type 1b. *Gene therapy*, 14(7):584–594, 2007.

[70] Gregory M Acland, Gustavo D Aguirre, Jharna Ray, Qi Zhang, Tomas S Aleman, Artur V Cideciyan, Susan E Pearce-Kelling, Vibha Anand, Yong Zeng, Albert M Maguire, et al. Gene therapy restores vision in a canine model of childhood blindness. *Nature genetics*, 28 (1):92–95, 2001.

[71] Michel Cayouette and Claude Gravel. Adenovirus-mediated gene transfer of ciliary neurotrophic factor can prevent photoreceptor degeneration in the retinal degeneration (rd) mouse. *Human gene therapy*, 8(4): 423–430, 1997.

[72] Sophia Millington-Ward, Brian O'Neill, Gearoid Tuohy, Najma Al-Jandal, Anna-Sophia Kiang, Paul F Kenna, Arpad Palfi, Patrick Hayden, Fiona Mansergh, Avril Kennan, et al. Strategems in vitro for gene therapies directed to dominant mutations. *Human Molecular Genetics*, 6(9):1415–1426, 1997.

[73] Matthew JA Wood, Barbara Trülzsch, Amr Abdelgany, and David Beeson. Therapeutic gene silencing in the nervous system. *Human molecular genetics*, 12(suppl 2):R279–R284, 2003.

[74] A Klug. Towards therapeutic applications of engineered zinc finger proteins. *FEBS letters*, 579(4):892–894, 2005.

[75] Scot A Wolfe, Lena Nekludova, and Carl O Pabo. Dna recognition by cys2his2 zinc finger proteins. *Annual review of biophysics and biomolecular structure*, 29(1):183–212, 2000.

[76] Jens Boch and Ulla Bonas. Xanthomonas avrbs3 family-type iii effectors: discovery and function. *Phytopathology*, 48(1):419, 2010.

[77] Joshua F Meckler, Mital S Bhakta, Moon-Soo Kim, Robert Ovadia, Chris H Habrian, Artem Zykovich, Abigail Yu, Sarah H Lockwood, Robert Morbitzer, Janett Elsäesser, et al. Quantitative analysis of tale–dna interactions suggests polarity effects. *Nucleic acids research*, 41(7):4118–4128, 2013.

[78] Rodolphe Barrangou, Christophe Fremaux, Hélene Deveau, Melissa Richards, Patrick Boyaval, Sylvain Moineau, Dennis A Romero, and Philippe Horvath. Crispr provides acquired resistance against viruses in prokaryotes. *Science*, 315(5819):1709–1712, 2007.

[79] Elitza Deltcheva, Krzysztof Chylinski, Cynthia M Sharma, Karine Gonzales, Yanjie Chao, Zaid A Pirzada, Maria R Eckert, Jörg Vogel, and Emmanuelle Charpentier. Crispr rna maturation by trans-encoded small rna and host factor rnase iii. *Nature*, 471(7340):602–607, 2011.

[80] Yanfang Fu, Jeffry D Sander, Deepak Reyon, Vincent M Cascio, and J Keith Joung. Improving crispr-cas nuclease specificity using truncated guide rnas. *Nature biotechnology*, 32(3):279–284, 2014.

[81] Morgan L Maeder, Samantha J Linder, Vincent M Cascio, Yanfang Fu, Quan H Ho, and J Keith Joung. Crispr rna-guided activation of endogenous human genes. *Nature methods*, 10(10):977–979, 2013.

[82] Fahim Farzadfard, Samuel D Perli, and Timothy K Lu. Tunable and multifunctional eukaryotic transcription factors based on crispr/cas. *ACS synthetic biology*, 2(10):604–613, 2013.

[83] Lei S Qi, Matthew H Larson, Luke A Gilbert, Jennifer A Doudna, Jonathan S Weissman, Adam P Arkin, and Wendell A Lim. Repurposing crispr as an rna-guided platform for sequence-specific control of gene expression. *Cell*, 152(5):1173–1183, 2013.

[84] Wusheng Liu, Joshua S Yuan, and C Neal Stewart Jr. Advanced genetic tools for plant biotechnology. *Nature Reviews Genetics*, 14(11):781–793, 2013.

[85] Claudio Mussolino, Daniela Sanges, Elena Marrocco, Ciro Bonetti, Umberto Di Vicino, Valeria Marigo, Alberto Auricchio, Germana Meroni, and Enrico Maria Surace. Zinc-finger-based transcriptional repression of rhodopsin in a model of dominant retinitis pigmentosa. *EMBO molecular medicine*, 3(3):118–128, 2011.

[86] Fatih Ozsolak and Patrice M Milos. Rna sequencing: advances, challenges and opportunities. *Nature reviews genetics*, 12(2):87–98, 2011.

[87] Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by rna sequencing. *Science*, 320 (5881):1344–1349, 2008.

[88] Shawn C Baker, Steven R Bauer, Richard P Beyer, James D Brenton, Bud Bromley, John Burrill, Helen Causton, Michael P Conley, Rosalie

Elespuru, Michael Fero, et al. The external rna controls consortium: a progress report. *Nature methods*, 2(10):731–734, 2005.

[89] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628, 2008.

[90] John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517, 2008.

[91] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel Gaffney, Laura L Elo, Xuegong Zhang, et al. A survey of best practices for rna-seq data analysis. 2016.

[92] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1): 57–63, 2009.

[93] Carsten A Raabe, Thean-Hock Tang, Juergen Brosius, and Timofey S Rozhdestvensky. Biases in small rna deep sequencing data. *Nucleic acids research*, 42(3):1414–1426, 2014.

[94] Shujun Luo. Microrna expression analysis using the illumina microrna-seq platform. *Next-Generation MicroRNA Expression Profiling Technology: Methods and Protocols*, pages 183–188, 2012.

[95] Piero Carninci, Albin Sandelin, Boris Lenhard, Shintaro Katayama, Kazuro Shimokawa, Jasmina Ponjavic, Colin AM Semple, Martin S Taylor, Pär G Engström, Martin C Frith, et al. Genome-wide analysis

of mammalian promoter architecture and evolution. *Nature genetics*, 38(6):626–635, 2006.

[96] Tatiana Borodina, James Adjaye, and Marc Sultan. A strand-specific library preparation protocol for rna sequencing. *Methods Enzymol*, 500: 79–98, 2011.

[97] Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E Mason, Nicholas D Socci, and Doron Betel. Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. *Genome Biol*, 14(9):R95, 2013.

[98] Sonia Tarazona, Fernando García-Alcalde, Joaquín Dopazo, Alberto Ferrer, and Ana Conesa. Differential expression in rna-seq: a matter of depth. *Genome research*, 21(12):2213–2223, 2011.

[99] David G Robinson and John D Storey. subseq: determining appropriate sequencing depth through efficient read subsampling. *Bioinformatics*, 30(23):3424–3426, 2014.

[100] Picard. URL `http://picard.sourcefoge.net/`.

[101] Simon Andrews. Fastqc: A quality control tool for high throughput sequence data. http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/.

[102] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, page btu170, 2014.

[103] Felix Krueger. Trimgalore! http://www.bioinformatics.bbsrc.ac.uk/projects/.

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

[105] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[106] Ben Langmead, Cole Trapnell, Mihai Pop, Steven L Salzberg, et al. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biol*, 10(3):R25, 2009.

[107] Michael Burrows and David Wheeler. A block-sorting lossless data compression algorithm. In *DIGITAL SRC RESEARCH REPORT*. Citeseer, 1994.

[108] Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, et al. De novo transcript sequence reconstruction from rna-seq using the trinity platform for reference generation and analysis. *Nature protocols*, 8(8):1494–1512, 2013.

[109] Yinlong Xie, Gengxiong Wu, Jingbo Tang, Ruibang Luo, Jordan Patterson, Shanlin Liu, Weihua Huang, Guangzhu He, Shengchang Gu, Shengkang Li, et al. Soapdenovo-trans: de novo transcriptome assembly with short rna-seq reads. *Bioinformatics*, 30(12):1660–1666, 2014.

[110] Bo Li and Colin N Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323, 2011.

[111] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals

unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.

[112] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[113] Xiaobei Zhou, Helen Lindsay, and Mark D Robinson. Robustly detecting differential expression in rna sequencing data using observation weights. *Nucleic acids research*, 42(11):e91, 2014.

[114] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biol*, 11(10):R106, 2010.

[115] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*, 15(12):550, 2014.

[116] Mark D Robinson, Alicia Oshlack, et al. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biol*, 11(3):R25, 2010.

[117] Peter McCullagh and James A Nelder. Generalized linear models, no. 37 in monograph on statistics and applied probability, 1989.

[118] Michael Love, Simon Anders, and Wolfgang Huber. Differential analysis of count data–the deseq2 package. *Genome Biology*, 15:550, 2014.

[119] Wolfgang Huber, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, et al. Orchestrating high-throughput genomic analysis with bioconductor. *Nature methods*, 12(2):115–121, 2015.

[120] Rickard Sandberg. Entering the era of single-cell transcriptomics in biology and medicine. *Nature methods*, 11(1):22–24, 2014.

[121] Arjun Raj, Patrick Van Den Bogaard, Scott A Rifkin, Alexander Van Oudenaarden, and Sanjay Tyagi. Imaging individual mrna molecules using multiple singly labeled probes. *Nature methods*, 5(10): 877, 2008.

[122] Kiyomi Taniguchi, Tomoharu Kajiyama, and Hideki Kambara. Quantitative analysis of gene expression in a single cell by qpcr. *Nature methods*, 6(7):503, 2009.

[123] Oliver Stegle, Sarah A Teichmann, and John C Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, 2015.

[124] Spyros Darmanis, Steven A Sloan, Ye Zhang, Martin Enge, Christine Caneda, Lawrence M Shuer, Melanie G Hayden Gephart, Ben A Barres, and Stephen R Quake. A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences*, 112(23):7285–7290, 2015.

[125] Philip Brennecke, Alejandro Reyes, Sheena Pinto, Kristin Rattay, Michelle Nguyen, Rita Küchler, Wolfgang Huber, Bruno Kyewski, and Lars M Steinmetz. Single-cell transcriptome analysis reveals coordinated ectopic gene-expression patterns in medullary thymic epithelial cells. *Nature immunology*, 2015.

[126] Sameer S Bajikar, Christiane Fuchs, Andreas Roller, Fabian J Theis, and Kevin A Janes. Parameterizing cell-to-cell regulatory heterogeneities via stochastic transcriptional profiles. *Proceedings of the National Academy of Sciences*, 111(5):E626–E635, 2014.

[127] Antoine-Emmanuel Saliba, Alexander J Westermann, Stanislaw A Gorski, and Jörg Vogel. Single-cell rna-seq: advances and future challenges. *Nucleic acids research*, 42(14):8845–8860, 2014.

[128] Vipul Bhargava, Pang Ko, Erik Willems, Mark Mercola, and Shankar Subramaniam. Quantitative transcriptomics using designed primer-based amplification. *Scientific reports*, 3, 2013.

[129] Vipul Bhargava, Steven R Head, Phillip Ordoukhanian, Mark Mercola, and Shankar Subramaniam. Technical variations in low-input rna-seq methodologies. *Scientific reports*, 4, 2014.

[130] Daniel J Woodsworth, Mauro Castellarin, and Robert A Holt. Sequence analysis of t-cell repertoires in health and disease. *Genome Med*, 5(10): 98, 2013.

[131] Maria A Turchaninova, Olga V Britanova, Dmitriy A Bolotin, Mikhail Shugay, Ekaterina V Putintseva, Dmitriy B Staroverov, George Sharonov, Dmitriy Shcherbo, Ivan V Zvyagin, Ilgar Z Mamedov, et al. Pairing of t-cell receptor chains via emulsion pcr. *European journal of immunology*, 43(9):2507–2515, 2013.

[132] Tamar Hashimshony, Florian Wagner, Noa Sher, and Itai Yanai. Cel-seq: single-cell rna-seq by multiplexed linear amplification. *Cell reports*, 2(3):666–673, 2012.

[133] Saiful Islam, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex rna-seq. *Genome research*, 21(7):1160–1167, 2011.

[134] Simone Picelli, Åsa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. Smart-seq2 for sensitive full-

length transcriptome profiling in single cells. *Nature methods*, 10(11): 1096–1098, 2013.

[135] Simone Picelli, Omid R Faridani, Åsa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length rna-seq from single cells using smart-seq2. *Nature protocols*, 9(1):171–181, 2014.

[136] Illumina ®. Single cell research - an overview of recent single cell research publications featuring illumina® technology. http://www.illumina.com.

[137] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods*, 9(1):72–74, 2012.

[138] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33(2):155–160, 2015.

[139] Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. The role of spike-in standards in the normalization of rna-seq. In *Statistical Analysis of Next Generation Sequencing Data*, pages 169–190. Springer, 2014.

[140] Jakob Lovén, David A Orlando, Alla A Sigova, Charles Y Lin, Peter B Rahl, Christopher B Burge, David L Levens, Tong Ihn Lee, and Richard A Young. Revisiting global gene expression analysis. *Cell*, 151 (3):476–482, 2012.

[141] Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. Normalization of rna-seq data using factor analysis of control genes or samples. *Nature biotechnology*, 32(9):896–902, 2014.

[142] Catalina A Vallejos, John C Marioni, and Sylvia Richardson. Basics: Bayesian analysis of single-cell sequencing data. *PLoS Comput Biol*, 11(6):e1004333, 2015.

[143] Michael Hecker, Sandro Lambeck, Susanne Toepfer, Eugene Van Someren, and Reinhard Guthke. Gene regulatory network inference: data integration in dynamic models—a review. *Biosystems*, 96 (1):86–103, 2009.

[144] Riet De Smet and Kathleen Marchal. Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*, 8 (10):717–729, 2010.

[145] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[146] Huai Li, Yu Sun, and Ming Zhan. Exploring pathways from gene co-expression to network dynamics. *Computational Systems Biology*, pages 249–267, 2009.

[147] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):1, 2008.

[148] Vincenzo Belcastro, Velia Siciliano, Francesco Gregoretti, Pratibha Mithbaokar, Gopuraja Dharmalingam, Stefania Berlingieri, Francesco Iorio, Gennaro Oliva, Roman Polishchuck, Nicola Brunetti-Pierri, et al.

Transcriptional gene network inference from a massive dataset elucidates transcriptome organization and gene function. *Nucleic acids research*, 39(20):8677–8688, 2011.

[149] Mukesh Bansal, Vincenzo Belcastro, Alberto Ambesi-Impiombato, and Diego Di Bernardo. How to infer gene networks from expression profiles. *Molecular systems biology*, 3(1):78, 2007.

[150] Christopher A Penfold and David L Wild. How to infer gene networks from expression profiles, revisited. *Interface Focus*, 1(6):857–870, 2011.

[151] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo D Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(Suppl 1):S7, 2006.

[152] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian von Mering, et al. String v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(D1):D808–D815, 2013.

[153] Irina Voineagu, Xinchen Wang, Patrick Johnston, Jennifer K Lowe, Yuan Tian, Steve Horvath, Jonathan Mill, Rita M Cantor, Benjamin J Blencowe, and Daniel H Geschwind. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*, 474(7351):380–384, 2011.

[154] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[155] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[156] Sara Mostafavi, Debajyoti Ray, David Warde-Farley, Chris Grouios, Quaid Morris, et al. Genemania: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol*, 9(Suppl 1):S4, 2008.

[157] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *The Journal of Machine Learning Research*, 7:551–585, 2006.

[158] S Ballouz, W Verleyen, and J Gillis. Guidance for rna-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics*, 31(13):2123–2130, 2015.

[159] Federico M Giorgi, Cristian Del Fabbro, and Francesco Licausi. Comparative study of rna-seq-and microarray-derived coexpression networks in arabidopsis thaliana. *Bioinformatics*, 29(6):717–724, 2013.

[160] Gökhan Gün and Wilfried A Kues. Current progress of genetically engineered pig models for biomedical research. *BioResearch open access*, 3(6):255–264, 2014.

[161] MJ Chandler, PJ Smith, DA Samuelson, and EO MacKay. Photoreceptor density of the domestic pig retina. *Vet Ophthalmol*, 2(3):179–184, 1999.

[162] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[163] Ulrich Bodenhofer, Andreas Kothmeier, and Sepp Hochreiter. Apcluster: an r package for affinity propagation clustering. *Bioinformatics*, 27(17):2463–2464, 2011.

[164] Anand Swaroop, Douglas Kim, and Douglas Forrest. Transcriptional regulation of photoreceptor development and homeostasis in the mammalian retina. *Nature Reviews Neuroscience*, 11(8):563–576, 2010.

[165] Isabelle Rivals, Léon Personnaz, Lieng Taing, and Marie-Claude Potier. Enrichment or depletion of a go category within a class of genes: which test? *Bioinformatics*, 23(4):401–407, 2007.

[166] Douglas Forrest and Anand Swaroop. Minireview: the role of nuclear receptors in photoreceptor differentiation and disease. *Molecular Endocrinology*, 26(6):905–915, 2012.

[167] Akishi Onishi, Guang-Hua Peng, Erin M Poth, Daniel A Lee, Jichao Chen, Uel Alexis, Jimmy de Melo, Shiming Chen, and Seth Blackshaw. The orphan nuclear hormone receptor errβ controls rod photoreceptor survival. *Proceedings of the National Academy of Sciences*, 107(25): 11579–11584, 2010.

[168] Sofia Henriksson and Marianne Farnebo. On the road with wrap53β: guardian of cajal bodies and genome integrity. *Frontiers in genetics*, 6, 2015.

[169] Leila Farajzadeh, Henrik Hornshøj, Jamal Momeni, Bo Thomsen, Knud Larsen, Jakob Hedegaard, Christian Bendixen, and Lone Bruhn Madsen. Pairwise comparisons of ten porcine tissues identify differential transcriptional regulation at the gene, isoform, promoter and transcription start site level. *Biochemical and biophysical research communications*, 438(2):346–352, 2013.

[170] Deyi Duan, Yuhong Fu, George Paxinos, and Charles Watson. Spatiotemporal expression patterns of pax6 in the brain of embryonic, new-

born, and adult mice. *Brain Structure and Function*, 218(2):353–372, 2013.

[171] Michael H Farkas, Gregory R Grant, Joseph A White, Maria E Sousa, Mark B Consugar, and Eric A Pierce. Transcriptome analyses of the human retina identify unprecedented transcript diversity and 3.5 mb of novel transcribed sequence via significant alternative splicing and novel genes. *BMC genomics*, 14(1):486, 2013.

[172] Sohail Malik and Robert G Roeder. The metazoan mediator co-activator complex as an integrative hub for transcriptional regulation. *Nature Reviews Genetics*, 11(11):761–772, 2010.

[173] Valentina Perissi, Kristen Jepsen, Christopher K Glass, and Michael G Rosenfeld. Deconstructing repression: evolving models of co-repressor action. *Nature Reviews Genetics*, 11(2):109–123, 2010.

[174] Kenneth P Mitton, Prabodh K Swain, Shiming Chen, Siqun Xu, Donald J Zack, and Anand Swaroop. The leucine zipper of nrl interacts with the crx homeodomain a possible mechanism of transcriptional synergy in rhodopsin regulation. *Journal of Biological Chemistry*, 275 (38):29794–29799, 2000.

[175] J Lee, CA Myers, N Williams, M Abdelaziz, and JC Corbo. Quantitative fine-tuning of photoreceptor cis-regulatory elements through affinity modulation of transcription factor binding sites. *Gene therapy*, 17(11):1390–1399, 2010.

[176] Jerome E Roger, Avinash Hiriyanna, Norimoto Gotoh, Hong Hao, Debbie F Cheng, Rinki Ratnapriya, Marie-Audrey I Kautzmann, Bo Chang, and Anand Swaroop. Otx2 loss causes rod differentiation

defect in crx-associated congenital blindness. *The Journal of clinical investigation*, 124(2):631, 2014.

[177] Krzysztof Palczewski. Chemistry and biology of vision. *Journal of Biological Chemistry*, 287(3):1612–1619, 2012.

[178] Marijana Samardzija, Hedwig Wariwoda, Cornelia Imsand, Philipp Huber, Severin R Heynen, Andrea Gubler, and Christian Grimm. Activation of survival pathways in the degenerating retina of rd10 mice. *Experimental eye research*, 99:17–26, 2012.

[179] Liang Li, Bin Li, Hao Zhang, Shuwei Bai, Yichen Wang, Bowen Zhao, and Jost B Jonas. Lentiviral vector-mediated pax6 overexpression promotes growth and inhibits apoptosis of human retinoblastoma cells. *Investigative ophthalmology & visual science*, 52(11):8393–8400, 2011.

[180] Jianhai Du, Whitney M Cleghorn, Laura Contreras, Ken Lindsay, Austin M Rountree, Andrei O Chertov, Sally J Turner, Ayse Sahaboglu, Jonathan Linton, Martin Sadilek, et al. Inhibition of mitochondrial pyruvate transport by zaprinast causes massive accumulation of aspartate at the expense of glutamate in the retina. *Journal of Biological Chemistry*, 288(50):36129–36140, 2013.

[181] Gail D Zeevalk and William J Nicklas. Lactate prevents the alterations in tissue amino acids, decline in atp, and cell damage due to aglycemia in retina. *Journal of neurochemistry*, 75(3):1027–1034, 2000.

[182] Andrei O Chertov, Lars Holzhausen, Iok Teng Kuok, Drew Couron, Ed Parker, Jonathan D Linton, Martin Sadilek, Ian R Sweet, and James B Hurley. Roles of glucose in photoreceptor survival. *Journal of Biological Chemistry*, 286(40):34700–34711, 2011.

[183] Salvatore Botta, Elena Marrocco, Nicola de Prisco, Fabiola Curion, Mario Renda, Martina Sofia, Mariangela Lupo, Annamaria Carissimo, Maria Laura Bacci, Carlo Gesualdo, et al. Rhodopsin targeted transcriptional silencing by dna-binding. *eLife*, 5:e12242, 2016.

[184] Monica Doria, Antonella Ferrara, and Alberto Auricchio. Aav2/8 vectors purified from culture medium with a simple and rapid protocol transduce murine liver, muscle, and retina efficiently. *Human gene therapy methods*, 24(6):392–398, 2013.

[185] Tiansen Li, Wendy K Snyder, Jane E Olsson, and Thaddeus P Dryja. Transgenic mice carrying the dominant rhodopsin mutation p347s: evidence for defective vectorial transport of rhodopsin to the outer segments. *Proceedings of the National Academy of Sciences*, 93(24):14176–14181, 1996.

[186] C Mussolino, M Della Corte, S Rossi, F Viola, U Di Vicino, E Marrocco, S Neglia, M Doria, F Testa, R Giovannoni, et al. Aav-mediated photoreceptor transduction of the pig cone-enriched retina. *Gene therapy*, 18(7):637–645, 2011.

[187] F Krueger. Trim galore!: A wrapper tool around cutadapt and fastqc to consistently apply quality and adapter trimming to fastq files, 2015.