Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN

SCIENZE E TECNOLOGIE DELLA SALUTE

Ciclo 34

Settore Concorsuale: 06/D6 - NEUROLOGIA

Settore Scientifico Disciplinare: MED/26 - NEUROLOGIA

INTEGRATING OMICS IN PRION DISEASES AS A MODEL TO EXPLORE THE STRAIN PARADIGM IN NEURODEGENERATIVE DISEASES.

Presentata da: Martina Tarozzi

Coordinatore Dottorato

Marco Viceconti

Supervisore

Sabina Capellari

Co-supervisore

Gastone Castellani

Esame finale anno 2022

Contents

ABSTRACT

This project aims at deepening the understanding of the molecular basis of the phenotypic heterogeneity of prion diseases. Prion diseases represent the first and clearest example of “protein misfolding diseases”, that are all the neurodegenerative diseases caused by the accumulation of misfolded proteins in the central nervous system. In the field of protein misfolding diseases, the term “strain” describes the heterogeneity observed among the same disease in the clinical and pathologic progression, biochemical features of the aggregated protein, conformational memory and pattern of lesions. In this work, the two most common strains of Creutzfeldt-Jakob Disease (CJD), named MM1 and VV2 were analyzed. This thesis investigates the strain paradigm with the production of new multi omic data, and, on such data, appropriate computational analysis combining bioinformatics, data science and statistical approaches was performed. In this work, genomic and transcriptomic profiling allowed an improved characterization of the molecular features of the two most common strains of CJD, identifying multiple possible genetic contributors to the disease and finding several shared impaired pathways between the VV2 strain and Parkinson Disease. On the epigenomic level, the tridimensional folding of the chromatin in peripheral immune cells of CJD patients at onset and healthy controls was investigated with Hi-C. While being the first application of this very advanced technology in prion diseases and one of the first in general in neurobiology, this work found a significant and diffuse loss of genomic interactions in immune cells of CJD patients at disease onset, particularly in the *PRNP* locus, suggesting a possible impairment of chromatin conformation in the disease. The results of this project represent a novelty in the state of the art in this field, both from a biomedical and technological point of view.

INTRODUCTION

CHAPTER 1: SEQUENCING TECHNOLOGIES

Since the discovery of the molecular structure of DNA in 1953 by Watson and Crick[1], determining the order of nucleic acids in polynucleotide molecules and its functional meaning ultimately represented one of the major goals of biological research to decrypt the properties of life on Earth. Only twenty-four years later, in 1977, the first method for DNA sequencing was published by Fredrick Sanger and colleagues, who developed the “chain-termination" or dideoxy technique[2]. This method used chemical analogues of the deoxyribonucleotides (dNTPs), the Dideoxynucleotides (ddNTPs), which have the important feature of lacking the 3′ hydroxyl group that is necessary for extension of DNA chains, thus avoiding the creation of a bond with the 5′ phosphate of the next dNTP. Using radio-labelled ddNTPs in an appropriate ratio (from 10:1 to 300:1, depending on the desired read length) together with standard dNTPs in a DNA extension reaction, DNA strands of each possible length are generated, as ddNTPs get randomly incorporated in the extending strand, preventing further extension. Performing four parallel reactions containing each individual ddNTP base, and running the results on four lanes of a polyacrylamide gel, autoradiography used to be used to infer at each site which chain terminator was incorporated and thus what the nucleotide sequence in the original template was[2]. Improvement of this groundbreaking invention represented the first generation of DNA sequencing technologies, which managed to produce reads nearly of one kilobase (kb) in length[3]. The following development of other molecular biology techniques such as polymerase chain reaction (PCR)[4] in 1983 and recombinant DNA technologies[5] provided the means for generating the high quantities of DNA required by first generation technologies and triggering the genomic revolution, ultimately leading to completing the first draft of the human genome in 2001[6].

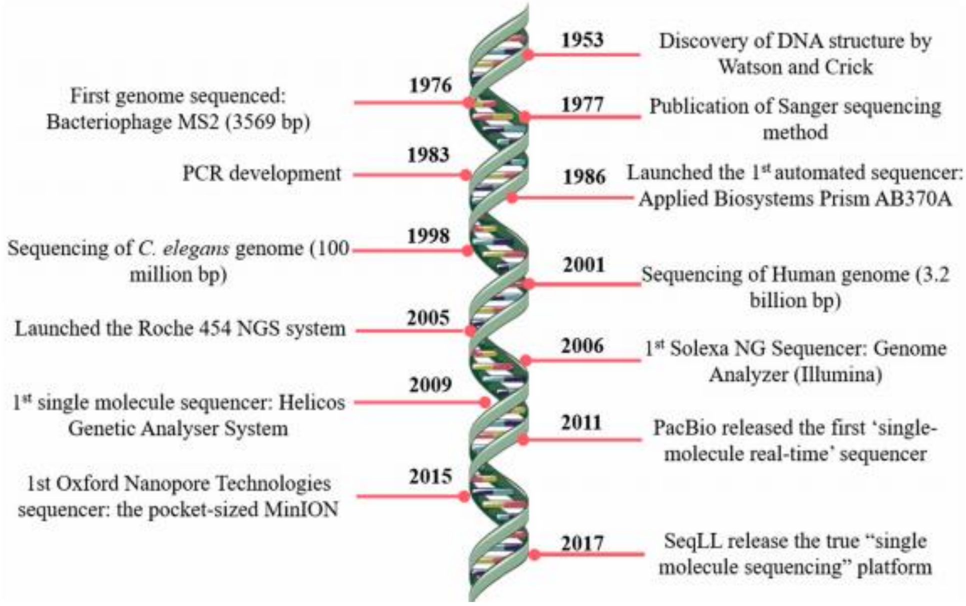
1953
Discovery of DNA structure by Watson and Crick
1976
First genome sequenced: Bacteriophage MS2 (3569 bp)
1977
Publication of Sanger sequencing method
1983
PCR development
1986
Launched the 1st automated sequencer: Applied Biosystems Prism AB370A
1998
Sequencing of C. elegans genome (100 million bp)
2001
Sequencing of Human genome (3.2 billion bp)
2005
Launched the Roche 454 NGS system
2006
1st Solexa NG Sequencer: Genome Analyzer (Illumina)
2009
1st single molecule sequencer: Helicos Genetic Analyser System
2011
PacBio released the first 'single-molecule real-time' sequencer
2015
1st Oxford Nanopore Technologies sequencer: the pocket-sized MinION
2017
SeqLL release the true "single molecule sequencing" platform

Figure 1: Main milestones of the technological improvement in DNA sequencing. Picture taken from Pereira et Al., 2020[7].

The technological development of sequencing methods moved at a tremendous rate, passing from sequencing only short oligonucleotides to DNA molecules of millions of bases, from struggling to acquire the coding sequence of a single gene to technologies able to provide rapid and affordable whole genome sequencing (Figure 1). A pivotal turning point was achieved in 2005, with the birth of Next Generation Sequencing (NGS) technologies[3,7] (sometimes referred to as second generation sequencing), that by means of different reactions depending on the specific technology, allow massive and parallel sequencing even of whole genomes. In the last decade, sequencing technologies expanded also to methods for RNA sequencing[8,9] -giving birth to the transcriptomic field- and to methods for unveiling the structural features and environment-mediated modifications of chromatin and DNA[10,11] - representing the epigenomic field-, as well as all the single-cell omics technologies starting from 2009 [12]. In this chapter, the most important features of NGS technologies in the genomic, transcriptomic and epigenomic fields will be presented, focusing on the three methods used in this thesis: DNA-targeted sequencing, RNA-sequencing, Hi-C.

- *GENOMICS: DNA LIBRARIES AND ILLUMINA SEQUENCING*

Next Generation Sequencing refers to modern high throughput sequencing technologies that can be applied to DNA or RNA. Given the complexity of the topic, today this term refers to technologies that can use different sequencing reactions, classified in two main categories as sequencing by ligation (SBL) and sequencing by synthesis (SBS). SBS approaches are the most common and, despite of their differences, share the feature of depending on DNA-polymerase reactions. In this thesis will be presented in detail only Illumina sequencing, since this is the technology used in this work. Illumina sequencing by synthesis technology is inspired by improvements of Sanger sequencing, in which radio labelled di-deoxynucleotides chain terminators are replaced by reversible terminators combined with fluorophores. The workflow starts with library preparation that can vary significantly in different protocols, here we consider a standard workflow for a targeted sequencing library preparation with hybridization capture-based enrichment, summarized in figure 2. Genomic DNA is always fragmented either mechanically, enzymatically or with transposons in fragments of appropriate length depending on the experimental design, typically around 400bp. Next, blunt ends at both extremities are repaired: usually 5’ ends are phosphorylated and 3’ ends are repaired with Adenine residues. Subsequently, with a PCR reaction, adapters are ligated to the extremities. Adapters are crucial short oligonucleotides that contain a platform specific binding sequence that mediate the binding to the flow cell on which the sequencing reaction will take place and a unique index sequence that allows for the identification of each sample. Depending on the experimental design, it is possible to enrich for a selection of regions that are of interest (like for the exome, that account for ~2% of the genome or for a more restricted selection of genes) or maintain the whole genome. In exome or targeted sequencing, regions of interest are hybridized with biotinylated probes complementary to the target, and subsequently probes are captured using a biotin-streptavidin pull down. Enriched libraries are finally amplified with PCR and quantified. Before being loaded on the sequencer, libraries need to be denatured and drastically diluted to a final concentration around 0.2-0.8 picoM that can vary slightly depending on the sequencer and the specific library.

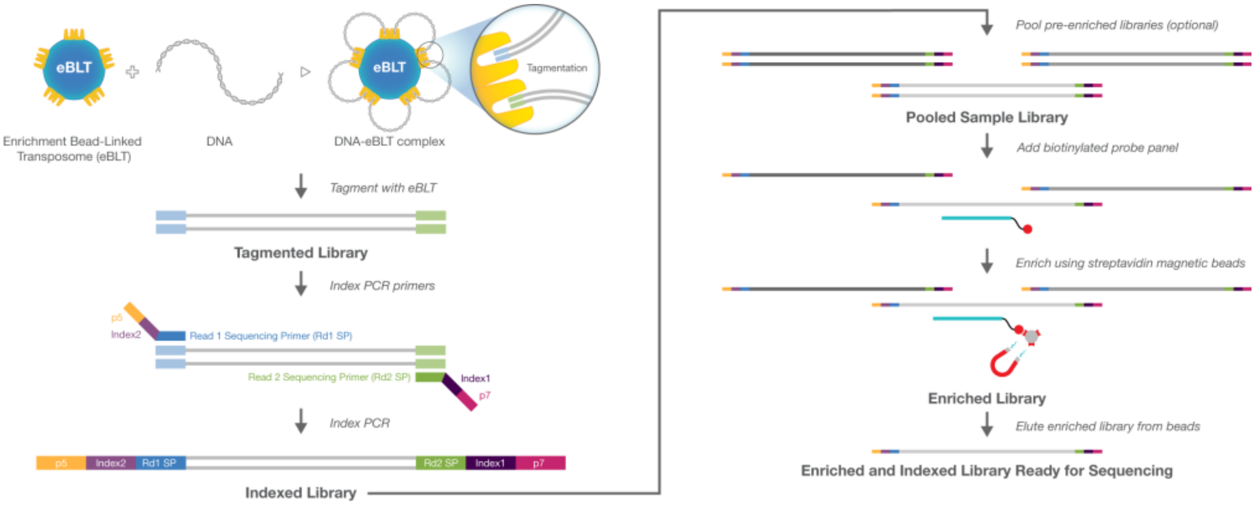
eBLT
eBLT
Tagmentation
Enrichment Bead-Linked Transposome (eBLT)
DNA
DNA-eBLT complex
Tagment with eBLT
Tagmented Library
Index PCR primers
p5
Index2
Read 1 Sequencing Primer (Rd1 SP)
Read 2 Sequencing Primer (Rd2 SP)
Index1
p7
Index PCR
p5
Index2
Rd1 SP
Rd2 SP
Index1
p7
Indexed Library
Pool pre-enriched libraries (optional)
Pooled Sample Library
Add biotinylated probe panel
Enrich using streptavidin magnetic beads
Enriched Library
Elute enriched library from beads
Enriched and Indexed Library Ready for Sequencing

Figure 2: Overview of a workflow for library preparation of DNA samples with enrichment for Illumina platforms, summarizing the main steps of fragmentation, end repair, adapter ligation and enrichment. The picture is taken from the reference manual “Illumina DNA prep with enrichment” which is the protocol used in this thesis for acquiring genomic data (see materials and methods), nevertheless the basic principles are shared among other protocol of the same type of libraries preparation.

Once the sequencer is loaded, all further reactions are automatized: libraries are pumped on the flow cell, a glass device divided in lanes on which are harbored adapter oligonucleotides complementary to those attached to each DNA fragment. Each single stranded DNA fragment hybridize to the flow-cell adapters on one extremity and subsequently to the second too, forming a “bridge” structure. Next follows the bridge amplification step, in which a complementary strand is synthetized based on the nucleotide sequence of the sample. The double stranded fragments are then denatured, leading to single strand fragments anchored to the flow cell only by one adapter. This process of bridge amplification is repeated several times, generating thousands of copies of each initial fragment and millions of clusters anchored to each lane of the flow cell. After cluster generation, sequencing by synthesis takes place. Proprietary nucleotides, modified in the 3′ hydroxyl group to carry a reversible chain terminator and a fluorophore corresponding to each of the four nucleotides, are incorporated in the growing complementary strand by DNA polymerase and detected by a camera. The nucleotides, acting as reversible chain terminators, block the synthesis of the strand until after the detection of the fluorescent label is acquired, and then a next round of synthesis starts. The fluorophore wavelength together with its intensity determines the base call. All clusters are sequenced simultaneously in parallel reactions and each genomic region is sequenced in several different fragments, quantified by the “coverage” metric. The reactions are repeated for 150 rounds (75 rounds in case of short reads protocols, like RNAseq) for read 1. Indexes are then sequenced and, in paired-end sequencing, a further single bridge amplification is performed. The template fragment is washed away, and another read of the nucleotide sequence from the other extremity of the fragment is acquired, generating read 2. The acquired signals for each lane of the flow cell are converted into base call format files (*bcl* file) that represent the output of sequencing and the first raw input for bioinformatic analysis (figure 3).

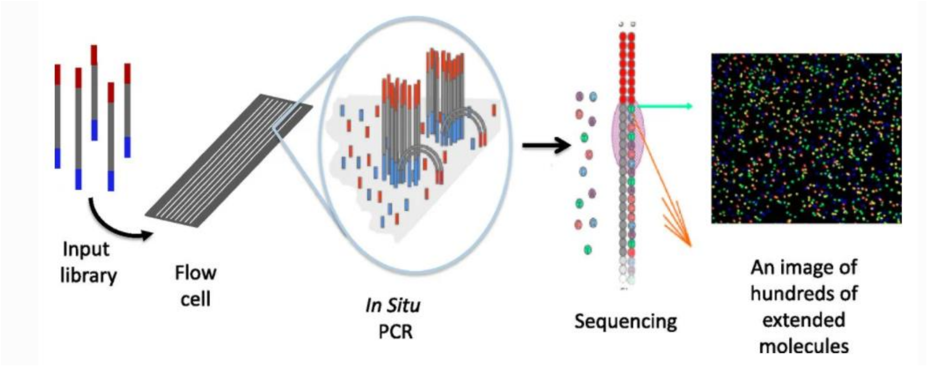

Figure 3: Overview of sequencing reaction on Illumina platforms. Denatured library is first bound by means of adapters to the flow cell, where it is amplified through bridge amplification creating millions of

clusters. Single strands fragments are then sequenced with a sequencing by synthesis mechanism, in which modified reversible chain terminator nucleotides are added once per round after detecting the characteristic signal for each base incorporated. The acquired images are converted into a *bcl* file that represent the raw output of sequencing and the first input for bioinformatic analysis.

Despite library preparation vary depending on the experimental design and on the nucleic acid used as input, the sequencing workflow for Illumina platforms follows the presented steps independently of the library, therefore this applies to all the omics reported in this work, namely DNA target sequencing, RNAseq and Hi-C.

- ### *TRANSCRIPTOMIC: RNA SEQUENCING LIBRARIES*

The information encoded in DNA is expressed through transcription into coding and non-coding RNA molecules, the transcriptome. The transcriptome of the same individual changes in different tissues and during different stages of life since different cells express different mRNAs and regulatory RNAs. Transcriptomic technologies are the techniques used to acquire a snapshot of the RNAs expressed at a given time in a tissue/cell type, and to study how the expression profile changes in different conditions. Such technologies have been and still are an extremely powerful tool to understand gene functions, cell differentiation and development or pathways affected in a given condition. The two most relevant technologies in this field are microarray and RNAseq, the first nowadays quite obsolete but extremely important in the first stages of this field during the decade 2000-2010[13]. RNA sequencing is a very versatile high throughput sequencing technique introduced in 2008[14–17], that allows in a single experiment to investigate not only gene expression, but also alternative splicing[18], allele specific expression[19], variation in linear nucleotide sequence[20,21], novel transcript expression[22] and gene fusion events[23]. The appropriate library preparation protocol for RNA sequencing needs to be carefully chosen depending on the study, since important bias can be introduced at the library preparation step. There are four main categories of possible library preparations that answer different question (Figure 4):

- **total RNAseq**: all structural, regulatory, coding and non-coding RNA are sequenced.
- **RNAseq with ribosomal RNA reduction**: only rRNAs useful for phylogenic reconstruction are kept together with regulatory and coding RNAs.
- **cDNA capture**: only coding RNAs are enriched using probes targeting exon sequences. It is possible though to detect some non-coding RNAs similar enough to the probe to be captured.
- **polyA selection**: only mature mRNAs are isolated through poly-T probes that bind the 3’ poly-A tail of mRNAs.

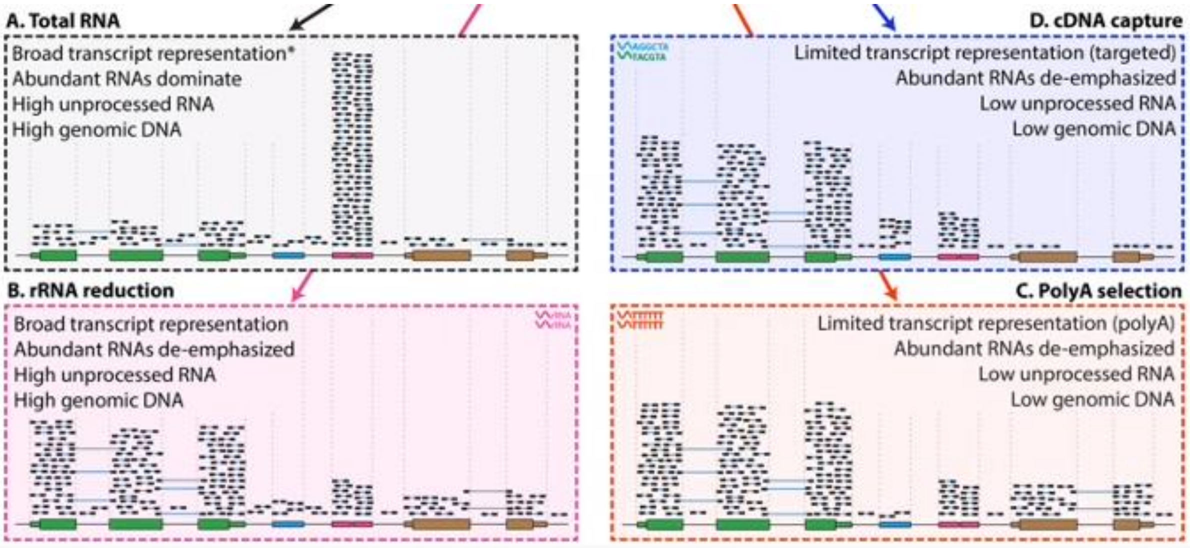
A. Total RNA
Broad transcript representation*
Abundant RNAs dominate
High unprocessed RNA
High genomic DNA
D. cDNA capture
Limited transcript representation (targeted)
Abundant RNAs de-emphasized
Low unprocessed RNA
Low genomic DNA
B. rRNA reduction
Broad transcript representation
Abundant RNAs de-emphasized
High unprocessed RNA
High genomic DNA
C. PolyA selection
Limited transcript representation (polyA)
Abundant RNAs de-emphasized
Low unprocessed RNA
Low genomic DNA

Figure 4: Overview of different approaches for RNAseq library preparation. Picture taken from the course material of the workshop “Informatics for RNA-seq analysis” provided by the Canadian Bioinformatics Workshops (https://bioinformatics.ca/workshops/2016-informatics-rna-seq-analysis)

In this work, a cDNA enrichment protocol based on hybridization with probes targeting exon was used. This approach is particularly suited for RNAs extracted from degraded samples (such as postmortem tissues) that show a significant level of degradation at the moment of total RNA extraction[24,25]. In this case thus, it was not adequate to use a poly-T enrichment for the isolation of mRNA, since it would lead to a strong overrepresentation of the 3’ extremities due to the experimental procedure, while representation of 5’ sequences would be extremely poor. During library preparation, total RNA is fragmented and retrotranscribed into cDNA with random primers. Next, adapters are ligated to the fragment’s extremities and subsequently the enrichment of coding regions is performed with biotinylated exons probes, similarly to target sequencing, described in the previous paragraph. After an amplification step with PCR, in Illumina platforms libraries are sequenced similarly to DNA libraries, with the only difference of having a paired end sequencing of 75bp per reads instead of 150bp. The sequencing depth, expressed in millions of reads per sample, is a very important feature to consider at the experimental design step. Depending on the amount of reads per sample, it will be possible to acquire information with variable degree of confidence about non-common transcripts, which will be adequately represented with sequencing depths around 25 million reads/sample. Given the numerous information that can be extracted from this type of data, the bioinformatic analysis is particularly challenging and heterogeneous. This very important part of transcriptomics is treated extensively in “Chapter 2: bioinformatics and omics data analysis”.

- ### *EPIGENOMICS: STRUCTURAL GENOMICS AND HI-C*

The eukaryotic genome is both very dynamic and highly condensed in the nucleus: in humans, nearly 6 billion nucleotides (that correspond of 2 meters in length in its linear form) are organized in 46 chromosomes packed within a nucleus of 6µm of diameter on average.[26] This high level of compactness must anyway satisfy functional requirements to allow gene expression and transcription regulation. How chromatin is structurally and functionally organized in the nucleus is therefore extremely important, since it determines changes in gene expression that happen without altering the linear DNA sequence, that is indeed what Epigenetics deals with[10,27]. Gene expression in mammals strongly depends on epigenetic regulation, in terms of modification of the accessibility of genomic regions trough DNA methylation, histone modifications and three-dimensional chromatin organization, and of post-transcriptional regulation mechanisms such as noncoding RNA–mediated regulation and RNA editing. These modifications respond to environmental stimuli and regulate major processes such as cell development and differentiation, cell cycle regulation, and ultimately the responses in health and disease[28–31].

Already in the first half of the 20th century the initial observation of a genome divided in euchromatin and heterochromatin and the hypothesis of chromosome territories in the nucleus was proposed using light microscopy and chromatin dyes [30,32,33], and it was validated in 1980s after the development of more advanced imaging techniques, most remarkably by fluorescence in situ hybridization (FISH)[34]. In 1974 Kornberg published the first observation of the chromatin structure[35]: in the nucleus, DNA is first packed into nucleosomes, composed of 147 bp of DNA coiled around a histone octamer containing two copies of four different histone proteins, H2A, H2B, H3, and H4[36]. At this level of organization several modifications may happen: histone proteins can harbor a plethora of modifications that mediate transcriptional regulation trough methylation, acetylation, phosphorylation, ubiquitination and sumoylation, allowing the epigenetic regulation of most biological processes including DNA repair, transcription, and chromatin remodeling. Non-histone modifications are also possible, like shift of nucleosome position which changes the DNA region accessible to the transcription apparatus, or methylation of DNA itself in C5 of Cytosine residues, that is associated to reduced gene expression as a consequence of the recruiting of transcriptional repression proteins and of reduced accessibility to transcription factor of promoter regions[37–39]. Both histones and DNA modifications are extremely important and thoroughly studied epigenetic regulators of cell differentiation, cell cycle and response to environmental stress and disease, nevertheless they will not be further discussed here since the experimental work of this thesis deals with the three-dimensional organization of chromatin assessed with Hi-C experiments.

The structural organization of the genome can be studied at different levels of resolution, depending on the chromosome conformation technique and for Hi-C depending also on the sequencing depth. At the highest scale, the nucleus is functionally compartmentalized[40,41], with regions enriched for euchromatin and gene-rich regions localized at the center of the nucleus (compartment A), whereas heterochromatin- rich regions are found near the nuclear margins (compartment B)[11,42–45]. Lamina-associated domains (LADs) refer to genomic regions in contact with the nuclear lamina, these domains show typical features of heterochromatin, such as low gene density, reduced gene expression, and enrichment for histone marks associated to reduce transcription rates. Compartment A contains predominantly gene-rich regions, is enriched in H3K36me3, marked of open chromatin, and for DNase I hypersensitive sites[46]. In contrast, compartment B represents typically heterochromatic regions and overlaps with the lamina-associated domains (LADs)[47]. Chromosome compartments can be further separated into Topologically associating domains (TADs)[48]. TADs represent the basic unit of the chromatin three-dimensional organization[11,44,49,50], they are delimited by convergent binding sites for CCCTC-binding factor (CTCF) and Cohesin and show high degree of internal self-interactions. Differently from other structural features like chromosome compartments, TADs are relatively stable in different cell types and seem to not be affected by tissue-specific gene expression or histone modifications. TADs are furtherly organized in chromatin nanodomains (CNDs) and chromatin loops that represent loci of stable intra- interactions within TADs [51–53] (Figure 5).

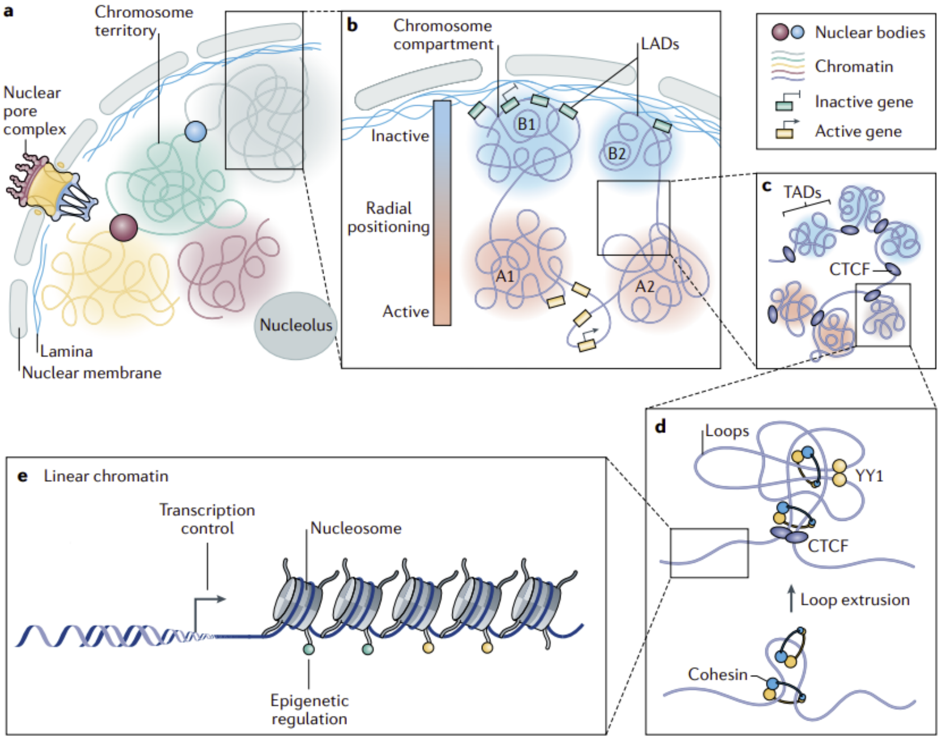

Figure 5: Levels of 3D organization of Eukaryotic genomes inside the nucleus. At the lowest resolution (panel 'a'), chromosome territories are visible: chromosome occupy a well-defined territory where intra-chromosomal interaction are more frequent than inter-chromosomal interactions. Zooming inside a single chromosome territory (panel 'b'), compartments A and B can be identified based on the radial positioning, chromatin compactness and the transcriptional activity. Inside a single compartment (panel 'c'), Topologically Associating Domains (TADs) are delimited by convergent binding sites for CCCTC-binding factor (CTCF) and Cohesin, representing the basic unit of the chromatin 3D organization and showing a high degree of internal self-interactions. Zooming inside a single TAD (panel 'd'), at the highest level of resolution achievable with Hi-C, we find single chromatin loops, that represent loci of stable intra-interactions. Panel 'e' shows DNA structure in the 30-2nm range, where DNA is wrapped around an octamer of histones, and both DNA and histones can harbor modifications able to modify transcriptional activity of that locus. Picture taken from Wang et Al., 2021, Nature Reviews.

Despite all these level of organization have been observed in different cell lines and species, the mechanisms that regulate chromatin folding are just beginning to be understood[54–56]. Currently, Loop Extrusion is the accepted model for chromatin folding at the Mb scale: in interphase, CTCF and Cohesin complexes reach their binding sites and begin to extrude a loop symmetrically until convergently oriented CTCF motifs are recognized. Here the protein complex stops the extrusion, resulting in a DNA loop with cohesin and convergently oriented CTCF present at its base, physically demarcating TADs borders (Figure 6).

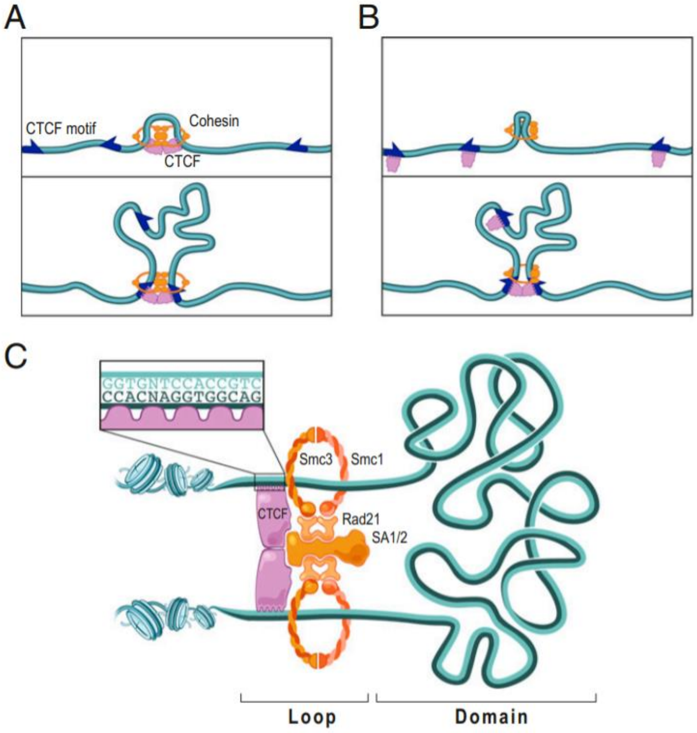

Figure 6: Illustration of two possible models of TAD formation according to the loop extrusion model. CTCF protein could slide together with cohesin until convergent CTCF motifs are identified and the loop is stabilized, or be bound to the CTCF binding motif and assemble to stabilize the loop when two motifs with correct orientation meet. Illustration taken from Sanborn et al. (PNAS; 2015)[57]

Inside TADs, enhancer-promoter interactions are facilitated and stabilized: disruption of TADs border leads to the loss of such crucial regulatory elements and is associated to impairment of cell fate commitment and to several human diseases, both with inherited [58,59] and acquired origin[60–62]. The study of 3D genomics is a very promising field, which provides also new insight for the functional interpretation for the enormous amount of intronic variants associated to complex diseases with Genome-Wide Association Studies (GWAS). Many of such variants in fact are in noncoding regions or near cis-regulatory elements, and for most of them the target gene is unknown. Frequently, it is assumed that SNPs associated to a disease located in introns of a gene related to affected pathways must act on that gene or in the closest plausible, but there are many examples of how misleading this assumption could be [63–66]. In fact enhancers frequently regulate gene expression of distal genomic loci, both backwards and forwards in the linear sequence[63,66], interacting through chromatin loops. Improving the understanding of the three-dimensional folding of the genome will help to discover the complex regulation networks of gene expression in physiological conditions and how they are altered in disease.

From a technological point of view, a turning point in chromatin biology was the development in 2002 of chromosome conformation capture (3C) technique that represented the first genomic method for the investigation of chromosome conformation in-nuclei[67], then implemented in 3C-derived methods such as 4C, 5C and Hi-C[42,68–71]. All these techniques provide information on the three-dimensional conformation of genomes by relying on the ligation of distal genomic fragments that have been crosslinked together due to their spatial proximity in the nucleus. 3C-derived methods differ in the way the interactions are detected and quantified. Hi-C is so far the most powerful conformation capture technique since it is an high-throughput method that allows to detect genome wide interactions between genomic loci (all-to-all)[42,50,72,73] whereas the original 3C method only allows to detect if a specific region of interest interacts with another (one to one)[67], typically to identify specific enhancer-promoter interactions. The results of Hi-C experiments are the pairwise frequencies of interactions of distinct genomic elements in the same sequencing read, which reflect their average spatial distances in the nuclei of the analyzed cell population, thus providing information on the 3D conformation of the genome[11,67,74]. The standard Hi-C workflow is summarized in figure 7 and starts with a formaldehyde crosslinking of the chromatin on a high number of cells. Formaldehyde permeates cell membranes and forms covalent bonds between

DNA, proteins and other reactive molecules in close proximity[42], stabilizing the three-dimensional organization of DNA in the nucleus. DNA is then digested with restriction enzymes and biotinylated nucleotides are incorporated into the restriction fragment ends. Next, restriction fragments are ligated, so that two fragments which are crosslinked together because close one to another in the 3D space are joined by the ligase enzyme, even though they may be kilobases away in the genome.

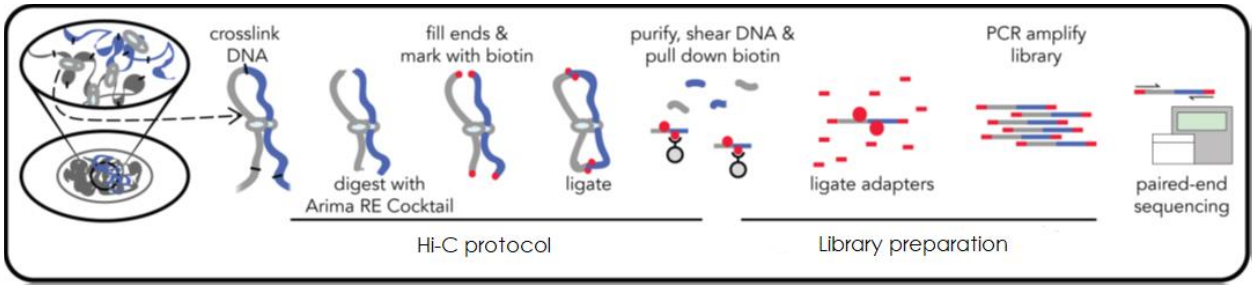

Figure 7: Standard Hi-C workflow from cells to library preparation and sequencing. Adapted from the reference manual of the commercial product "Mammalian cell lines kit-ARIMA technologies".

The biotinylated ligation products are then pulled down with streptavidin beads and with this input, DNA libraries are generated for whole genome sequencing, with a workflow similar to the one described in the *Genomics* paragraph. Sequencing data, after the appropriate bioinformatic analysis, will provide the relationship between interaction frequency and spatial distance in all chromosomes, thus allowing for the computational modelling of the 3D organization of chromatin. Computational aspects of the analysis of this type of data are presented in "Chapter 2: bioinformatics and analysis omic data".

CHAPTER 2: BIOINFORMATICS AND -OMICS DATA ANALYSIS

- ## *PRIMARY DATA ANALYSIS AND BIOINFORMATIC PIPELINES*

Due to the increasing affordability, popularity and optimization of sequencing experiments in the last decade, the amount of sequencing data produced exploded[75]: the bottleneck is not the production of sequencing data anymore, but rather the analysis and interpretation of such complex and high-dimensional data. Bioinformatic analysis involve several steps of computationally intensive data transformations, each requiring specific a tool. A lot of effort has been put in developing reliable software to perform such data transformations and to make computation analysis of NGS data reproducible. Package and environment management system such as CONDA allow the installation and management of software developed in different languages for different operating systems. Through an organization into channels where packages are collected and maintained, it allows an efficient distribution of tools for different fields of data science. Bioconda[76] is arguably the most common CONDA channel for bioinformatic tools. Primary analysis for Illumina raw data consists in demultiplexing *bcl* data into FASTQ files. This process matches the index sequences of each sequenced fragment to group all calls belonging to a sample into a text file in which the nucleotide sequence is accompanied by a PHRED score defining the quality of the nucleotide call. Quality checks follow this conversion, and after, adapter reads and base calls with low quality are trimmed from the sequence that will undergo the proper analysis. Workflow management systems such as Snakemake[77] address the need to concatenate the heterogenous steps required by bioinformatic pipelines while providing an efficient usage of computational power through process parallelization[78]. Snakemake pipelines are contained in ‘Snakefile’ files which are written in a domain specific language that uses a syntax very similar to Python, in which each step is expressed as a rule having a specific input, output, environment and by the proper command which can be coded in Shell, Python, R, or other languages according to the user needs. Concatenating the output of a rule to the input of the following, it allows to create automatized workflows. With the use of wildcards identifying samples names, it allows for the parallel analysis of multiple samples at once[77]. In this work, a customized pipeline for each omics data type was written to perform appropriate secondary analysis.

- *TARGETED SEQUENCING DATA ANALYSIS*

Secondary analysis of genomic data is the most mature and standardized workflow within the considered data types of this thesis, already used in diagnostics as well as in research. It generally comprehends three main steps: mapping on the reference genome, post-alignment processing and variant calling. A standard workflow of secondary analysis of DNA data is reported in figure 8.

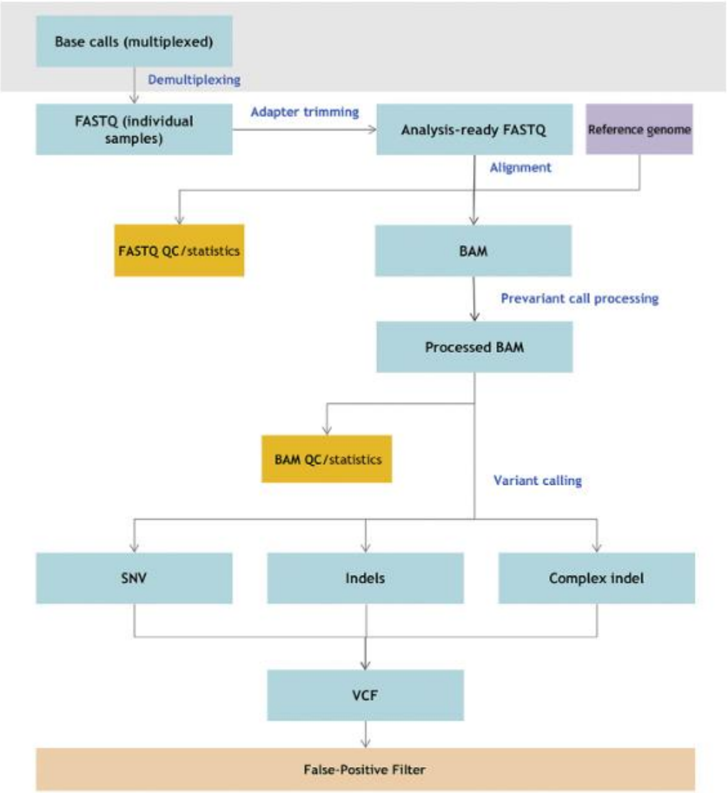

Figure 8: Standard workflow of secondary analysis of next generation sequencing genomic data, consisting in demultiplexing, trimming, mapping, post-alignment processing and variant calling. Picture taken from Roy et al., 2018[79].

After demultiplexing, sample's sequence is expressed in two FASTQ files containing nucleotide sequence and quality score derived from millions of short reads. Mapping algorithms aim at identifying a location in the reference genome that matches the experimental read, tolerating some mismatches and extra spaces to allow for variants detection. The most recent and accurate mapping algorithms are based on the Burrows-Wheeler Transform (BWT)[80], that is particularly appropriate for the compression and alignment of lowly divergent strings. One of the most used tools for mapping NGS sequences to the reference genome is indeed the Burrows-Wheeler Alignment[81], used with *mem* algorithm that provides high speed and accuracy of mapping. The output of mapping algorithms is *sam* files (sequence alignment map), which store all the information about the mapping procedure that generated them in the metadata section, together with information about the mapped genomic region and the accuracy of the mapping. The binary counterpart of *sam* files is *bam* files, that represent the type of data that will undergo the further post alignment processing and variant calling. Post alignment processing consists in sorting, marking duplicates (optical duplicates and PCR artifacts) and indexing the *bam* file. At this point, coverage, that is the number of reads that "cover" a specific genomic region and the average number of reads for the whole target, can be computed on the marked *bam* file. Next follows the variant calling step, aimed at the identification of single nucleotide variants and small insertions and deletions. Several tools with different assumptions and features are available, the most appropriate caller should be chosen depending on how the data were produced and on the experimental design. In this work, germline variants were called with Strelka2[20], a variant caller optimized for analysis of germline SNV and indel in small cohorts. Its outputs are Variant Call Format (VCF) files, which store in the metadata section information about tools, time of the analysis, support files (such as reference genome version, targeted regions if provided, assembly, etc.) used in the analysis together with the list of variants detected, expressed through the eight columns: chromosome, position, *rs* identifier of the variant (if exist), reference nucleotide, alternative nucleotide, quality, filters, and other info. Secondary analysis of genomic data ends with variant annotation, aimed at acquiring functional information about the type of nucleotide substitution and it predicted effect.

At this point, a classical approach is to consider only variants that affect the primary structure of the coded protein, like missense or truncating variants, especially if predicted as likely pathogenic. In recent years, data science methods applied to genomic data proved to be a resourceful approach for acquiring a more complete understanding of polygenic contributions in complex diseases[82] and in the context of precision medicine[83], resulting in an increase of the use of machine learning methods (ML) in the genomic field[84,85]. The term machine learning refers to several algorithms able to perform and evaluate pattern recognition, classification and prediction tasks based on models derived from existing data. The key feature of these methods is the absence of coded instructions given by the developer to describe the steps towards which input data are transformed in output results. Thus, the algorithm computes a model trained on the input data to address a specific task. There are two main types of ML methods: supervised and unsupervised learning. In supervised learning, objects are classified using a set of attributes, labels or features given by the operator. Examples of such methods are classifiers like decision trees, random forests, support vector machines and neural networks. The result of the classification process is a set of rules that assigns objects to classes based on the provided labels. These rules can display new insights on the relevant features used to correctly identify the studied classes. On the other hand, in unsupervised learning no predefined labels are provided for the objects under study. Here, the goal is to explore the data and discover similarities between objects that will cluster together based solely on the input data. Clustering and dimension reduction techniques such as principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) are suitable unsupervised ML methods for visualizing high dimensional data such as genomic data[86,87]. As the biomedical sector becomes more data-intensive, these approaches are bound to play a pivotal role both in research activity and in precision medicine, as the amount of data generated even by target sequencing panels for each patient is already hardly handled by human operators.

- ### *TRANSCRIPTOMICS: RNA SEQUENCING ANALYSIS*

Even though RNA sequencing is a widely used, mature technology, on the computational side bioinformatic methods are still rapidly evolving and there are no absolute gold standards. Different approaches in terms of mapping strategy, normalization and statistical modelling for differential gene expression exist, the most appropriate should be chosen depending on the experimental design and questions. Here will be introduced only the final workflow setup that was chosen in this work. RNA sequencing data analysis starts with demultiplexing, quality check of FASTQ files and trimming, similarly to DNA sequencing data analysis. Mapping RNA sequencing can be done using as reference either genome or the annotated transcriptome. Depending on the reference of choice, the number of multi-mapped reads will change, and the quality parameters cut-offs should change accordingly. After mapping, post alignment processing is performed similarly to DNA pipelines. During mapping, reads are quantified, and raw read counts will represent the input for differential gene expression (DGE) analysis. Raw reads quantifications must be normalized for technical factors that influence the number of detected reads per genomic region (such as sequencing depth in different samples, gene length, RNA and cell type composition) to compare different samples. DeSeq2[88] is one of the most used tools to perform DGE analysis, its normalization implies the following steps. At first, a pseudo-reference sample derived from rowwise geometric means of all biological replicates is created. Then, for each gene in a sample and for every sample, the ratios to the pseudo-reference are calculated. For each sample, the median value of all ratios is taken as the scaling factor. Raw counts are then divided by the sample's scaling factor to normalize raw counts. On normalized data, sample level quality checks must be performed to evaluate overall similarity between samples, to identify possible outliers and experimental biases. Principal Component Analysis (PCA) and hierarchical clustering are unsupervised methods commonly used to explore how well replicates cluster together and to identify which other parameters (in addition to the object of the comparative analysis) represent a major source of variation in the dataset, that therefore should be used as covariates in the model. The identification of differentially expressed genes (DEGs) implies determining which genes have a significantly different mean expression between groups of interest, given the intra-group variation of biological replicates of the same group. This intra-group variation needs to account for the fact that variance (Var) increases linearly with the mean (μ) expression and proportionally to a coefficient of variation *f*.

$$Var = f\ \mu$$

Therefore, DeSeq2 introduces a new measure of variation to avoid DEGs to be overestimated in lowly expressed genes. Gene-wise dispersion is defined as follows and is computed on normalized read counts and used as a measure of variation in the data.

$$\alpha = \frac{Var - \mu}{\mu^2}$$

Dispersion increases together with variance and decreases with high mean expression values (figure 9). In highly or moderately expressed genes, the square root of dispersion equal the coefficient of variation, $\sqrt{\alpha} = f$. Therefore, the value of dispersion represents the expected variation around the mean across biological replicates. Dispersion values represent the variance for a given mean in gene expression, since the dispersion estimate for genes with the same μ will differ solely based on their variance.

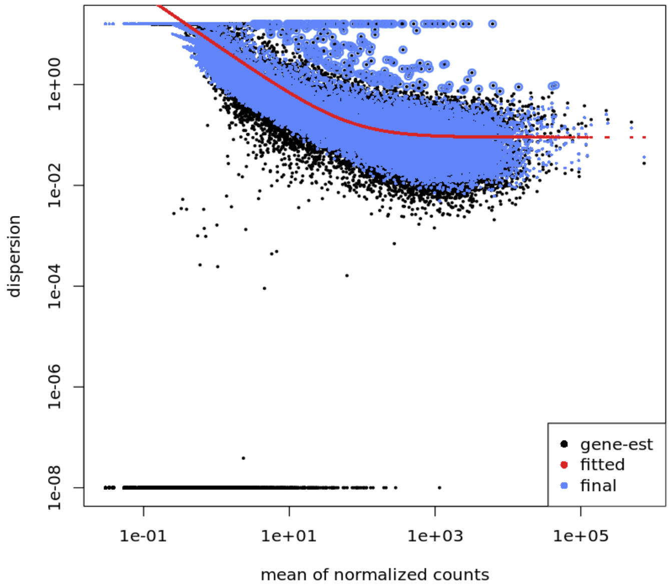

Figure 9: Dispersion plot computed with DeSeq2. Different genes have different of biological variability, but, among the whole transcriptome, there will be a distribution of reasonable estimates of dispersion. The red line represents the estimate for the expected dispersion given an expression rate. Black dots the mean expression level of each gene and blue dots are the maximum likelihood estimation of the dispersion after the shrinking towards the curve.

Once gene-wise dispersion estimates are computed, a curve is fitted to the data to express the expected relationship between dispersion and gene expression. To reduce the false positives in DGE analysis, dispersion values are shrunken towards the curve, with the exception of extremely high values that are not shrunken since those genes could not follow the modelling assumption for biological or technical reasons. DeSeq2 models normalized counts on a negative binomial distribution to compute gene’s model coefficients for log2foldchanges (LFC) for each group of using the following equation:

$$K_{ij} \sim NB(s_{ij}q_{ij}, \alpha_i)$$

K_{ij} = Raw counts for gene i, sample j

$s_{ij}q_{ij}$= Normalized counts scaled by a normalization factor

α_i= Dispersion for gene i

The model should be customized adding the most relevant covariates identified in the sample level QC step, while comparing the classes of interest. Normalized count data are fit to this equation and coefficients (estimates for the LFC) along with their standard error are estimated for each class involved in the comparison. LFC of transcripts with either low counts or high dispersion are likely to be inaccurate, therefore, DeSeq2 performs a shrinkage of the estimated value towards zero, using the distribution of LFC estimates for all genes as a priori knowledge to shrink the LFC estimates of genes with little information toward more likely LFC estimates. Accurate LFC values are crucial since the null hypothesis under which differentially expressed genes are identified is that LCF = 0, that means that no gene expression changes are present between conditions. Hypothesis testing is usually performed either with Wald test or Likelihood Ratio Test.

The result of DGE analysis is a list of all transcripts associated to a log2foldchange and adjusted p value. On this output, functional enrichment analysis can be performed to gain more biological insight about the affected pathways. There are several different but complementary ways to perform functional enrichment analysis. Functional analysis based only on a list of differentially expressed genes is referred to as Over Representation (OR) analysis: in this case only genes below a chosen p value cutoff are use used for querying databases containing information about biological functions, involvement in pathways and physical interactions like Gene Ontology[89] or KEGG[90] to determine if input genes belonging to a functional category are overrepresented with Hypergeometric test. Another way to explore the biological interplay of protein coding transcripts is with Protein-Protein interaction networks (PPIn), that provide information about the physical interactions between proteins coded by differentially expressed genes and the relative biological pathways associated to their physical interactions. A different approach to perform functional analysis is with Functional class scoring (FCS) tools, such as GSEA (Gene Set Enrichment Analysis)[91]. The hypothesis of FCS methods is that small but coordinated changes in functionally related genes can have biological consequences. Thus, rather than setting an arbitrary p value cut-off to identify "significant genes", the log2foldchange of all genes are considered. The gene-level statistics from the dataset are aggregated querying the database of choice to generate pathway-level statistic, providing a p value expressing the probability that by chance the observed enrichment of a given pathway happened. Arguably the main result of GSEA is the enrichment score (ES), which quantifies the degree to which a gene set associated to a given pathway is overrepresented at the top (overexpressed genes) or bottom (under expressed genes) of the experimental ranked list of genes. GSEA calculates the ES by walking down the experimental ordered list of genes and increasing ES score every time there is a match with gene set of the pathway and decreasing it in case of mismatches. The higher the ES, the higher the pathway is affected in the tested phenotype. These three computational methods can be used both singularly and together, to identify altered pathways from different perspectives.

• *EPIGENOMICS: HI-C DATA ANALYSIS*

Hi-C analysis provides the pairwise frequencies of interactions of distinct genomic loci which reflect their spatial proximity in the nuclei of the analyzed cells. To get to these structural features, Hi-C experimental data must be filtered, normalized, and quantified by several steps of bioinformatic analysis. Among the considered technologies, Hi-C is the most recent and consequently the computational analysis of this type of data is still rapidly evolving but several tools are already available[92]. In this work, TADbit[41] was used to perform the analysis of Hi-C interaction datasets covering all the steps from aligning the sequenced reads (paired-end reads) up to interaction matrices. Similarly to the previous data types, the primary analysis starts with demultiplexing of the raw sequencing data to FASTQ files. The main steps of this pipeline are summarized in figure 10 and comprehend quality control of the reads, mapping to the reference genome, filtering of artifactual reads, normalization of experimental biases, generation of binned interaction maps, statistical analysis of the differences and consistencies between experiments, identification and comparison of the structural features in the interaction maps.

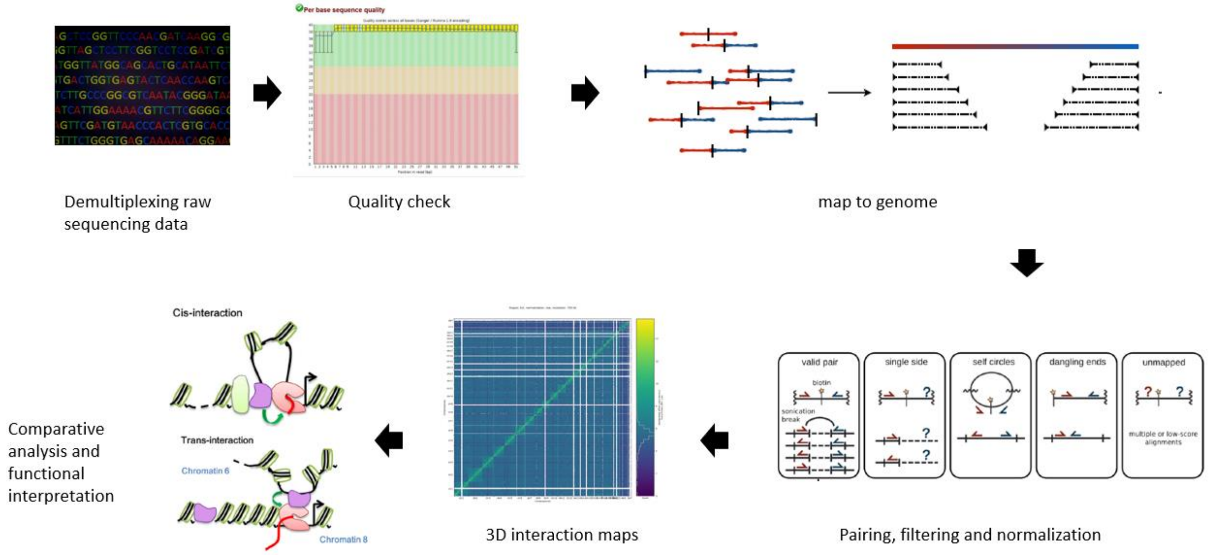
Per base sequence quality
Demultiplexing raw sequencing data
Quality check
map to genome
Cis-interaction
Trans-interaction
Chromatin 6
Chromatin 8
Comparative analysis and functional interpretation
3D interaction maps
valid pair
biotin
sonication break
single side
self circles
dangling ends
unmapped
multiple or low-score alignments
Pairing, filtering and normalization

Figure 10: Schematic representation of a Hi-C data analysis workflow, consisting in demultiplexing, quality controls of input reads, mapping, filtering, normalization and computation of the interaction matrix.

In addition to the standard quality checks on the FASTQ files, here is assessed also the quality of the various steps of the Hi-C experiment, such as the number of undigested and unligated restriction sites, the overall percentage of digested sites, the percentage of digested but non-ligated extremities (dangling-ends) and the percentage of read-ends with a ligation site. Similarly to previous pipelines, the mapping step aims at uniquely aligning the input reads to the reference genome, allowing for chimeric reads resulting from the ligation of distal genomic regions. After mapping, if needed, biological replicates can be merged into a single file representing the group of biological interest. Next, reads that have been uniquely mapped on both ends together with their genomic coordinates and their position relative to the closest RE site are kept. Reads are then filtered, keeping only structurally informative reads, that are fragments in which both read-ends are close to a Restriction Enzyme (RE) site on different fragments and in facing orientation (valid pairs), and represent typically around 36-70% of uniquely mapped reads. Other mapped pairs will be classified in the following categories, represented also in figure 11:

Valid pairs: Read-ends are mapped to different RE fragments in facing orientation, and the ligation event is larger than the longest ligation fragment, defined as the upper limit of the library size distribution.

Dangling-end: both read-ends are mapped to the same RE fragment in opposite orientation (~15%).

Extra dangling-end: Read-ends are mapped to different RE fragments in facing orientation but are close enough from the RE site to be considered part of adjacent RE fragments that were not cut by digestion (~20%).

PCR artifacts or duplicated: Start positions, mapped length, and strands of both read-ends is the same, therefore only one copy is kept (~5%).

Random breaks: The start position of one read-end is too far from the RE cut site. These breaks are likely produced by non-specific enzyme activity or by random breakage of the chromatin (~1%).

Errors: both read-ends are mapped to the same RE fragment in the same orientation (<1%).

Valid pairs will be used to assess the number of interactions between two loci. Due to the polymeric structure of the chromatin, the number of interactions and genomic distance between two interacting sequences is inversely proportional. Interaction matrices are generated by dividing the genome in equally long loci (bins) and assigning each end of the read to its binned genomic location. This process is called binning and will define the resolution of the Hi-C matrix.

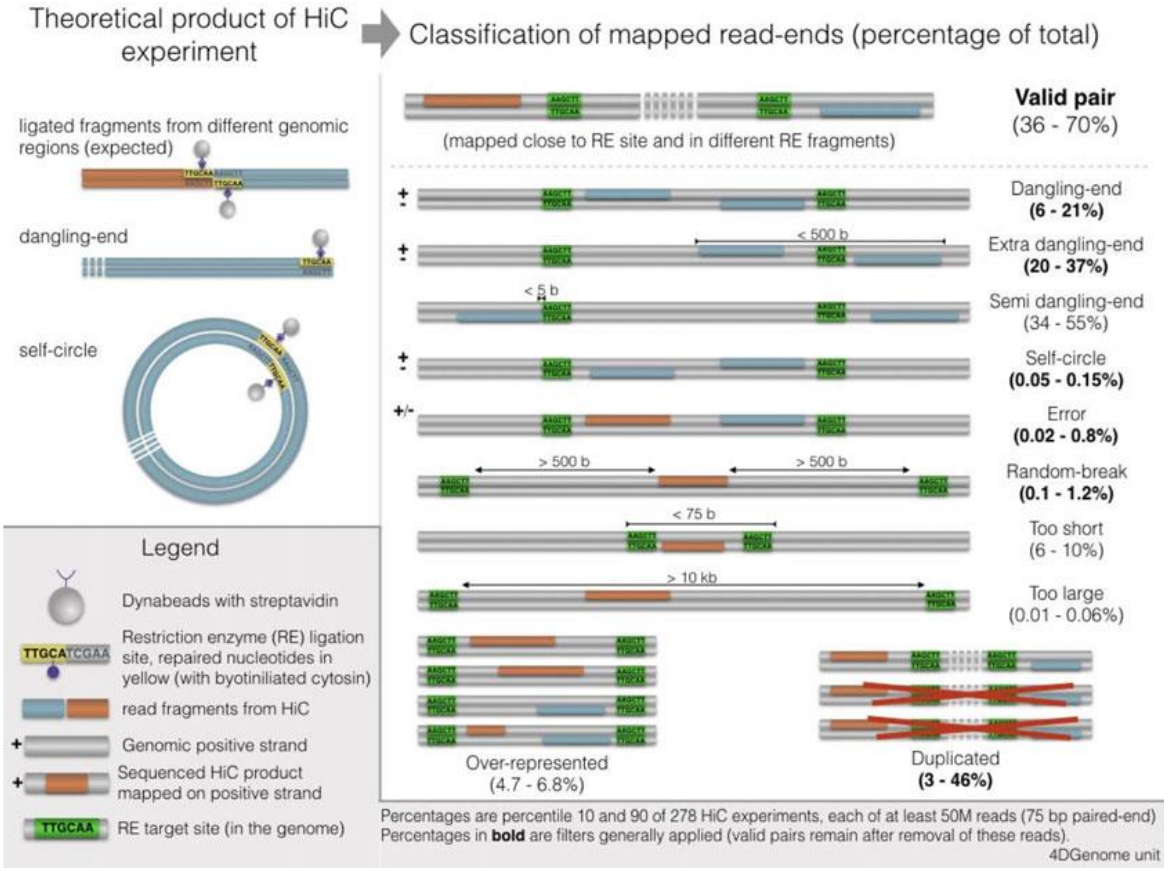
Theoretical product of HiC experiment
Classification of mapped read-ends (percentage of total)
ligated fragments from different genomic regions (expected)
TTGCAA
dangling-end
self-circle
Legend
Dynabeads with streptavidin
TTGCATCGAA
Restriction enzyme (RE) ligation site, repaired nucleotides in yellow (with byotiniliated cytosin)
read fragments from HiC
+
Genomic positive strand
+
Sequenced HiC product mapped on positive strand
TTGCAA
RE target site (in the genome)
Valid pair
(36 - 70%)
(mapped close to RE site and in different RE fragments)
+
-
Dangling-end
(6 - 21%)
< 500 b
Extra dangling-end
(20 - 37%)
< 5 b
Semi dangling-end
(34 - 55%)
Self-circle
(0.05 - 0.15%)
+/-
Error
(0.02 - 0.8%)
> 500 b
> 500 b
Random-break
(0.1 - 1.2%)
< 75 b
Too short
(6 - 10%)
> 10 kb
Too large
(0.01 - 0.06%)
Over-represented
(4.7 - 6.8%)
Duplicated
(3 - 46%)
Percentages are percentile 10 and 90 of 278 HiC experiments, each of at least 50M reads (75 bp paired-end)
Percentages in bold are filters generally applied (valid pairs remain after removal of these reads).
4DGenome unit

Figure 11: Summary of the different products of a Hi-C experiment (top left) and the categories of mapped reads (right).

Interaction matrices derived from Hi-C experiments contain different types of biases that can be removed with normalization. In TADbit, Vanilla normalization is the default setting. It is based on Iterative Correction and Eigenvector decomposition (ICE), that assumes equal experimental visibility of each bin. Vanilla normalization performs a single iteration in which each cell is divided by the product of the sum of counts in its row times the sum of counts in its column. This process equalizes the sum of counts per bin in the matrix, creating a matrix in which all bins have the same sums. On the normalized matrix, structural features such as A/B Compartments and Topologically Associating Domains (TADs) can be called. At this point, comparative analysis between different conditions is performed to possibly unveil structural differences in three-dimensional chromatin organization. Next, quantitative comparison and feature extraction from can be performed to identify and quantify differences and similarities between conditions.

CHAPTER 3: PRION DISEASES

- *PRIONS AND PROTEINOPATIES*

Prion diseases, also known as Transmissible Spongiform Encephalopathies (TSE), are rare, invariably lethal and rapidly progressive neurodegenerative disorders that affect humans and other species of mammals[93]. Examples of human prion diseases are Creutzfeldt-Jakob disease (CJD), fatal familial insomnia (FFI) and Gerstmann-Straussler-Scheinker disease (GSS), whereas a few examples of TSE in other mammals are scrapie in sheep and goats -where this kind of diseases were originally first described- and bovine spongiform encephalopathy (BSE) in cows, which became sadly famous worldwide due to the late ‘90s outbreak in the United Kingdom (commonly known as mad cow disease), that caused more than three hundred human deaths of variant CJD due to dietary intake of infected meat and the preventive slaughter of 4.4 million bovines to contain the epidemic[94–96]. Human prion diseases have an annual incidence of on average 1.5 person/million people[97] and have a complex origin: similarly to other neurodegenerative diseases, the majority of patients show a sporadic form of the diseases while in 10-15% of cases the condition has a genetic origin[98,99]. Differently to any other neurodegenerative disease, TSEs can also be acquired with medical procedures involving infected tissue or through food. Despite of their etiology, the hallmark of this heterogeneous group of diseases is the accumulation in the central nervous system of the misfolded form of the cellular prion protein, the “scrapie” prion protein (PrP^{Sc}), creating aggregates that lead to spongiform change, gliosis, and neuronal loss. The term “prion” was coined by Prusiner in 1982 to define the properties of the proteinaceous infectious particle found in scrapie, a TSE affecting sheep[93]. PrP^{Sc} acts as an unconventional infectious agent that replicates in absence of nucleic acids and is able to induce the same pathologic conformational change in other physiological prion protein (PrP^{c})[93], which are present in the central nervous system and particularly abundant the neuronal cell membrane. Even though no changes in the primary structure of the prion protein occur, PrP^{Sc} proteins gain new biochemical characteristic compared to the physiological form: they are insoluble in water and non-denaturing detergents and are partially resistant to protease degradation[99–101]. They are also characterized by an increased percentage of β sheets that contribute to the tendency of the misfolded protein to form aggregates in fibrillar structures[93,102]. Despite decades of studies on this topic, it is still not fully understood what causes to the first conformational change of PrP^{c} in PrP^{Sc} in sporadic cases[103–106], but once the misfolded conformation is present at a minimum infectious dose, it propagates its altered structure. This will lead to the generation of aggregates and fibrils, neuronal death, spongiosis and astrocytosis, responsible for the associated clinical signs such as dementia, motor deterioration and eventually death. Prion diseases are associated with a long and silent incubation period that lasts several years followed by an aggressive clinical phase that typically leads to death in few months. From a molecular point of view, the clinically silent phase is characterized by an exponential growth of the PrP^{Sc} through seeded protein polymerization. During this phase PrP^{c} monomers are recruited and structurally converted in PrP^{Sc}, establishing a positive feedback that will generate new seeds and ultimately will cause aggregation in fibrils[107] (Figure 12). This phase is followed by a plateau phase whose length is inversely proportional to PrP^{c} expression level and continues until symptoms onset[108].

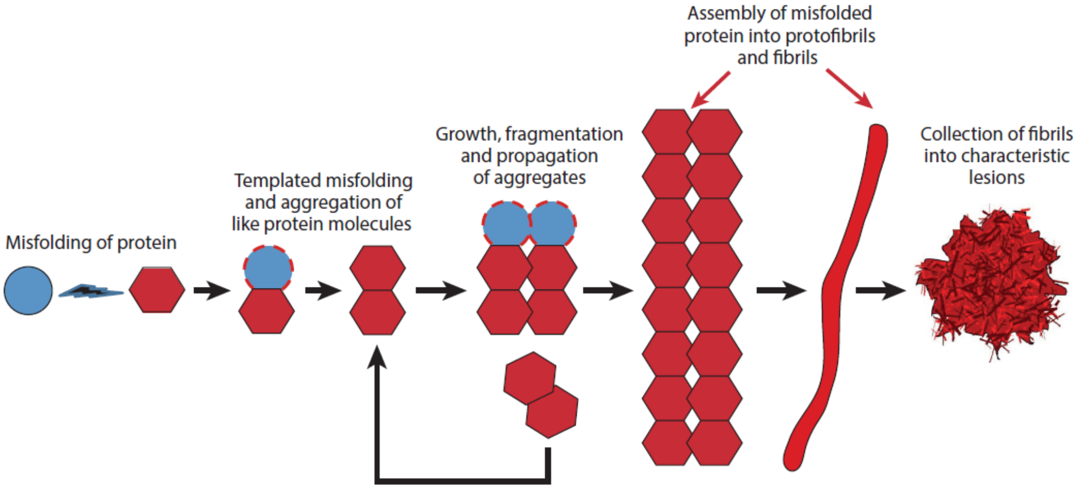
Assembly of misfolded protein into protofibrils and fibrils
Misfolding of protein
Templated misfolding and aggregation of like protein molecules
Growth, fragmentation and propagation of aggregates
Collection of fibrils into characteristic lesions

Figure 12: Schematic representation of the prion paradigm of seeded protein aggregation. The diagram shows the predicted series of events starting from a misfolded monomer to the self-aggregation of protein into fibrillar molecules and the final formation of characteristic lesions. Picture taken from Walker & Jucker, 2015[107].

PrP misfolding and accumulation causes microscopic histopathologic lesions including PrP^{Sc} aggregate deposition (associated to amyloid plaques in nearly 10% of cases), spongiform change, synaptic and neuronal loss, astrocytic gliosis and microglial activation[109]. Prion diseases represent the clearest example of "protein misfolding diseases", a classification that indicates all the neurodegenerative diseases that are caused by the accumulation of misfolded proteins in the central nervous system[82,110–113], such as Alzheimer's Disease, Parkinson's Disease, Amyotrophic Lateral Sclerosis and Frontotemporal dementias. Even though these diseases vary greatly in their pathological and clinical features, from a molecular point of view they all share the same phenomenon: the seeded aggregation of the disease-specific protein, known as the prion paradigm[107]. The mechanisms of neurotoxicity resulting from misfolding and protein aggregation in prion disease as well as in others neurodegenerative diseases is still an open question, since in none of the protein misfolding diseases the mechanism of toxicity is completely clear. While there are evidence indicating that protein aggregates are the cause of the neurodegeneration, many studies link toxicity to intermediate structures, such as for example the oligomeric state prior to the fiber formation that may act as mediator of the toxic signal through the interaction of other cellular components[106,108]. In recent years, also another feature initially typical of prion disease has been described in more common neurodegenerative disease: the phenotypic heterogeneity typical of prion strains[114].

- *PRION STRAINS*

Protein misfolding could result in different structural conformations, that are associated to different clinical profiles of the same disease or even different diseases[106,115–117]. The different conformations of PrP^{Sc} and their associated phenotypes are referred to as strains, term borrowed from virology in a time in which prion diseases were thought to be caused by slow viruses[118]. Different strains are characterized by different incubation times, profile of histological damage, PrP^{Sc} deposition pattern and clinical signs that remain stable after several passage of inoculation in different syngeneic hosts[119–121]. Strains arise from different conformations of the misfolded protein which are a consequence of many factors, such as the genetic background of the host, post-translational modifications, physical and chemical features of the cellular microenvironment. Who are the other players in this process and what is their role is currently largely unknown. Sporadic Creutzfeldt-Jakob disease (sCJD) is the most common and best studied human prion disease. As will be described in detail below, sCJD is characterized by a wide phenotypic spectrum regarding first symptoms, rate of progression, and appearance of other clinical signs. At onset and during early stages of the disease, symptoms are variable and often nonspecific: impairment of higher cognitive functions is the most frequent neurologic symptom, but visual symptoms are also quite frequent due to the major involvement of the occipital lobe in the most common subtype[109,121]. Early stages of the disease are often characterized by psychiatric features, where patients may show hallucinations, psychosis and disorientation. During the progression of the disease cognitive deficits increase and further neurologic signs appear, typically including dementia, apraxia and mutism, visual signs, and movement impairment due to cerebellar, pyramidal and extrapyramidal signs. Ultimately, patients lose contact with the environment and are usually bedridden, mute and akinetic.

Current sCJD classification recognizes six main clinical and pathological phenotypes that largely correlate at the molecular level with the genotype at *PRNP* codon 129 (Met/Met, Met/Val, or Val/Val) and the protein type (type 1 or type 2), defined by the length of the main PrP^{Sc} fragment after proteinase K digestion[121]. The phenotypic variant or “subtype” results from the combination of codon 129 genotype and protein, with two exceptions: The MM(V)1 subtype includes MM1 and MV1 cases since they are phenotypically indistinguishable, while in the MM2 group two subtypes show distinctive histopathological features, either affecting mainly the cerebral cortex (MM2-Cortical or MM2C) or the thalamus (MM2-Thalamic or MM2T)[109,121]. Among the six sCJD subtypes, MM1 and VV2 are the most common and only these strains have been considered in this work. Phenotype MM1 is the most common of all, accounting for ~65% of sCJD cases. It is associated with a rapidly progressive disease that leads to death on average in four months and has an average age of onset of 70 years. At onset, most patients suffer from cognitive decline and in one third of the cases of ataxia. It is also associated with language ability impairment, visive deficits and myoclonus or other involuntary movements. Spongiform lesions are mainly localized in the cortex (more prominently in the occipital lobe), striatum, thalamus and limitedly in the molecular layer of the cerebellum while hippocampus is not affected[120,122]. The VV2 phenotype accounts for 15-20% of sCJD cases and is associated with a slightly longer disease (six months on average) that appears earlier (64.5 years on average). It is associated to stronger movement impairment, since at onset it shows prominent ataxia, that is followed by cognitive decline and dementia, and sometimes also myoclonus. Spongiform degeneration affects cerebellum, limbic structures, striatum, thalamus, hypothalamus and brainstem while neocortex is relatively spared[122,123] (Figure 13).

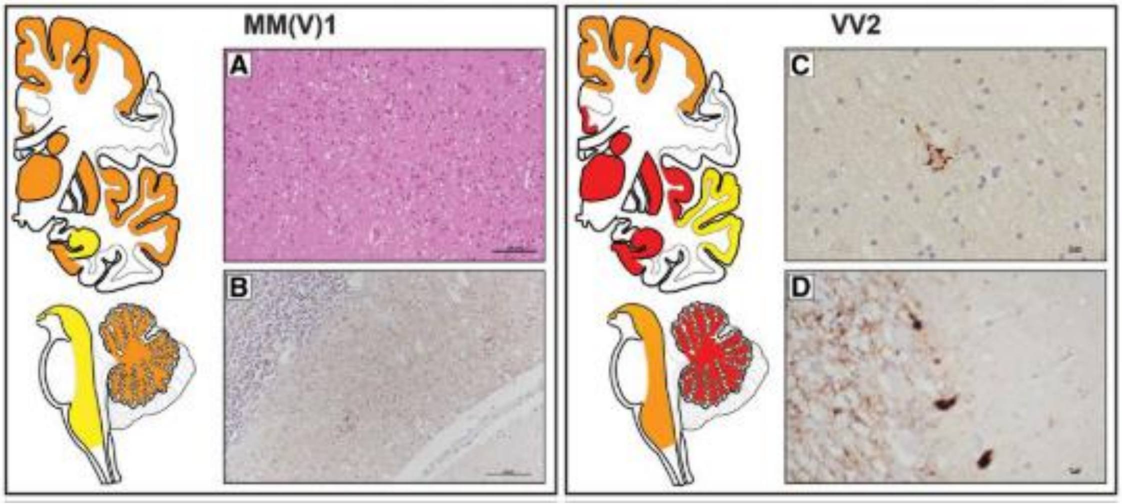

Figure 13: Lesion profiles and histopathological hallmarks of sCJD subtypes. Brain regions have been highlighted in yellow to indicate "minimal to mild" (score ≤1), in orange for "mild to moderate" (score >1–2) and red for "moderate to severe" (score >2) neuropathological damage. Adapted from Baiardi et Al. 2019[122].

- *MOLECULAR BIOLOGY OF THE PRION PROTEIN*

The prion protein is encoded by the *PRNP* gene, located in chromosome 20 in the human genome. The human *PRNP* gene is 16Kb long and made up of two exons, the second containing the whole open reading frame, thus excluding the possibility for alternative splicing. The native 253 amino acid long PRNP protein undergoes post-translational modifications resulting in a mature form of 208 amino acids, lacking a 22 amino acid signal peptide in the N-terminal region and with modifications in the C-terminal regions represented by the cleavage of other 23 residues and attachment of a glycosil-phosphatidyl-inosytol (GPI) moiety, necessary for its anchoring to the plasma membrane. The prion protein is made up of three domains: at the N-term, ~100 amino acids constitute an intrinsically disordered domain, also referred to as flexible tail, a central hydrophobic domain and a globular C-terminal domain. The globular domain contains two asparagine residues, N181 and N197, that can be N-glycosylated, determining three possible isoforms (mono/di glycosylated or unglycosylated), and two cysteines which form a disulfide bond, which play an important role in the correct protein folding. The secondary structure of the globular domain of PrP^c contains three α-helices and two short β-sheets regions, whereas the misfolded PrP^{Sc} shows a secondary and tertiary structure made of mainly of β-sheets (Figure 14).

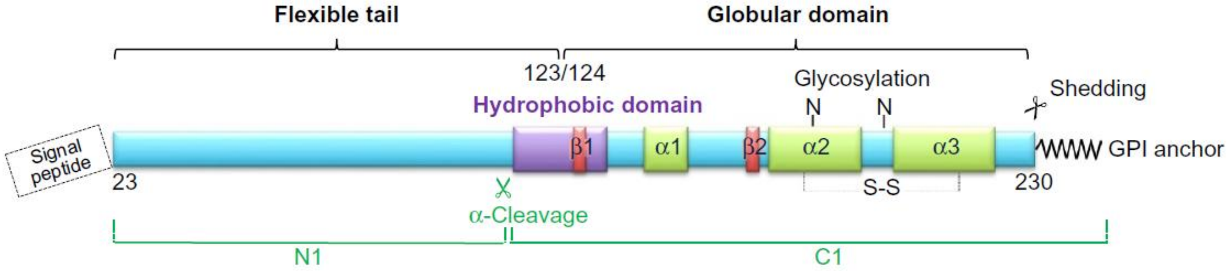

Figure 14: schematic representation of the secondary structure of PrP^c, emphasizing the N-terminal flexible tail and the C-terminal globular domain. β-sheets and α-helices are represented by rectangles in red and green, respectively, while the hydrophobic domain is highlighted in purple. N-glycosylation sites and cysteine disulfide bridge formation are indicated. Figure taken from Hirsch et al, 2017 [124].

PRNP is physiologically expressed in several tissues and is particularly abundant in the central nervous system in neurons, glia and in the endothelial cells of the blood-brain barrier. The functional role of the physiological form of the prion protein is still quite elusive: from an evolutionary perspective, both the *PRNP* gene family and the structural features of the resulting PRP protein are conserved in vertebrates and particularly in mammals, suggesting a functional importance[125]. Knockout mice lacking the *PRNP* gene are resistant to prion diseases but, except for this remarkable trait, they do not show any relevant impairment in development or behaviour[126]. Despite being dispensable for life and for proper development, *PRNP* is involved in a multitude of functions: in the CNS, it is involved in the maintenance of long-term memory[127], circadian rhythm regulation[128], synaptic activity and maintenance of white matter[124]. It is also probable an implication in anti-apoptotic mechanisms and a protective role towards oxidative stress, among other putative interplay in diverse cellular functions[124]. Recent Genome Wide Association Studies (GWAS) performed on large cohort of sCJD patients confirmed a significant association with codon 129 as the strongest risk factor and identified two other loci associated to an increased risk of sCJD, in Syntaxin-6 (*STX6,* rs3747957) and in Galactosylceramide sulfotransferase (*GAL3ST1,* rs2267161)[129], underlying intracellular trafficking and sphingolipid metabolism as probable triggering mechanisms and corroborating the likely shared underlying molecular dysregulation with other prion-like disorders. The most important known genetic risk factor and phenotypic modifier is the polymorphism at the codon 129 of the *PRNP* gene, that can result either in Methionine or Valine. Homozygotes are overrepresented in the population affected by prion diseases, while heterozygosity has a protective role. Epidemiologic studies performed on Caucasians showed that in the heathy population MV genotype was found in 50% of the population, MM in 39% and VV in 11%[130] whereas in the affected cases MM genotype represents the 69%, VV 18% and MV for 13%. This trend is coherent with the protein-only hypothesis of the spreading of PrP^{Sc}, that would be facilitated when both alleles code for a protein with the same primary sequence. However, it is still an open question how this polymorphism modulates the phenotypic outcome of the disease, that is shown in its clearest form in the cases of the pathogenic mutation D178N, that combined with 129V in cis results in Creutzfeldt-Jakob disease whereas results in Fatal Insomnia when in cis with 129M.

In terms of gene expression, microarray and RNA sequencing technologies have been applied to try to determine the most affected biological processes and molecular pathways at various stages of the disease. Most of the currently available knowledge comes from murine models: according to current literature[131–133] in the early stage of the disease the most prominent changes in gene expression are associate to immune response through the complement system and leukocyte infiltration, associated to microglia and astrocyte activation. During the intermediate stages of PrP^{Sc} accumulation, the transcriptional profile seems to change towards pathways involving membrane regulation and vesicle traffic, with the activation of sphingolipid, glycosaminoglycans, cholesterol metabolisms. In the final stage of disease, a transcriptional down regulation of genes associated with synaptic transmission and axonal growth occurs, followed by activation of cellular processes associated with apoptosis. Only few studies about gene expression changes carried out on human samples exist[25,134–136]. Despite being able to provide information only about the final stage of the disease, these studies allow to get insights also on sporadic forms of the disease, differently from animal model that describe the acquired forms. These works as well highlight a prominent impairment of gene expression profiles seem to parallel processes observed in animal models, providing anyway further insights towards peculiarity of the human disease and also allowing parallelisms with other human proteinopaties and aging. Since in all these conditions clinical symptoms appear only when neuronal death has already occurred and given the increasing number of people expected to suffer from this kind of pathologies, studies focused on the improvement of our understanding of the molecular pathways associated to phenotypic outcomes of protein misfolding diseases are very much needed. The previously described technologies and computational approaches represent an incredibly resourceful opportunity to address this challenge with modern approaches.

AIM OF THE PROJECT

This project was carried out in collaboration with the Neuropathology group at the Istituto delle Scienze Neurologiche of Bologna and the Applied Physics group of University of Bologna, and with the external collaboration of the Structural Genomics group of the Centre de Regulació Genòmica-Centre Nacional d'Anàlisi Genòmica (CRG-CNAG) of Barcelona. The aim of this project is to improve the current understanding of the molecular biology underlying phenotypic heterogeneity in prion diseases and the strain phenomenon with a multi-omics approach.

Specifically:

1. Produce, analyze, and compare genomic targeted sequencing data of a carefully selected cohort of forty-eight samples belonging to the two strains MM1 and VV2, to identify possible genetic modifiers and recurrent genetic patterns in the two classes.
2. Produce, analyze, and compare transcriptomic data of brain tissue in a subset of the previous cohort to characterize differentially expressed genes and altered pathways in the two classes at the end of the neurodegenerative process.
3. Produce, analyze, and compare three-dimensional chromatin organization data in patients at disease onset and in heathy controls of the same age to explore possible epigenomic causes or response to the disease in peripheral immune cells.

MATERIALS AND METHODS

GENOMICS: DNA TARGET SEQUENCING

- ### *SAMPLES*

Forty-eight patients diagnosed with definite sporadic CJD, accordingly to the updated clinical diagnostic criteria for sporadic Creutzfeldt-Jakob disease[137], afferent to the IRCCS Institute of Neurological Sciences of Bologna, either as outpatients, inpatients or sent for genetic analysis, were selected. Among the several samples available, selected samples belonging to patients with genotype Met-Met or Val-Val at the codon 129 of the *PRNP* gene, without any co-pathology, with a detailed clinical description of the disease course, with age of onset of the disease as similar as possible to the average (~60 years) and with as much other clinical and pathologic information as possible.

- ### *WET LAB*

All experimental steps were performed in the neuropathology laboratory of the IRCCS Institute of Neurological Sciences of Bologna.

- #### DNA EXTRACTION

Genomic DNA from peripheral blood or brain tissue was isolated using the Maxwell 16 extractor (Promega, Madison, WI, USA). Extracted genomic DNA was quantified using the Quantus Fluorometer (Promega) with QuantiFluor double-stranded DNA system.

- #### DNA TARGET SEQUENCING

DNA libraries were prepared with DNA Prep with Enrichment kit (Illumina, CA, USA) performing enrichment with Illumina Neurodegeneration panel which covers over 8.7 Mb in 118 genes, including introns, exons, untranslated regions and promoters to support surveys of both coding and regulatory regions. Library prep was performed following instruction provided by the vendor (Illumina DNA Prep with Enrichment Reference Guide (1000000048041)). As suggested by the vendor, a quantity between 50-1000 ng of genomic DNA was used as input material. Tagmentation of genomic DNA was performed with Enrichment Bead-Linked Transposomes (eBLT) with a five-minute incubation at 55°C. Tagmentation performed with eBLT allows the simultaneous fragmentation of genomic DNA and the attachment of adapter sequences to fragments. DNA fragments anchored to the beads were washed with the appropriate buffer to get rid of residues of the previous reaction before Polymerase Chain Reaction (PCR) amplification. During PCR amplification, tagmented DNA still bound to eBLT beads is amplified and i7 and i5 index sequences are added at the extremities of each fragment. In this case, “IDT for Illumina Nextera DNA UD Indexes (96 Indexes) Set A” were used as indexes. Cleaved DNA anchored to eBTL beads will work as a template while the amplification products will end up in solution. Given the not limiting amount of input DNA and to avoid PCR duplicates in the sequencing reads, only 9 cycles were performed using the following PCR program.

Initial denaturation	72°C 98°C	3 minutes 3 minutes	
Denaturation Annealing Extension	98°C 60°C 72°C	20 seconds 30 seconds 1 minute	9 cycles
Final extension	72°C	3 minutes	

PCR products are then separated from eBLT beads, and the supernatant will undergo a clean-up step with AMPure XP Beads (Beckman Coulter) in which only the amplified DNA with the desired length (300-400 bp) will be selected for the further steps.

PCR products were then quantified with Quantus Fluorometer (Promega) with QuantiFluor double-stranded DNA system and fragments length was checked with trough electrophoresis with 1.3% agarose gel. Indexed samples were then pooled together in four pools with twelve samples each. Library normalization by mass was performed using whenever possible 450ng of each library in the pool (three pools, 36 libraries), in one pool the normalization was performed using 230 ng per library due to lower PCR yield. Pooled libraries underwent enrichment using biotinylated oligos (TruSeq Neurodegeneration - Enrichment Oligos only), that capture and enrich for the genomic regions covered by this 8.7 Mb gene panel. For each pool, hybridization was performed with 30 µl of pre-enriched libraries, 10 µl of probe panel and 60 µl of a mix of Illumina buffer provided with the library preparation kit using the following touchdown program:

Initial denaturation	95°C	5 minutes	
Denaturation Touchdown hybridation	94°C ↓ 62°C	1 minute	16 cycles, Decreasing 2°C per cycle
Hybridation	62°C	90 minutes/overnight	

Probes are then captured using streptavidin beads to use the affinity of biotin-streptavidin bond. Enriched beads-ligated libraries are then extensively washed with buffer specific for this Illumina library kit to avoid carryover of the enrichment step. Afterwards, enriched libraries are eluted from the beads and are then further amplified with PCR using the following program.

Initial denaturation	98°C	3 minutes	
Denaturation Annealing Extension	98°C 60°C 72°C	10 seconds 30 seconds 30 seconds	14 cycles
Final extension	72°C	5 minutes	

Amplified libraries went through a final clean up with AMPure XP beads to remove any PCR-reagent carryover. Final libraries were then quantified with Quantus Fluorometer (Promega) using QuantiFluor double-stranded DNA system and fragments length was checked with trough electrophoresis with 1.3% agarose gel. Paired end sequencing was performed directly in the lab with a NextSeq 500 (Illumina) sequencer following the instruction provided by the manual “NextSeq System Denature and Dilute Libraries Guide”. Pooled libraries were normalized at first at 4nM in a unique 5 µl final pool. Denaturation was performed with 5 µl NaOH 0.2N for 5 minutes and then stopped with 5 µl Tris HCl pH7 200µM. Pooled libraries were afterwards diluted to concentration of 20pM with the Illumina HT1 buffer. Finally, libraries were furtherly diluted and then uploaded with the final concentration of 1.2 pM on an Illumina cartridge and flow cell High Output (300 Cycles) on which a 150bp paired-end sequencing was performed.

- *DRY LAB:*

Computational analysis was performed on Linux servers of the Applied Physics group of University of Bologna.

- PIPELINE DEVELOPMENT

To perform bioinformatic analysis of DNA sequencing data, a specific pipeline was developed. All steps were performed using packages installed through the open-source package and environment management system miniconda, using the Snakemake workflow management system to organize the several steps required in this pipeline. As first step, the entire sequencing run folder containing the raw sequencing reads in Base Call (*.bcl*) format was demultiplexed to FASTQ, that underwent a quality control step in which low quality bases and sequencing adapters were removed. Subsequently, FASTQ files were mapped to the reference genome (GRCh37) with BWA aligner[81] using *mem* algorithm. Aligned files were sorted and marked for PCR duplicates. Sequencing coverage was computed with GATK DepthOfCoverage[138]. Variant calling of Single Nucleotide Polymorphisms (SNPs) and small indels was performed using Strelka2[20], setting the analysis for germline variant discovery. Detected variants were subsequently filtered according to quality parameters and constrains provided by the genomic regions covered by the target sequencing panel. All tools used in the pipeline are summarized in table 1.

TOOL NAME	**AIM**
Fastqc	Quality check
Trimmomatic	Trimming
BWA mem	mapping to reference genome
Picard SortSam	Sorting
Picard MarkDuplicates	Marking PCR duplicates
Picard BuildBamIndex	Building index
GATK DepthOfCoverage	Compute sequencing coverage
Strelka2	Variant detection
GATK Select Variants	Filtering low quality and off target calls

Table 1: summary of the bioinformatic tools used in the custom pipeline built for the secondary analysis of DNA target sequencing data.

o VARIANT ANNOTATION AND EFFECT PREDICTION

VCF files were then annotated with BaseSpace Variant Interpreter (Illumina). To exclude technical errors in handling several samples at once, in each sample variant discovery was validated on already genotyped loci, such as codon 129 in the *PRNP* gene and/or *APOE* genotype. Missense variants effect prediction was estimated with SIFT[139] and Polyphen2[140].

o STATISTICAL ANALYSIS

The genetic information contained in Variant Call Format (VCF) files was transformed into binary data through an in-house python script (https://github.com/UniboDIFABiophysics/binaryVCF), generating a matrix in which each row represents a variant reported in the provided VCF files at least once (which is encoded as chromosome number-position-reference allele- alternate allele) and each column is named after an ID assigned to each sample. In each cell of the matrix is reported the number of alternative alleles for each locus, thus 0 indicates that the variant is not present in the VCF file of the patient whereas 1 indicates its presence in heterozygosity and 2 the presence of the variant in homozygosity. This matrix was used as input for machine learning methods and statistical analysis. Allele frequencies of each variant described at least in one patient of our cohort were calculated and then compared with those reported in the GnomAD database[141] for the non-Finnish European population using Fisher's exact test and Benjamini-Hochberg multiple test correction. The same statistical test was used to compare allele frequencies between samples from MM1 and VV2 strains.

- FUNCTIONAL ANALYSIS

Functional analysis was performed with over-representation analysis (ORA), using g:Profiler[142] on two different gene sets: one derived from genes harboring variants predicted as “Pathogenic” or “Likely Pathogenic” by both SIFT and Polyphen2, and the second derived by genes harboring at least one variant with allele frequency significantly (padj <0.05) higher/lower compared to the European population.

- MACHINE LEARNING APPROACHES

The overall genomic information of the dataset carried in the ternary matrix was used as input for supervised and unsupervised analysis, using SciKit-Learn[143], Seaborn [144], Plotly express [145], pandas, Numpy[146] and SciPy[147] packages on Jupyter notebooks. Both supervised and unsupervised methods were used to extract as much valuable information as possible from the transformed data. To visualize such high dimensional data, different dimensionality reduction techniques were tested, such as Principal component analysis (PCA), t-distributed stochastic neighbour embedding (t-SNE) using Jaccard similarity as metric and Uniform Manifold Approximation and Projection (UMAP). As clustering methods, SciKit Learn dendrograms and K-Means were used. As supervised methods, decision trees on binary data labelled accordingly to the different strains of each patient were used. The classifier was trained on a random selection of 2/3 of the dataset and adequate branching depth was set to avoid overfitting. The classification rules were tested on a validation set represented by the remaining 1/3 of the dataset.

TRANSCRIPTOMICS: RNA SEQUENCING

- *SAMPLES*

Twenty brain samples of sCJD patients with strain MM1 or VV2 were selected. Sporadic cases of sCJD were classified as MM1 or VV2 according to histopathological criteria and PrP^{Sc} typing. All cases used were selected from previously used cases in the genomic analysis experiment with tissue suitable for analysis (body kept refrigerated (2-4°C) before autopsy and with a post-mortem < 36 h to minimize RNA degradation) with mild to moderate lesions in the frontal cortex on histopathological examination. The frontal cortex was chosen as the area of interest because it is pathologically involved in all CJD subtypes and usually available for sampling.

- *WET LAB:*

All experimental steps were performed in the neuropathology laboratories of the IRCCS Institute of Neurological Sciences of Bologna.

- RNA EXTRACTION AND QUALITY ASSESMENT

Total RNA was extracted from 50mg of frozen frontal cortex with RNeasy Lipid Tissue Mini Kit (Qiagen) following the protocol provided by the vendor. Given the infectivity of tissues affected by prion diseases, RNA extraction was performed in a laboratory with biosafety level 3. Total RNAs were then quantified with NanoDrop 2000 (Thermo Scientific). One µg of total RNA was subsequently treated with DNase I, RNase-free (Thermo Scientific), following the protocol provided by the vendor. Quality assessment of total RNA’s quality was performed trough capillary electrophoresis with Fragment Analyzer system (Agilent Technologies) with RNA Kit (15nt) (Agilent Technologies), following the protocol provided. This methodology establishes the degradation level of RNA molecules based on their length. For the protocol we selected for

RNAseq library preparation, the most important parameter was $DV_{200,}$ which express the percentage of RNA molecules longer than 200nt, that must be at least above 30% and optimally above 70%.

- RNA SEQ

RNA libraries were prepared with Truseq RNA Exome (Illumina) performing enrichment with Illumina Exome Panel – Enrichment Oligos Only. To avoid errors derived by pipetting very small volumes, 1µl of total RNA was used as input for each sample. Thus, the quantity of total RNA used ranged between 79-110 ng per sample. According to the DV_{200} value of each sample and to the separation profiles on the electropherogram, an optional round of fragmentation was performed thru incubation at 98°C for a time ranging between 0 and 8 minutes as reported in table 2.

ID	DV200 (%)	Minutes
#2	90.6	8
#33	76.8	5
#32	77.5	0
#13	95.1	8
#15	90.9	0
#12	88.6	5
#36	86.7	0
#40	73	0
#14	69.8	0
#43	90.9	5
#18	89.4	5
#17	79	0
#44	85	0
#45	87.9	5
#46	83.8	0
#22	86.4	0

Table 2: Quality metrics, expressed in the percentage of RNA fragments longer than 200nt and the incubation times for RNA fragmentation of each sample.

After fragmentation, RNA fragments primed with random hexamers were retrotranscribed with SuperScript VILO reverse transcriptase (Thermo Fisher) into first strand cDNA. The provided buffer contains Actinomycin D to prevent spurious DNA-dependent synthesis, while allowing RNA-dependent synthesis and improving strand specificity, was added with SuperScript VILO at a ratio of 1:10. To each sample, 8µl of the prepared first strand synthesis buffer was added, and the following incubations were performed:

Temperature	Time of incubation
25 °C	10 minutes
42 °C	15 minutes
70 °C	15 minutes

Subsequently, the second strand of cDNA was produced. In this step, the RNA template is displaced, and a replacement strand is synthetized incorporating uridine in place of Deoxythymidine triphosphate (dTTP) to generate double strand cDNA. The incorporation of uridine prevents the second strand synthesis during amplification. 20µl of Illumina second strand master mix and 5µl of resuspension buffer were added to each sample and after, a 1-hour incubation at 16°C was performed to complete the second strand synthesis. The obtained ds cDNAs were then purified with AMPure XP, washed with 80% ethanol, and eluted with resuspension buffer. cDNAs underwent a 3’ adenylation step, in which a single 'A' nucleotide was added to the 3′ ends of the blunt fragments to prevent them from ligating to each other forming chimeric artifacts during the following adapter ligation step. To each sample, 12.5µl of Illumina A-tailing mix was added and the following incubation steps were performed:

Temperature	Time of incubation
37 °C	30 minutes
70 °C	5 minutes
4 °C/ on ice	1 minute

“Illumina TruSeq RNA Single Indexes (12 indexes, 24 samples) Set A” were used in the adapter ligation. To each sample, 2.5µl of RSB, 2.5µl RNA adapters and 2.5µl Ligation enzyme mix was added, and then a 10-minute incubation at 30°C performed. Samples were then chilled on ice and the ligation reaction was stopped with 5µl Stop Ligation Buffer. Indexed cDNAs were then purified with AMPure XP and washed with 80% ethanol in two consecutive rounds and eluted in 20µl of resuspension buffer.

DNA fragments carrying adapters at both extremities were then amplified trough PCR using the provided PCR master mix and the following protocol:

Initial denaturation	98°C	30 seconds	
Denaturation	98°C	10 seconds	14 cycles
Annealing	60°C	30 seconds	
Extension	72°C	30 seconds	
Final extension	72°C	5 minutes	

Amplified libraries were cleaned up with AMPure XP and washed with freshly prepared 80% ethanol to remove any PCR-reagent carryover. Libraries were quantified with Quantus Fluorometer (Promega) using QuantiFluor double-stranded DNA system and fragments length was checked trough electrophoresis with 1.3% agarose gel.

Based on the quantification, for each library a volume corresponding to 200ng of dsDNA was pooled into 4-plex libraries pools. Pooled libraries underwent a first hybridization with biotinylated oligos recognizing coding exons (Illumina Coding Exons Oligos) to enrich only for regions corresponding to the initial mRNA. 45µl of pooled libraries were mixed with 50µl of Illumina Capture Target Buffer and 5µl of Coding Exons Oligos, hybridization was performed with the following touchdown program:

Initial denaturation	95°C	10 minutes	
Denaturation Touchdown hybridation	94°C ↓ 62°C	1 minute	18 cycles, Decreasing 2°C per cycle
Hybridation	58°C	90 minutes	

Enriched libraries were then mixed with streptavidin magnetic beads (SMB) to capture hybridized probes. Two heated washes performed with Illumina washing buffers remove nonspecific binding from the beads, afterwards the enriched libraries were eluted from the beads and prepared for a second round of hybridization, with the same conditions used in the first one, to ensure high specificity of the captured regions. After another round of streptavidin capture and heated washes, libraries were purified with AMPure XP beads and amplified trough PCR with the following program.

Initial denaturation	98°C	30 seconds	
Denaturation Annealing Extension	98°C 60°C 72°C	10 seconds 30 seconds 30 seconds	10 cycles
Final extension	72°C	5 minutes	

Amplified libraries went through a final clean up with AMPure XP beads to remove any PCR-reagent carryover. Final libraries were then quantified with Quantus Fluorometer (Promega) using QuantiFluor double-stranded DNA system and fragments length was checked with electrophoresis in 1.3% agarose gel. Paired end sequencing was performed with a NextSeq 500 (Illumina) sequencer following the instruction provided by the manual “NextSeq System Denature and Dilute Libraries Guide”. Pooled libraries were normalized at first at 4nM in a unique 5 µl final pool. Denaturation was performed with 5 µl NaOH 0.2N for 5 minutes and then stopped with 5 µl Tris HCl pH7 200µM. Pooled libraries were afterwards diluted to concentration of 20pM with the Illumina HT1 buffer. Finally, libraries were furtherly diluted and then uploaded at a final concentration of 1.4 pM on an Illumina Cartridge and Flowcell High Output (150 Cycles) on which a 75bp paired-end sequencing was performed.

- DdPCR

Differential Gene Expression results obtained by RNAseq were validated using droplet digital polymerase chain reaction (ddPCR). Five genes with the most extreme Log2FoldChange were selected for validation among those differentially expressed: Antileukoproteinase (*SLPI*), Interleukin-1 receptor antagonist protein (*IL1RN*), Cytochrome P450 3A5 (*CYP3A5*), Macrophage mannose receptor 1 (*MRC1*) and Olfactory receptor 2M2 (*OR2M2*). Reverse transcription ware carried out on 1μg of DNAse I-treated total RNA with SuperScript VILO cDNA Synthesis Kit (Thermo Fisher Scientific). Reactions for ddPCR assay were performed using the QX200 Droplet Generator, the QX200 Droplet Reader, the C1000 Touch Thermal Cycler, and the PX1 PCR Plate Sealer (Bio-Rad, Hercules, CA, USA) following the manufacturer's instructions. Reactions were carried out in triplicate using the ddPCR Supermix for Probes (no dUTP). The cDNA copies/unit were quantified using the QuantaSoft Software (Bio-Rad). The following three reference genes were chosen as housekeeping reference for brain tissue based on available literature research[148,149]: *XPNPEP1* (X-prolyl aminopeptidase P1), considered a gold standard reference gene for RNA expression studies in post-mortem human brain tissue, *UBE2D2* (ubiquitin-conjugating enzyme E2D 2), and *CYC1* (cytochrome c1).

- *DRY LAB:*

Computational analysis was performed on Linux servers of the Applied Physics group of University of Bologna.

- PIPELINE DEVELOPMENT

To perform secondary analysis of RNA sequencing data, a specific pipeline was developed. As in the previous pipeline, all bioinformatics steps were performed using packages installed through miniconda and Snakemake workflow management system was used to organize the several steps required in this pipeline. Table 3 presents the main tools used in the pipeline.

TOOL NAME	**AIM**
Fastqc	Quality check
Trimmomatic	Trimming
STAR	mapping to reference genome and quantification
Picard SortSam	Sorting

Picard MarkDuplicates	Marking PCR duplicates
Samtools index	Building index

Table 3: summary of the bioinformatic tools used in the custom pipeline built for the secondary analysis of DNA target sequencing data.

After demultiplexing, FASTQ files underwent a quality control step in which low quality bases and sequencing adapters are removed. Subsequently, FASTQ files were aligned to the reference genome (GRCh37) with STAR aligner[150] with default setting and read counts quantification. Aligned *bam* files are sorted and marked for PCR duplicates.

- DIFFERENTIAL GENE EXPRESSION ANALYSIS WITH DESEQ2

Differential gene expression (DGE) was computed on read counts files output of STAR, using DeSeq2[88] on Jupyter notebooks[151]. Prior to DGE analysis, normalization of read counts between samples was performed creating scaling factors (see introduction). On normalized counts, quality checks were performed to explore sample-level and gene-level features. Principal component Analysis (PCA) and hierarchical clustering were used as sample-level quality checks to assess which samples showed the greatest similarity and to identify major sources of variation in the data, and among the available metadata of the analyzed samples. Gene-level QC was performed with default settings. This step is required to omit genes with zero counts in all samples, with an extreme count outlier or with a low mean normalized counts. This gene-level QC step is important to remove gene that would otherwise reduce the specificity in correctly identifying differentially expressed genes. Based on the previous steps, the design formula of DeSeq2 was written using “biological sex”, “experimental batch”, “disease duration” and “post-mortem conservation” values as parameters of internal sources of variation while testing for differences between the MM1 and

VV2 strains. Differential gene expression was then tested with Wald test on gene model coefficients (LFC), using as null hypothesis that no differential expression is present between groups. Shrinkage parameters of $\log_2$FoldChange were estimated with “apeglm”[152] method on the comparison between VV2 and MM1, that in this work represent the baseline. Significance cut-off of $p < 0.05$ was used and multiple test correction was performed with Benjamini-Hochberg. Results of DGE analysis were the annotated using the R package “Annotables” on human genome GRCh37.

- FUNCTIONAL ENRICHMENT

To gain more functional insights from the list of differentially expressed genes in the VV2 strain compared to MM1, functional analysis was performed with over-representation analysis (ORA) and Functional Class Scoring (FCS) approaches (see introduction, chapter 2). Functional analysis with both approaches was performed using the Bioconductor packages “ClusterProfiler”[153], “DOSE”[154] and “Pathview”[155] on Gene Ontology [89]categories. Representation of enriched over and under expressed pathways was performed using EnrichmentMap and Annotables on Cytoscape. To further explore the biological interplay of differentially expressed genes, we performed protein-protein interaction analysis (PPI) mapping under and over expressed genes on STRING interactome, considering only physical interactions. Results were functionally interpreted with over representation analysis using KEGG database to discover pathways significantly enriched based on the physical interactome.

EPIGENOMICS: HI-C

- *SAMPLES*

To analyze 3D organization of the genome, it is compulsory to use samples with intact structural organization of the nucleus and the nucleic acids in it. To preserve and stabilize the 3D structural features of the samples, Hi-C protocol require an in-nuclei chromatin crosslinking prior to snap freezing. The crosslinking is performed with formaldehyde, which permeates cell membranes and leads the formation of covalent bonds between DNA, proteins and other reactive molecules in close proximity. This procedure is not usually performed on standard routine since standard NGS protocols study the nucleic acids linear sequence that is perfectly preserved simply storing samples at -20°C for DNA and -80°C for RNA. Formalin-fixed paraffin-embedded samples could not have been and option for Hi-C experiments because, despite a work from Troll et al[156] showed that in principle this kind of samples have some applications for structural studies, brains of patients affected by prion diseases, after formalin fixation, require an additional treatment to strongly reduce prion infectivity with formic acid, which solubilize proteins through protonation, destabilize hydrogen bonds and interacts with hydrophobic residues. Therefore, it was not possible to use already collected CJD samples for this analysis, thus buffy coat samples of new sCJD patients had to be used for this Hi-C experiment. From the beginning of February to the end of April 2021, each blood sample of patients with rapidly progressive cognitive decline or suspected Creutzfeldt-Jakob Disease afferent to the Cognitive Disorders and Dementia Center of the UOC Clinica Neurologica, Bologna, either as outpatients, inpatients, or sent for genetic analysis, was treated with the following protocol. Among them, two samples with confirmed diagnosis of CJD via positive result at RT-QuIC analysis [not covered in this work] were compared with two samples of healthy volunteering controls chosen to obtain the same average age of 57 years both in cases and controls.

- *WET LAB*

  - CHROMATIN CROSSLINKING

Whole blood samples were centrifuged at 4000 rpm for 15 minutes to separate plasma, buffy coat and red blood cells. Only white blood cells were collected and washed with red blood lysis buffer (NH_4Cl 149.9 mM, $NaHCO_3$ 10 mM, EDTA 1.1 mM) to remove any residue of erythrocytes. Afterwards, buffy coats were washed with abundant Phosphate-Buffered Saline (PBS) buffer pH 7.4 (Thermo Fisher). Crosslinking of the cells was performed with a solution 1% final of Formaldehyde 37% (Sigma Aldrich) and PBS (45 ml PBS, 1.25 ml Formaldehyde 37%), cells were incubated for 10 minutes in rotation at room temperature. Formaldehyde quenching was performed adding 3.1ml of Glycine 2M (0.125 M final) and leaving in incubation at room temperature for 5 minutes and afterwards on ice for 15 minutes. Crosslinked cells were then centrifuged at 1200 rpm at 4°C for 10 minutes, the supernatant was discarded, and the cell pellet was resuspended in 2ml of PBS. Two further rounds of PBS wash with centrifugation at 1500 rpm, 4°C for 5 minutes were performed to remove any residue of formaldehyde and glycine. Crosslinked cells were then stored at -80°C.

Samples were then shipped in dry ice and the whole Hi-C protocol was performed in the laboratories of the Centre of Genomic Regulation (CRG) while sequencing and computational analysis were performed at the National Centre of Genomic Analysis (CNAG), both located in Barcelona, in collaboration with the structural genomics group, led by prof. Marc Marti-Renom.

- HI-C

The Hi-C protocol was performed with the ARIMA Hi-C kit for mammalian cell lines (Arima Technologies), following the protocol provided by the vendor. As extensively illustrated in the introduction section, this part of the protocol captures the sequence and the structure of the genome using the covalent bonds that are formed during the crosslinking step. The expected input of nuclei per aliquot was estimated on extra samples, since 3D chromatin structure is disrupted by thaw-freeze cycles. This step is necessary to determine how many crosslinked cells comprise 750ng-5µg of DNA required by this protocol. Crosslinked cells were resuspended in 500µL of freshly prepared ice-cold lysis buffer supplemented with protease inhibitors (10 mM Tris-HCl pH8.0; 10 mM NaCl; 0.2% NP40; 1X Protease inhibitors (Complete Protease inhibitor, EDTA Free (100X in nuclease-free water, Roche) and incubate on ice for 30 minutes. Aliquots of 70µl, 100µl and 150µl of lysis buffer and cells were collected and centrifuged for 5 minutes at 3,000 rpm at 4ºC. The pellet was then washed with 300µL of cold 1X NEBuffer2 (New England Biolabs) and resuspended in 190µL of 1X NEBuffer2 and 10µL of 10% SDS (final concentration 0.5% (vol/vol)). A 10-minute incubation at 65ºC was performed and afterwards the aliquot was put on ice. 400µL of 1X NEBuffer2 and 120µl of 10% Triton X-100 were added to quench the SDS. Afterwards, 20µl of proteinase K 20mg/ml was added to reverse crosslinking and tubes were left incubating at 65°C overnight. DNA was then purified with AMPure XP beads and quantified with Qubit dsDNA HS Assay Kit (Thermo Scientific). In this case, to satisfy the required input of nearly 2.5µg of DNA (accepted range 750ng-5µg) in no more than 20µl of volume, the amount of cells contained in 100µl of the ice-cold lysis buffer and crosslinked cells solution were necessary. The four selected samples were thus resuspended in 500µl of freshly prepared ice-cold lysis buffer supplemented with protease inhibitors and incubated on ice for 30 minutes. 100µl were taken and centrifuged for 5 minutes at 3,000 rpm at 4ºC. The supernatant was discarded and 20µL of ARIMA Lysis Buffer was added, samples were incubated at 4°C for 15 min. To each sample 24µL of ARIMA Conditioning Solution was added and a 10-minute incubation at 62°C was performed. Then, 20µl of Stop Solution was added and another incubation at 37°C for 15 minutes was performed. At this point, the ARIMA restriction enzymes mix was added to each sample together with the appropriate buffers and the following incubations were performed.

Temperature	**Time of incubation**
37 °C	60 minutes
65 °C	20 minutes
25°C	10 minutes

After the RE digestion, 5'-overhangs are filled in and marked with biotin using biotinylated nucleotides, adding to each tube the appropriate ARIMA buffer and enzyme and with a 45-minutes incubation at room temperature for 45 min. Extremities of crosslinked DNA fragments are then ligated with the provided ligase enzyme and with a 15-minutes incubation at room temperature. Crosslinking is then reversed with the provided ARIMA enzyme and samples were incubated as follows.

Temperature	**Time of incubation**
55 °C	30 minutes
68 °C	90 minutes
25°C	10 minutes

DNA was then purified with AMPure XP beads and quantified with Qubit dsDNA HS Assay Kit (Thermo Scientific). Intermediate quality checks were performed to quantify the fraction of proximally-ligated DNA that has been labelled with biotin, that is the fraction of DNA that will be used in the next generation sequencing library preparation following the Hi-C protocol.

- HI-C LIBRARY PREPARATION

For each sample, 1µg in 100µl of Hi-C product was fragmented trough sonication (Bioruptor Pico sonicator) to obtain DNA fragments of 300-400 bp. In this case, 5 cycles of 20 seconds ON and 60 seconds OFF resulted to be the best condition for obtaining fragments of the desired length, which was assessed through electrophoresis. Library preparation was performed using NEBNext DNA Library Prep Master Mix Set for Illumina (New England BioLabs) following instructions provided by the vendor. Hi-C biotinylated fragments were isolated from the rest the fragments with a biotin-streptavidin pull down using 20µl of Dynabeads Streptavidin T1 beads (Invitrogen) and incubation for 30 minutes in rotation at room temperature. Two rounds of washing with biotin binding buffer followed the pulldown. Afterwards, samples were resuspended in 100 µL of end repair mix (NEBNext End repair module E6050L, 10 µL of 10X NEBNext End Repair buffer, 5 µL of NEBNext End Repair Enzyme Mix, 85 µL H_2O) and incubated for 30 minutes at room temperature. Beads were then separated by the supernatant on a magnetic stand and washed with biotin binding buffer. The A-tailing step was performed resuspending the beads with 100 µL dATP attachment mix (NEBNext A-tailing module E6053L, 10 µL of 10X NEBNext dA-tailing buffer, 5 µL of 5 U/µL NEBNext dA-tailing enzyme mix, 85 µL of H_2O) and incubating at 37°C for 30 minutes. Beads were again washed with biotin binding buffer and index adapters were added to the extremities of each fragment. Beads were resuspended in 50µL of Adapters ligation mix (10 µL of 5x NEB Quick ligation reaction buffer, 2.5 µL of NEBNext Multiplex Oligos for Illumina (96 Unique Dual Index Primer Pairs), 2 µL of NEB T4 DNA ligase, 35.5 µL H_2O) and incubated at room temperature for 15 minutes. To each library, 3 µl of NEBNext USER (from 96 Unique

Dual Index Primer Pairs) were added and samples were incubated for 15 minutes at 37ºC. Two rounds of washing with biotin binding buffer followed the index ligation, streptavidin beads were then resuspended in 20 µl of sterile H_2O. At this point, libraries were amplified trough PCR using 25 µl NEBNext High-Fidelity 2X PCR Master Mix (M0541S), indexes from NEBNext Multiplex Oligos for Illumina (96 Unique Dual Index Primer Pairs) with the following program:

Initial denaturation	98°C	30 seconds	
Denaturation	98°C	10 seconds	8 cycles
Annealing	65°C	30 seconds	
Extension	72°C	30 seconds	
Final extension	72°C	5 minutes	

In this on-beads amplification, the DNA template will stay attached to the DynaBeads while the amplified library will be in the supernatant. Amplified libraries were then separated from the beads on a magnetic stand and supernatant went through a final clean up with AMPure XP beads to remove any PCR-reagent carryover. Libraries were eluted from AMPure beads in 30µl of Tris Buffer EB (Qiagen) and quantified with a fluorometric method (Qubit, Thermo Fisher) while libraries length was checked trough electrophoresis with 1.3% agarose gel. All libraries satisfied all the quantity and quality check required prior to sequencing. Sequencing was performed by the sequencing facility of the National Centre of Genomic Analysis (CNAG) with Illumina NovaSeq6000 sequencer, performing paired end sequencing 2x150bp reads, with a minimum sequencing depth of 400 M PE reads/sample.

- *DRY LAB*

Computational analysis of Hi-C data was performed in the National Centre of Genomic Analysis (CNAG) of Barcelona. This analysis was performed with TADbit tools[41,157] for secondary analysis and matrices generation, while CHESS[158] was used for the comparative anlysis. Similarly to previous sequencing data, demultiplexing of raw data represents the first step. All following steps -unless stated- were performed with TADbit tools, setting appropriate options as shown in the available documentation. FASTQ files underwent a quality control step before being mapped to the reference genome (GRch38). TADbit uses GEM mapper[159] using and iterative mapping strategy and using the previous knowledge of the cutting sequences of the restriction enzymes used in the Hi-C library preparation. Next, mapped read are filtered to correct for experimental biases and errors, and to omit sites with many zero counts. A normalization step was then performed with default setting using Vanilla normalization[72]. This normalization is a variation of the ICE[160] method, in which a single iteration is performed dividing each element by the sum of counts in its row times the sum of counts in its column. At this point, read pairs are binned at a specified resolution, such as 50kb or 100kb for TADs and compartments detection, respectively. Biological replicates were also merged into cases and controls, respectively. On the merged data, normalized interaction matrices were generated for each biological replicate at 50kb and 100kb resolution for each chromosome, and at higher resolution (25, 20, 10 kb) for specific loci of interest. On this data, a quantitative comparative analysis was performed with CHESS[158] using default settings and tuning noise/signal score when necessary, according to visual inspection of the Hi-C maps produced. This tool relies on the concept of the structural similarity index (SSIM) (frequently used in image analysis) to Hi-C matrices, assigning a structural similarity score and an associated P value to pairs of genomic regions.

RESULTS

Forty-eight samples from patients affected by CJD MM1 and VV2, without any co-pathology, with a detailed clinical description of the disease course, with age of onset of the disease as similar as possible to the average (~60 years) and with as much clinical and pathologic information as possible were selected. Demographic and clinical information on the dataset is summarized in table 4. Males and females are equally distributed in the cohort and in the two subgroups, the average age of onset in the MM1 group was 65.8 years (σ = 7.6 years) and in the VV2 was 65.7 (σ = 8.5 years). Disease duration in the MM1 group was 3.1 (σ = 1.8 months) and in the VV2 6.8 months (σ = 2.4 months).

Sample	Strain	Disease Duration	Age of Onset	Co-pathology	Sex
#1	VV2	13	64	A-beta 0, tau 0	F
#2	VV2	6	57	A-beta 2, tau +	F
#3	VV2	7	78	A-beta 0, tau 1	M
#4	VV2	6	49	A-beta 0, tau 0	F
#5	VV2	5	75	A-beta 3, tau 0	M
#6	VV2	5	58	A-beta 0, tau 0	F
#7	VV2	6	79	A-beta 3, tau 1	F
#8	VV2	6	63	A-beta 0, tau 0	M
#9	VV2	5	60	A-beta 1, tau 1	M
#10	VV2	8	61	A-beta 2, tau 0, CAA	F
#11	VV2	10	71	A-beta 3, tau +	F
#12	VV2	4	78	A-beta 0, tau II	F
#13	VV2	5	49	A-beta 0, tau 0	F
#14	VV2	7,5	66	A-beta 0, tau 0	M
#15	VV2	6	74	A-beta 1, tau 1	M
#16	VV2	7	71	A-beta 0, tau +	F
#17	VV2	9	64	NA	F
#18	VV2	6	70	AGD, A-Beta neg	M
#19	VV2	9	65	A-beta 0, tau 1	F

#20	VV2	5,2	72	A-beta 3, tau +	M
#21	VV2	2,6	64	A-beta 0, tau ARTAG nel 13	M
#22	VV2	*14*	68	Neg	M
#23	VV2	5	61	NA	F
#24	VV2	7	59	A-beta 0, tau 0	F
#25	MM1	2	59	A-beta 0, tau 0	F
#26	MM1	10	70	A-beta 0, tau 0	F
#27	MM1	9	62	A-beta 0, tau +	F
#28	MM1	3	72	A-beta 0, tau 0	M
#29	MM1	2	59	A-beta 0, tau +	M
#30	MM1	5,5	64	A-beta 3, tau +	F
#31	MM1	1,5	69	A-beta 0, tau +	M
#32	MM1	2,5	76	A-beta 1, tau +	M
#33	MM1	4,5	68	A-beta 1, tau +	F
#34	MM1	3	62	A-beta 0, tau 1	M
#35	MM1	1	67	A-beta 0, tau +	M
#36	MM1	2	66	A-beta 2, tau 1	M
#37	MM1	2,5	65	A-beta 0, tau 0	M
#38	MM1	1,5	68	A-beta 1, tau 1	M
#39	MM1	3,5	43	Neg	M
#40	MM1	1,5	65	A-beta neg, tau 1	F
#41	MM1	1,5	67	A-beta 1, tau +	M
#42	MM1	1	61	A-beta 2, tau +	M
#43	MM1	2	74	A-beta 1, tau +	M
#44	MM1	2,5	63	A-beta neg, tau +	F
#45	MM1	1,5	67	A-Beta 1a, tau+	M
#46	MM1	3	57	Neg.	M
#47	MM1	3,1	74	A-beta 0, tau 0	M
#48	MM1	5,1	80	A-beta 1a, tau 1	M

Table 4: Clinical information on the dataset used in the genomic layer. For each sample, are reported strain type, disease duration expressed in months, age of onset of the disease (years), markers of co-pathology and biological sex.

GENOMICS: DNA TARGET SEQUENCING

Secondary analysis of target sequencing data provided information about single nucleotide polymorphisms (SNPs) and small indel variants stored in VCF files for the 48 samples analyzed. On these data, variant annotation, statistical analysis, and data science approaches were performed. The average mean coverage at a sample level in this dataset is 109.25X (σ = 46). After applying filters described in the materials and methods section, 57’005 different variants were identified in the 118 analyzed genes. In the MM1 group 42’169 different variants were identified, while in the VV2 47’594. Variants distribution on the 118 genes analyzed was compared between the two strains, showing no significant difference between conditions. On average, each sample carried 14’369 variants (σ =823). In this case as well no significant difference in terms of average number of variants was present in MM1 and VV2 patients, even though MM1 strain samples showed a higher homogeneity in this regard compared to VV2 samples (MM1: 14366 σ = 528, VV2: 14373 σ = 1051).

o VARIANT ANNOTATION

Variants were annotated with Base Space Illumina Variant Interpreter. Variants of interest were selected based on their consequence and predicted effect. In this dataset, 35 different missense variants were predicted as “probably/likely pathogenic” or “deleterious” by both Polyphen2 and SIFT predictors. These variants, reported in table 5, involve 26 genes. Seven variants were found more than once in the dataset, for a total of 62 findings of likely deleterious variants (complete list in appendix/supplementary), equally distributed between the two strains (31 in MM1 and 31 in VV2). No difference was found in terms of number of samples carrying at least one putative damaging mutation between the two strains, since 17 MM1 and 16 VV2 cases carried at least one of such variants.

VID	**Gene**	**HGVSC**	**HGVSP**	**SIFT Prediction**	**PolyPhen2 Prediction**
1:17312787:A	*ATP13A2*	c.3472C>T	p.(Arg1158Cys)	deleterious	possibly damaging
1:207680070:T	*CR1*	c.313C>T	p.(Arg105Cys)	deleterious	probably damaging
1:207739203:T	*CR1*	c.2537C>T	p.(Ser846Phe)	deleterious	possibly damaging
1:20976976:A	*PINK1*	c.1538G>A	p.(Gly513Asp)	deleterious	probably damaging
11:108117787:T	*ATM*	c.998C>T	p.(Ser333Phe)	deleterious	probably damaging
11:2189817:C	*TH*	c.484T>G	p.(Phe162Val)	deleterious	possibly damaging
11:88027209:C	*CTSC*	c.1357A>G	p.(Ile453Val)	deleterious	possibly damaging
14:92905737:C	*SLC24A4*	c.377T>C	p.(Leu126Pro)	deleterious	probably damaging
14:93119136:A	*RIN3*	c.1742G>A	p.(Arg581Gln)	deleterious	probably damaging
14:93142861:C	*RIN3*	c.2377T>C	p.(Tyr793His)	deleterious	possibly damaging
15:62269347:C	*VPS13C*	c.2342T>G	p.(Leu781Trp)	deleterious	probably damaging
15:89865073:C	*POLG*	c.2492A>G	p.(Tyr831Cys)	deleterious	possibly damaging
17:42427095:A	*GRN*	c.325G>A	p.(Gly109Arg)	deleterious	probably damaging
19:1046239:G	*ABCA7*	c.1456C>G	p.(Pro486Ala)	deleterious	probably damaging
19:1047336:A	*ABCA7*	c.2026G>A	p.(Ala676Thr)	deleterious	probably damaging
19:1058635:T	*ABCA7*	c.5168C>T	p.(Ser1723Leu)	deleterious	probably damaging
19:15273335:T	*NOTCH3*	c.5854G>A	p.(Val1952Met)	deleterious	probably damaging
19:15289863:A	*NOTCH3*	c.3691C>T	p.(Arg1231Cys)	deleterious	possibly damaging
19:15290917:A	*NOTCH3*	c.3293C>T	p.(Thr1098Ile)	deleterious	possibly damaging
19:45375208:T	*NECTIN2*	c.577C>T	p.(Arg193Trp)	deleterious	probably damaging
2:202587783:T	*ALS2*	c.3685T>A	p.(Trp1229Arg)	deleterious	possibly damaging
2:74759825:A	*HTRA2*	c.1195G>A	p.(Gly399Ser)	deleterious	probably damaging
20:3888719:A	*PANK2*	c.775G>A	p.(Gly259Arg)	deleterious	probably damaging
22:38539240:A	*PLA2G6*	c.481C>T	p.(Arg161Cys)	deleterious	probably damaging
4:170398474:C	*NEK1*	c.2235T>G	p.(Asn745Lys)	deleterious	probably damaging
5:126158560:T	*LMNB1*	c.1474G>T	p.(Ala492Ser)	deleterious	possibly damaging
6:161781201:A	*PRKN*	c.1204C>T	p.(Arg402Cys)	deleterious	probably damaging
7:100016781:C	*ZCWPW1*	c.314A>G	p.(Glu105Gly)	deleterious	probably damaging
7:143088584:T	*EPHA1*	c.2897G>A	p.(Arg966His)	deleterious	probably damaging
7:143092269:A	*EPHA1*	c.2090C>T	p.(Pro697Leu)	deleterious	possibly damaging
7:37923923:C	*NME8*	c.1013T>C	p.(Ile338Thr)	deleterious	probably damaging
7:37924854:A	*NME8*	c.1247G>A	p.(Ser416Asn)	deleterious	probably damaging
7:37936557:A	*NME8*	c.1630G>A	p.(Ala544Thr)	deleterious	probably damaging
9:132580901:G	*TOR1A*	c.646G>C	p.(Asp216His)	deleterious	possibly damaging
9:135202325:C	*SETX*	c.4660T>G	p.(Cys1554Gly)	deleterious	probably damaging

Table 5: list of 35 probably damaging missense variants identified in this sporadic cohort

To acquire more biological insights about this gene list, functional enrichment performed with over representation methods on this gene list showed an enrichment in pathways involved in stress response, like regulation of mitochondrial autophagy and mitochondrial depolarization, response to oxidative stress, selective autophagy and lysosomal transport, together with fewer pathways involved in dopaminergic synaptic transmission.

Of the 26 genes involved in these possibly damaging variants, seven are affected in both MM1 and VV2 patients (*CR1*, *EPHA1*, *NME8*, *NOTCH3*, *POLG*, *RIN3*, *TOR1A),* while nine genes were affected only in MM1 samples *(ABCA7, ALS2, ATP13A2, CTSC, PANK2, SETX, SLC24A4, TH, VPS13C*) and other ten only in VV2 samples (*ATM, GRN, HTRA2*, *LMNB1*, *NECTIN2, NEK1*, *PINK1*, *PLA2G6, PRKN, ZCWPW1*) (complete name of protein coded by each gene in appendix). Therefore, even though in terms of number of likely pathogenic variants no difference is visible between strains, specific genes seem to be affected exclusively in one strain and not in the other. As will be discussed in the following chapter, the ten genes affected only in the VV2 cohort include important genes involved in energetic metabolism and mitophagy that are also strongly associated to Parkinson Disease.

- STATISTICAL ANALYSIS

Allele frequencies were compared between the two considered strains and between the whole sCJD dataset and the European population (GnomAD database) with Fisher exact test and Benjamini-Hochberg multiple test correction. Comparing MM1 and VV2 strains, four variants showed a significantly different allele frequency between the two classes (Table 6). Notably, one of these variants (rs1799990) is the codon 129, which was used as a criterion to distinguish the two strains, and the remaining three intronic SNV are as well in the *PRNP* intronic region and are in linkage disequilibrium with codon 129. Therefore, these three variants carry the same information, that was expected given the experimental design.

ID variant	Allele freq. MM1	Allele freq. VV2	p-value	p-adj	Annotation	Rs Id
chr20-4671225-T-G	0,5	1	1,2E-07	0,002	c.-11+3843T>G	rs6052769
chr20-4667829-T-C	0,5	1	2,2E-06	0,027	c.-11+447T>C	rs35519959
chr20-4672816-A-G	0,5	1	4,2E-08	0,001	c.-11+5434A>G	rs6052771
chr20-4680251-A-G *	0,5	1	2,5E-09	0	c.385A>G	rs1799990

Table 6: Single nucleotide variants that showed a significantly different allele frequency in the two considered strains.

The same comparative analysis was performed considering the whole dataset of sCJD patients versus the allele frequencies referring to the European population reported in the GnomAD database. In this comparison, 238 variants distributed in 37 different genes showed a significantly different allele frequency between sCJD and reference European population ($p\text{-adj} < 0.05$). Functional analysis with over representation methods on these 37 genes showed an enrichment of chaperone binding functions. In terms of altered biological processes, cellular component maintenance and specifically synapse organization maintenance showed the lowest p-adj values, followed by regulation of cell migration and motility and regulation of protein stability. These results are reported in Table 7.

Term name	Term ID	padj
chaperone binding	GO:0051087	1.945×10-2
cellular component maintenance	GO:0043954	3.276×10-4
regulation of synapse organization	GO:0050807	5.839×10-3
positive regulation of cell migration	GO:0030335	1.975×10-2
positive regulation of cell motility	GO:2000147	2.668×10-2
regulation of calcium ion transmembrane transport	GO:1903169	3.065×10-2
regulation of protein stability	GO:0031647	4.403×10-2

Table 7: Single nucleotide variants that showed a significantly different allele frequency in the two considered strains.

Three of the overrepresented variants in the sCJD cohort were also predicted as probably pathogenic by SIFT and PolyPhen2 variant predictors: *GRN* p.Gly109Arg (p.adj = 0.02, sample #17, VV2), *NME8* p.Ser416Asn (p.adj = 0.03, sample #40, MM1) and *RIN3* Arg581Gln (p.adj = 0.03, sample #8, VV2). For none of these three variants is available a clinical significance value on ClinVar. Each one of these variants was found in only one sample in the analyzed cohort, thus a validation in a larger cohort would be necessary to confirm this preliminary finding.

- DATA SCIENCE: UNSUPERVISED MACHINE LEARING

In this context, the aim of the unsupervised analysis is to highlight possible recurrent genetic patterns that characterize the two strains that not necessarily involve coding variants. The binary transformation of the complete genetic information contained in VCF files led to the generation of a matrix with shape 57005×48 in which each row represents a variant found at least in one sample and each column stands for each sample. In the matrix, 0 indicates that the variant is not present in the VCF file of the patient whereas 1 indicates its presence in heterozygosity and 2 the presence of the variant in homozygosity. This matrix was used as input for machine learning methods. It is expected that most variants will account for the normal genetic differences that exist between different individuals and for shared genetic traits of Caucasians, thus it is assumed that a large number of variants will not cause significant functional variation. For these reasons, both linear and non-linear dimension reduction algorithms (such as PCA and t-SNE or UMAP, respectively) were tested on the matrix containing the whole genetic information, to explore possible patients or SNPs that cluster together. The first exploratory analysis was performed trough dimension reduction with Principal Component Analysis (PCA) (figure 15). In the plot each dot represents a sample. The first eighteen principal components explain 50% of the overall variance of the dataset (supplementary, table S3). PC1's main contributor is a variant in the genomic region of to the gene E3 ubiquitin-protein ligase parkin (*PARK2/PARKN*, chr6-163069504-G-A), the main contributor to the second principal component is a SNV in genomic region of the gene Phosphatidylinositol-binding clathrin assembly protein (*PICALM*, chr11-85710180-G-A).

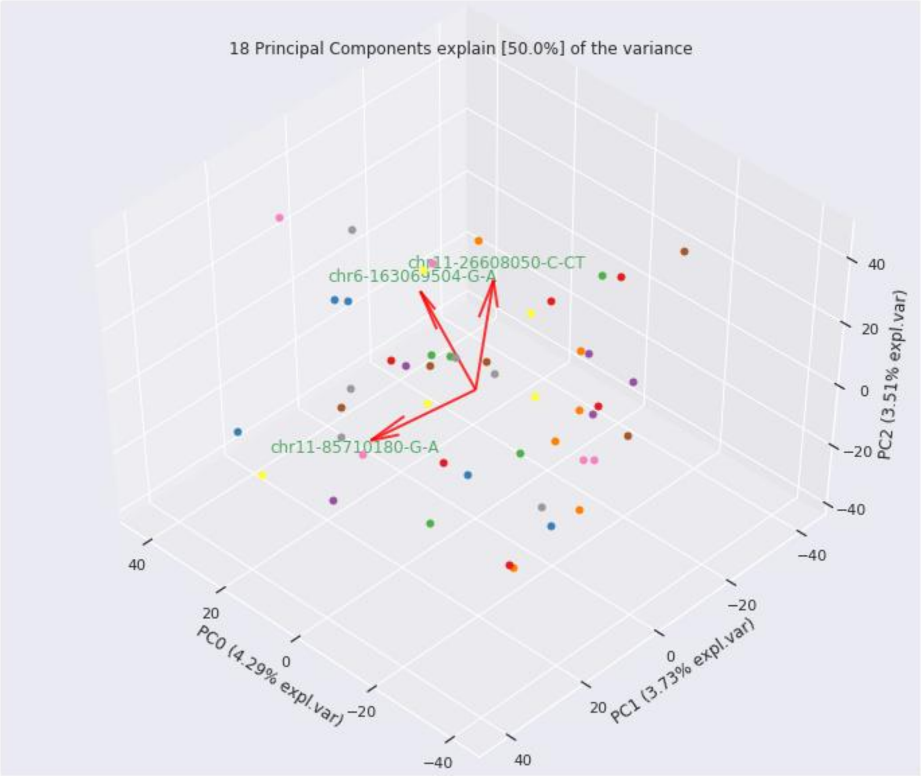

Figure 15: 3D plot of PCA of the 47005x48 matrix. In the plot, each point represents a sample. The three axes x, y, z represents the first three principal components, that contribute to the 4.29%, 3.73 and 3.51% of the explained variance, respectively. Red arrows represent the top contributing variants of the first three PC, harboured in intronic regions of the genes *PRKN, PICALM* and *ANO3*, respectively.

On the PCA plot (figure 12), no clear clusters were evident by eye. To check if the sample distribution based on the overall patient's genetic background matched biological features, labels matching sex, geographical origin of the patient and strain were combined with the bidimensional PCA plot, showing no specific clusterization based on the most likely confounder and the biological feature object of this study. On this matrix, two different unsupervised clustering methods were applied: hierarchical clustering with dendrograms and k-means (n=2). Hierarchical clustering did not identify two clusters, since it divided one sample from the rest of the dataset, therefore highlighting a possible outlier. K-means on the contrary, identified two numerically similar classes, separating samples along the first principal component (Figure 16), that as previously said had *PRKN* as main contributor.

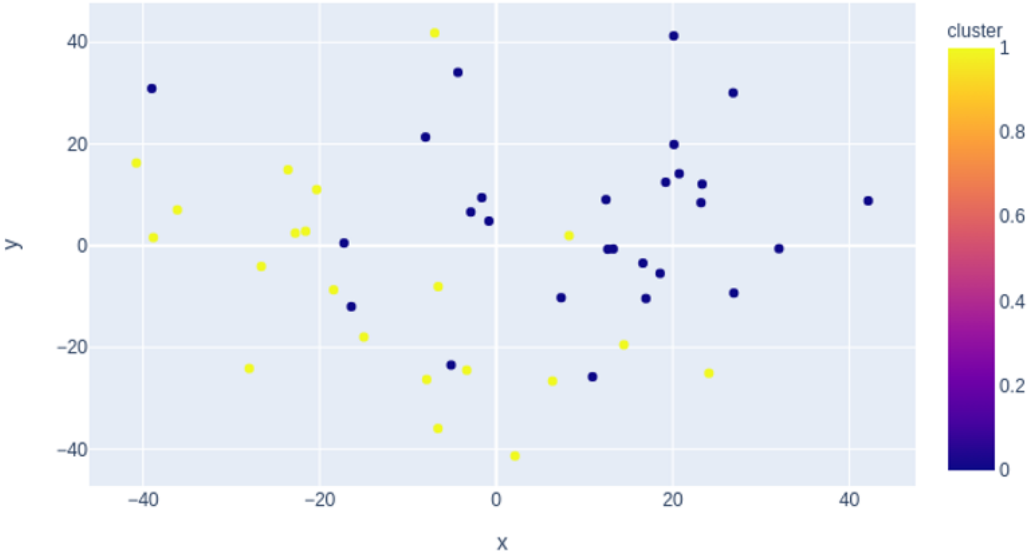

Figure 16: Plot of the K-Means clustering on the PCA plot. K-Means recognizes two clusters mainly distributed along the first principal component of the PCA plot.

Unfortunately, the two clusters did not match the strain type nor any of the available biological or clinical feature available for this dataset. Comments on these results are reported in the discussion chapter.

- DATA SCIENCE: SUPERVISED MACHINE LEARING

Supervised classifiers were used for automatic recognition of genetic patterns among the 57005 variants identified in this dataset. Decision trees have been previously used in clinical genomics and precision medicine applications to interpret the role of genetic variants in complex diseases[161,162]. In the 57005×48 matrix, to each sample a label corresponding to the strain (class: "MM1" or "VV2") was added. The classification was achieved perfectly, with 100% of accuracy (ratio of correctly predicted observation to the total observations) on the test set, basing the classification on the codon 129 (chr20-4680251-A/G). To test for other recurrent genetic patterns that could characterize the two phenotypic groups, we removed from the input data given to the classifier the row of the matrix indicating codon 129. As expected, accuracy decreased both in training set and in test set, but interestingly the classifier managed to distinguish the two diseases with a good accuracy (training = 1, test 0.81, table 8).

	Precision	**Recall**	**F1**	**Support**
MM1	0.73	1	0.84	8
VV2	1	0.62	0.77	8

Table 8: Classification metrics of decision trees algorithm. Precision is the ratio of correctly predicted observation to the total predicted positive observations (True Positive / True Positive + False Positive), Recall is the ratio of correctly predicted positive observations to all observations in actual class (True Positive / True Positive + False Negative), F1 Score is the harmonic mean of Precision and Recall (F1 Score = 2*(Recall * Precision) / (Recall + Precision)). Support indicates class numerosity.

The classification is based on two intronic variants, in *PRNP* and *FERMT2* genes (figure 17). According to common databases and genomic search engines such as VarSome44, OMIM45, ClinVar46 or HGMD47, the intronic variant in *FERMT2* was never previously reported as functional intronic variant. The intronic variant in *PRNP* is also referred to with the ID rs6037932, this variant was used in a phylogenetic study about founder effect in another prion disease (FFI) in 2008[163]. This SNP is in the intronic region between exon 1 and 2, 5kb away from codon 129. This SNP is not in complete linkage disequilibrium with the allele 129V, even though in this cohort as well as in the previously cited work it is more frequently associate to it. On the other hand, the

FERMT2 variant Chr14-53391236-T-G in this cohort is always associate to the allele 129V. These results will be discussed in the following chapter.

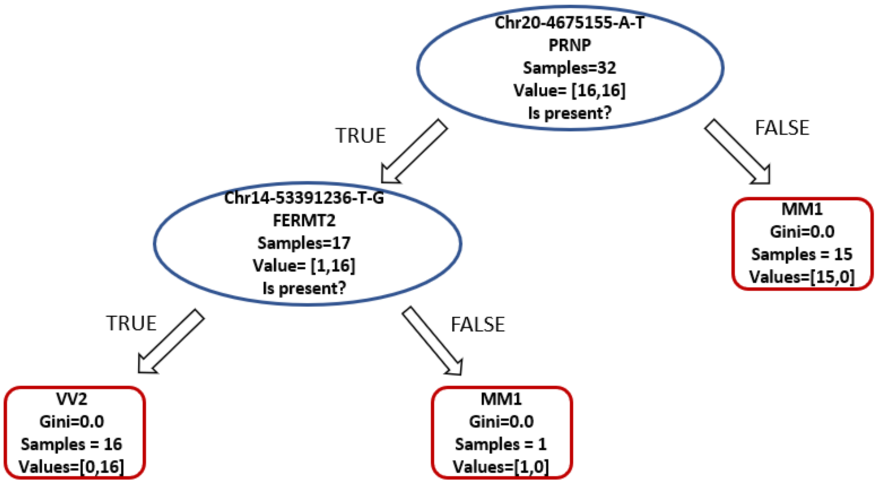

Figure 17: Decision tree graphical representation. The classification was performed using two variants as parameters. First, a sample did not carry the SNV A/T in position 4675155 of the *PRNP* gene, it was assigned to the MM1 class. Afterwards, if a sample carried that variant and also the SNV T/G in position 53391236 in the *FERMT2*, it was assigned to the class VV2.

TRANSCRIPTOMICS: RNA SEQUENCING

Total RNAs were extracted from frontal cortex of frozen postmortem brain tissue. RNA degradation was evaluated with Fragment Analyzer to acquire in each sample the percentage of RNA molecules longer than 200 nucleotides (DV_{200}). To be able to perform RNA sequencing with the chosen library preparation protocol, this parameter needs to be higher than 70% to consider the input material of high quality, 50-70% as medium quality 30-50% for low quality and under 30% results cannot be considered as reliable. All samples had a DV_{200} value higher than 70%, with an average DV_{200} in this dataset of 84.5% ($\sigma = 7.2$), as showed in table 9.

Sample ID	Strain	DV_{200}
#32	MM1	76.8
#33	MM1	77.5
#36	MM1	86.7
#40	MM1	73
#43	MM1	90.9
#44	MM1	85
#45	MM1	87.9
#46	MM1	83.8

Sample ID	Strain	DV_{200}
#2	VV2	90.6
#13	VV2	95.1
#15	VV2	90.9
#12	VV2	88.6
#14	VV2	69.8
#18	VV2	89.4
#17	VV2	79
#22	VV2	86.4

Table 9: Total RNA quality from extraction.

All intermediate quality checks in library preparation were satisfied and quality sequencing metrics were in the range recommended by Illumina (Recommended values: Cluster density =170-220 K/mm^2, Cluster passing filter: > 80%). During secondary analysis, uniquely mapped reads were quantified, with an average value of 81.39% ± 2.6 (70-90% range of optimal experimental data, ~63000 different transcripts). Sequencing depth was also quantified in each sample, on average 35.27M (Million reads) ± 13.5M (25M optimal according to Illumina guidelines to acquire a complete picture of also lowly expressed transcripts). In the whole dataset, more than 63’000 transcripts were annotated and quantified.

- DIFFERENTIAL GENE EXPRESSION ANALYSIS

Comparative analysis of transcriptomic profiles between the two strains was performed with DeSeq2. Based on the sample-level quality checks, the strongest covariates in the dataset were biological sex, experimental batch, post-mortem conservation time of the tissue before deep freezing, and disease duration. From DGE analysis, 1798 differentially expressed transcripts were identified in the comparison between VV2 and MM1, where MM1 are used as baseline in the contrast. Among them, 1196 transcripts are significantly over expressed in VV2 compared to MM1, while 602 are significantly under expressed in VV2 compared to MM1 (that is over expressed in MM1 compared to VV2, Figure 18). Differential expression of the six transcripts with the most extreme LFC values were validated with ddPCR. (in appendix, top 20 DGE per strain)

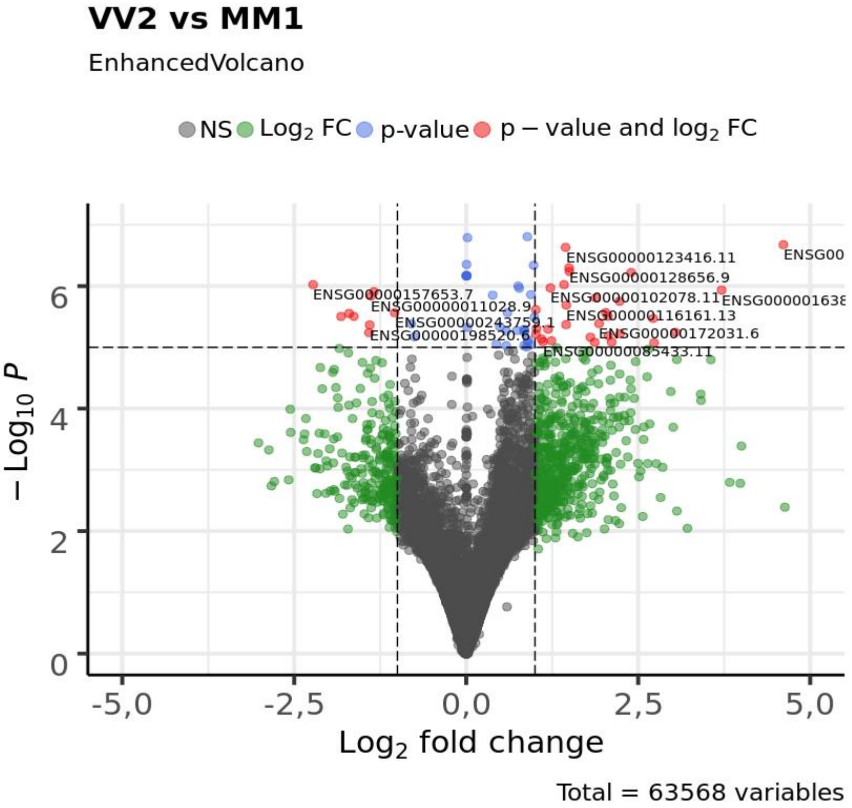

Figure 18: Volcano plot of the 63568 transcripts used in the DGE analysis. In grey are shown transcripts that do not vary between conditions, in green are highlighted the transcripts that show a relevant differential expression in terms of LFC (|ΔLFC| ≥ 1), in blue transcripts that show a statistically significant variation (Padj<0.05 after Wald test and BH correction) but low ΔLFC, while in red transcripts that exceed both thresholds.

Based on the significant differentially expressed genes, a heatmap with hierarchical clustering of the expression profiles is produced (Figure 19). In the heatmap, Z-scores are computed subtracting the mean expression value and dividing by the standard deviation. The Z-scores are computed after the clustering, therefore they affect only the graphical aesthetics, while the clustering is performed on normalized counts. Hierarchical clustering shows a good, even though not perfect, identification of the two strains based on the results of the differential gene expression analysis, with a smaller cluster made of five VV2 samples and one MM1 sample, and a larger cluster made of ten samples, seven of which are MM1 and three are VV2.

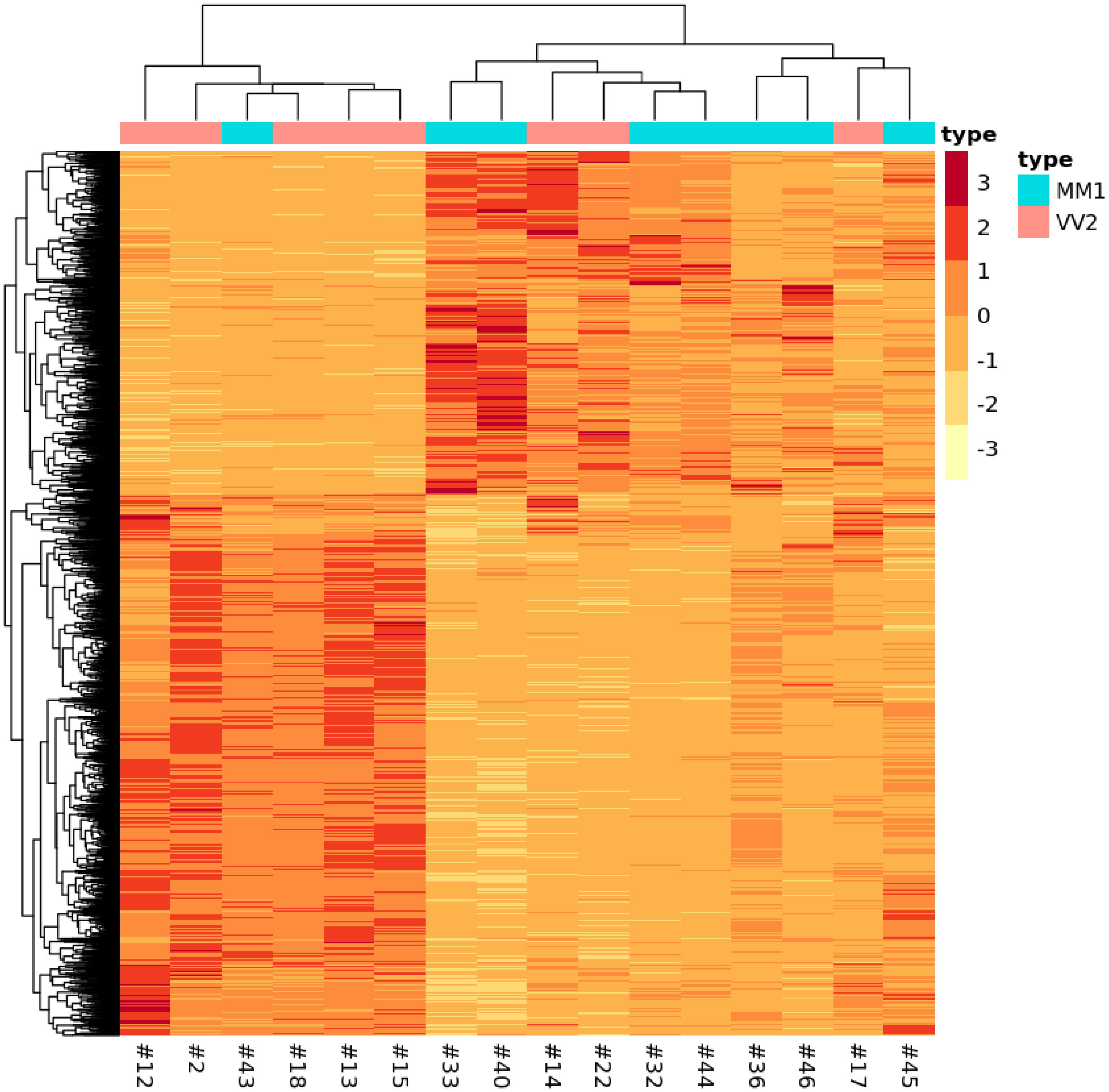

Figure 19: Heatmap of the 1798 differentially expressed genes in VV2 strain compared to MM1. Hierarchical clustering identified two clusters, one enriched in MM1 samples (identified by a light blue label) and the other with VV2 more represented (pink label).

- FUNCTIONAL ENRICHMENT

To gain functional insights about the consequences of differentially expressed genes, functional enrichment was performed with three different computational methods: over representation analysis, gene set enrichment analysis and protein-protein interaction networks. These methods, as explained in the introduction and in the materials and methods chapters, are based on different assumption and tests, nevertheless they provided shared or coherent results, expanding and reinforcing one another.

1. *OVER REPRESENTATION ANALYSIS*

This method uses only a subset of genes to find significantly enriched pathways and functional modules between a list of "significant" inputs. For this purpose, significantly over expressed genes in MM1 (p.adj <0.05, LFC < 0) and over expressed genes in VV2 (p.adj <0.05, LFC > 0) were considered separately. Twenty-one pathways showed a significant enrichment based on the MM1 overexpressed genes list, in table 10 and figure 20 are reported the top 10 results (p.adj <0.05).

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue
GO:0043087	regulation of GTPase activity	29/469	457/19179	3.11E+08	0.004	0.003477
GO:0043547	positive regulation of GTPase activity	26/469	385/19179	3.31E+08	0.004	0.003477
GO:0016054	organic acid catabolic process	22/469	294/19179	3.71E+08	0.004	0.003477
GO:0046395	carboxylic acid catabolic process	22/469	294/19179	3.71E+08	0.004	0.003477
GO:0046322	negative regulation of fatty acid oxidation	5/469	12/19179	5.88E+08	0.005	0.004414
GO:0008360	regulation of cell shape	15/469	160/19179	9.37E+08	0.006	0.005773
GO:0051056	regulation of small GTPase mediated signal transduction	23/469	338/19179	1.08E+09	0.006	0.005773
GO:0007265	Ras protein signal transduction	24/469	366/19179	1.29E+09	0.007	0.006045
GO:0040001	establishment of mitotic spindle localization	7/469	35/19179	1.86E+09	0.007	0.006771
GO:0030198	extracellular matrix organization	26/469	425/19179	1.91E+09	0.007	0.006771

Table 10: Top biological processes over expressed in the MM1 strain compared to VV2 samples.

The most affected pathways were regulatory pathways mediated by guanosine triphosphatases (GTPases), regulation of catabolic processes and maintenance of proper cell morphology and matrix organization. In the context of the comparative analysis object of this work, this means a significant enrichment in MM1 samples compared to VV2 mediated by the 543 genes with negative LFC identified from DGE analysis.

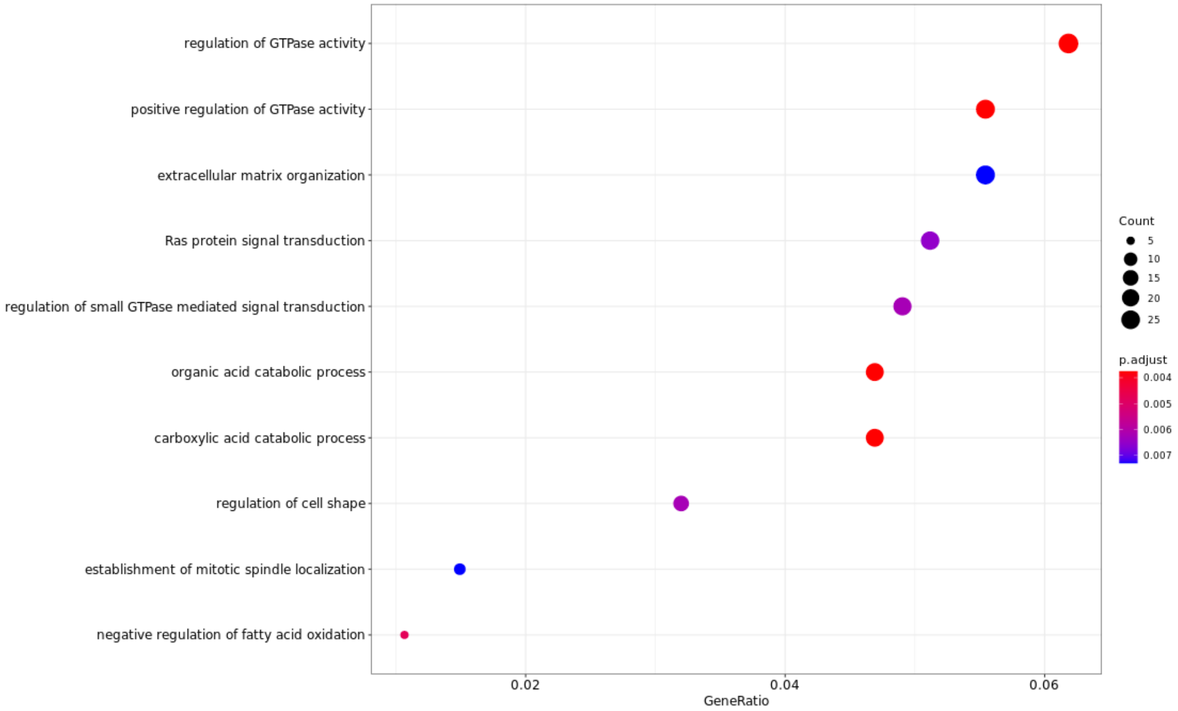

Figure 20: Dotplot of top 10 enriched pathways in MM1 strain result of the 602 over expressed genes in this group.

Considering now the VV2 group, based on the list of 1153 significantly overexpressed genes, 634 biological processes showed an enrichment, the top ten results are reported in table 11:

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue
GO:0050804	modulation of chemical synaptic transmission	91/1059	473/19179	3.43E-12	1.03E-08	8.44E-09
GO:0099177	regulation of trans-synaptic signaling	91/1059	474/19179	4.03E-12	1.03E-08	8.44E-09
GO:0099003	vesicle-mediated transport in synapse	55/1059	219/19179	6.96E-08	1.19E-04	9.74E-05
GO:0050808	synapse organization	79/1059	445/19179	1.37E-06	1.75E-03	1.43E-04
GO:0099504	synaptic vesicle cycle	49/1059	198/19179	2.43E-05	2.49E-02	2.04E-02
GO:0042391	regulation of membrane potential	77/1059	454/19179	6.93E-06	5.92E-02	4.85E-02
GO:1904062	regulation of cation transmembrane transport	62/1059	343/19179	8.63E-03	6.32E+00	5.17E-01
GO:0016079	synaptic vesicle exocytosis	35/1059	121/19179	1.79E-02	1.15E+01	9.41E+00
GO:0034765	regulation of ion transmembrane transport	76/1059	490/19179	2.11E-02	1.20E+01	9.83E-01
GO:0098693	regulation of synaptic vesicle cycle	33/1059	112/19179	7.15E-02	3.66E+01	3.00E+01

Table 11: Top biological processes over expressed in the VV2 strain.

To highlight the functional modules starting from hundreds of interconnected and overlapping pathways, Cytoscape was used to summarize non redundant functional modules and visualize interconnections between pathways. Results are reported in figure 21 and table 12.

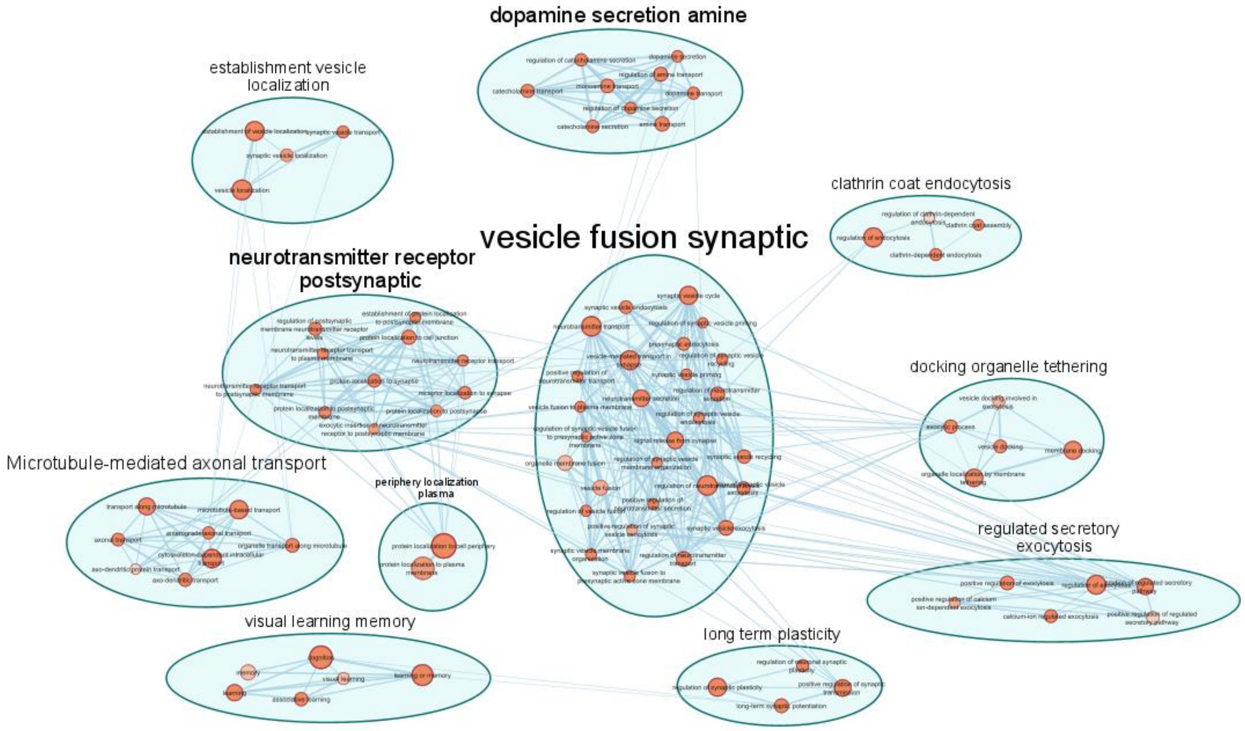
dopamine secretion amine
regulation of catecholamine secretion
dopamine secretion
regulation of amine transport
catecholamine transport
monoamine transport
dopamine transport
regulation of dopamine secretion
catecholamine secretion
amine transport
establishment vesicle localization
establishment of vesicle localization
synaptic vesicle transport
synaptic vesicle localization
vesicle localization
clathrin coat endocytosis
regulation of clathrin-dependent endocytosis
clathrin coat assembly
regulation of endocytosis
clathrin-dependent endocytosis
vesicle fusion synaptic
synaptic vesicle cycle
synaptic vesicle endocytosis
neurotransmitter transport
regulation of synaptic vesicle priming
presynaptic endocytosis
vesicle-mediated transport in synapse
regulation of synaptic vesicle recycling
positive regulation of neurotransmitter transport
synaptic vesicle priming
regulation of neurotransmitter secretion
neurotransmitter secretion
vesicle fusion to plasma membrane
regulation of synaptic vesicle endocytosis
regulation of synaptic vesicle fusion to presynaptic active zone membrane
signal release from synapse
organelle membrane fusion
regulation of synaptic vesicle membrane organization
synaptic vesicle recycling
vesicle fusion
regulation of neurotransmitter levels
regulation of synaptic vesicle exocytosis
positive regulation of neurotransmitter secretion
regulation of vesicle fusion
positive regulation of synaptic vesicle exocytosis
synaptic vesicle exocytosis
synaptic vesicle membrane organization
regulation of neurotransmitter transport
synaptic vesicle fusion to presynaptic active zone membrane
neurotransmitter receptor postsynaptic
establishment of protein localization to postsynaptic membrane
regulation of postsynaptic membrane neurotransmitter receptor levels
protein localization to cell junction
neurotransmitter receptor transport to plasma membrane
neurotransmitter receptor transport
neurotransmitter receptor transport to postsynaptic membrane
protein localization to synapse
receptor localization to synapse
protein localization to postsynaptic membrane
protein localization to postsynapse
exocytic insertion of neurotransmitter receptor to postsynaptic membrane
docking organelle tethering
vesicle docking involved in exocytosis
exocytic process
vesicle docking
membrane docking
organelle localization by membrane tethering
Microtubule-mediated axonal transport
transport along microtubule
microtubule-based transport
axonal transport
anterograde axonal transport
organelle transport along microtubule
cytoskeleton-dependent intracellular transport
axo-dendritic protein transport
axo-dendritic transport
periphery localization plasma
protein localization to cell periphery
protein localization to plasma membrane
regulated secretory exocytosis
positive regulation of exocytosis
regulation of exocytosis
regulation of regulated secretory pathway
positive regulation of calcium ion-dependent exocytosis
calcium-ion regulated exocytosis
positive regulation of regulated secretory pathway
visual learning memory
cognition
memory
visual learning
learning or memory
learning
associative learning
long term plasticity
regulation of neuronal synaptic plasticity
regulation of synaptic plasticity
positive regulation of synaptic transmission
long-term synaptic potentiation

Figure 21: Region of the network representing enriched pathways grouped by non-redundant functional modules in the VV2 group. Most of the biological processes can be grouped into the functional module of synaptic regulation and vesicle trafficking. Notably, in the upper part of the network a cluster representing pathways involved in dopamine secretion is represented.

Cluster	Nodes	Cluster	Nodes
vesicle fusion synaptic	28	regulation dendrite morphogenesis	4
activity transmembrane transporter	15	response metal substance	4
assembly synapse pre-synapse	12	central nervous neuron	3
muscle contraction cardiac	11	chemical postsynaptic excitatory	3
neurotransmitter receptor postsynaptic	11	regulation depolarization potential	3
dopamine secretion amine	9	amino acid starvation	2
membrane mitochondrial permeability	9	peptide hormone insulin	2
Microtubule-mediated axonal transport	8	periphery localization plasma	2
anion chloride transmembrane	7	RAC signal transduction	2
cellular response stimulus	7	regulation macro autophagy	2
regulation pH reduction	7	negative regulation cell	1
developmental growth extension	6	negative regulation microtubule	1
regulated secretory exocytosis	6	neurofilament cytoskeleton organization	1
visual learning memory	6	neuron apoptotic process	1

adult walking behaviour	5	neuron recognition	1
cytosol sarcoplasmic reticulum	5	post Golgi vesicle	1
docking organelle tethering	5	protein folding	1
positive protein intracellular	5	protein homo-oligomerization	1
sodium ion transmembrane	5	regulation dephosphorylation	1
clathrin coat endocytosis	4	regulation proteasomal protein	1
dendritic spine organization	4	response temperature stimulus	1
establishment vesicle localization	4	spontaneous synaptic transmission	1
long term plasticity	4	synaptic transmission GABAergic	1
potassium ion transmembrane	4	synaptic transmission glutamatergic	1

Table 12: Functional modules that represent hubs of the whole network, combined with the number of nodes for each hub.

Pathway affecting synaptic regulation show the strongest impact both in term of statistical significance (Table 11) and in term of number of interconnected pathways that end up under synaptic-related the functional modules (Table 12 and Figure 21). In addition to the high number of nodes of synaptic-related modules, it is worth noticing that these clusters show the highest rate of interconnections with other hubs, underling their pivotal functional role. This means that even if the baseline of this comparative analysis is represented by another group of CJD samples (MM1), in the VV2 group these pathways are significantly enhanced.

2. *GENE SET ENRICHMENT ANALYSIS (GSEA)*

Gene set enrichment analysis assumes that also weaker but coordinated changes in sets of functionally related genes can have significant effects, therefore in this type of functional analysis all transcripts are considered. Despite being based on different assumption compared to the previous approach, results reported in table 13 show several similarities in terms of altered pathways in the considered comparison.

ID	Description	EnrichmentScore	pvalue	p.adjust	qvalues
GO:0010975	regulation of neuron projection development	0.5549697	0,0001	0,0067	0,0056
GO:0015672	monovalent inorganic cation transport	0.5643432	0,0001	0,0067	0,0056
GO:0007409	axonogenesis	0.5099560	0,0001	0,0067	0,0056
GO:0072507	divalent inorganic cation homeostasis	0.4489722	0,0001	0,0067	0,0056
GO:0034765	regulation of ion transmembrane transport	0.6005878	0,0001	0,0067	0,0056
GO:0050769	positive regulation of neurogenesis	0.4842083	0,0001	0,0067	0,0056
GO:0050804	modulation of chemical synaptic transmission	0.6784846	0,0001	0,0067	0,0056
GO:0099177	regulation of trans-synaptic signalling	0.6782517	0,0001	0,0067	0,0056
GO:0072511	divalent inorganic cation transport	0.4676002	0,0001	0,0067	0,0056
GO:0070838	divalent metal ion transport	0.4744645	0,0001	0,0067	0,0056

Table 13: Top biological processes enriched in the VV2 strain according to Gene Set Enrichment Analysis.

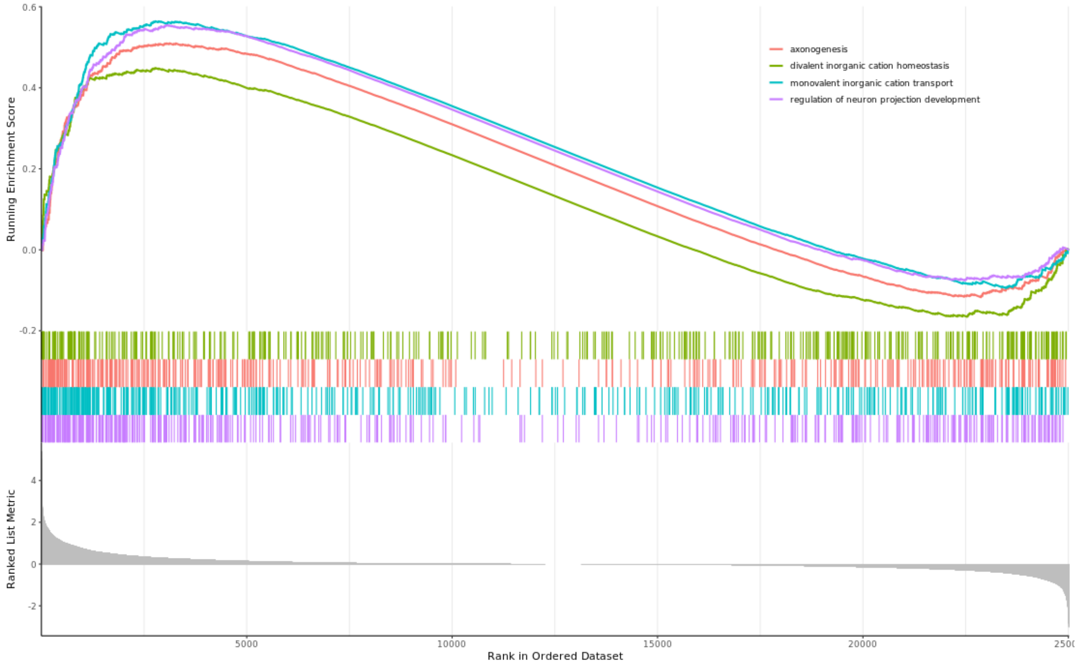

Figure 22: GSEA plot of the top four results of the analysis. In the upper park of the plot, the solid lines show the enrichments scores (ES), and peaks tells how over or under expressed each pathway is respect to the ranked list. The second part of the graph, (middle, with vertical bars) shows where genes related to each pathway are in the ordered ranking of input genes. The third part of the graph (bottom with grey curve) shows metric distribution in the input ranking list.

These results reinforce the previous findings of an increased regulation of synaptic functionality in the VV2 group compared to the MM1, both in structural terms affecting neuronal projection and axonogenesis, and in functional terms affecting synapse activity with altered regulation of mono and divalent cations transport. In addition, this analysis also provides information about the pathways affected by under expressed genes (that are, the overexpressed in the MM1 strain compared to the VV2), as shown in GSEA plot of figure 22 where a smaller but clear negative peak in enrichment score is visible also at the bottom of the ranked list in all top four results, indicating concordant changes in gene expression both in terms of overexpression of activating genes and under expression of negative regulator in those pathways.

3. PROTEIN-PROTEIN INTERACTION NETWORK

This last type of functional analysis approach provides information on physical interactions of proteins encoded by differentially expressed genes. Even though it is well known that alterations in mRNA quantity do not necessarily correlate with equal alteration of protein abundancy due to post transcriptional regulation, this analysis nevertheless provides important functional insights about the biological interplay of different proteins. In this analysis, only direct biophysical interactions were considered (i.e., molecular docking) mapped on STRING interactome and, based on these interactions, functional enrichment was performed on KEGG database. In table 14 are reported the top ten significant results from overexpressed genes in the VV2 strain, whereas in figure 23 enriched pathways are group in non-redundant functional modules.

Background Genes	Genes	Description	Fdr Value	P-Value
74,00	13,00	Synaptic vesicle cycle	4.0E-7	8.8E-10
355,00	21,00	Alzheimer disease	1.1E-4	4.96E-7
53,00	8,00	Endocrine and other factor-regulated calcium reabsorption	3.4E-4	4.49E-6
48,00	8,00	Vibrio cholerae infection	3.4E-4	2.32E-6
67,00	9,00	Epithelial cell signalling in Helicobacter pylori infection	3.4E-4	2.7E-6
208,00	14,00	cAMP signalling pathway	6.9E-4	1.06E-5
128,00	11,00	Dopaminergic synapse	6.9E-4	1.15E-5
46,00	7,00	Type II diabetes mellitus	8.6E-4	1.71E-5
102,00	9,00	C-type lectin receptor signalling pathway	0.0025	5.96E-5
135,00	10,00	Spinocerebellar ataxia	0.0034	9.25E-5

Table 14: functional enrichment of PPI using overexpressed genes from DGE analysis

Also with this approach, a different synapse regulation results to be the most evident pathway in this comparison between strain VV2 and MM1. Additionally, this analysis better underlines the different alteration of protein localization and of assembly of cellular components, as suggested by the high number of pathways that can be led back to these functional modules (figure 23). The decreased interconnection between functional modules of the network in figure 20 compared to the one reported for OR analysis is due to technical reasons, since the PPIn was restricted to the direct biophysical interactions coded by differentially expressed genes, avoiding the extension to neighbour interacting proteins as well.

PPIn provided few insights on the affected pathways by overexpressed genes in the MM1 group, summarised in table 15. Functional analysis of molecular pathways enriched through physical interaction between overexpressed protein coding genes in the MM1 group highlighted an impairment of the regulation of nucleotide binding mechanisms, which is coherent with the previously described result of a positive regulation of GTPase activity.

Cluster	**Nodes**
regulation nucleotide binding	11
egf domain extracellular	3
pleckstrin homology domain	1
sh3 domain	1
syndromic deafness	1

Table 15: summary of the most enriched functional modules through physical interaction between protein coded by overexpressed genes in the MM1 group.

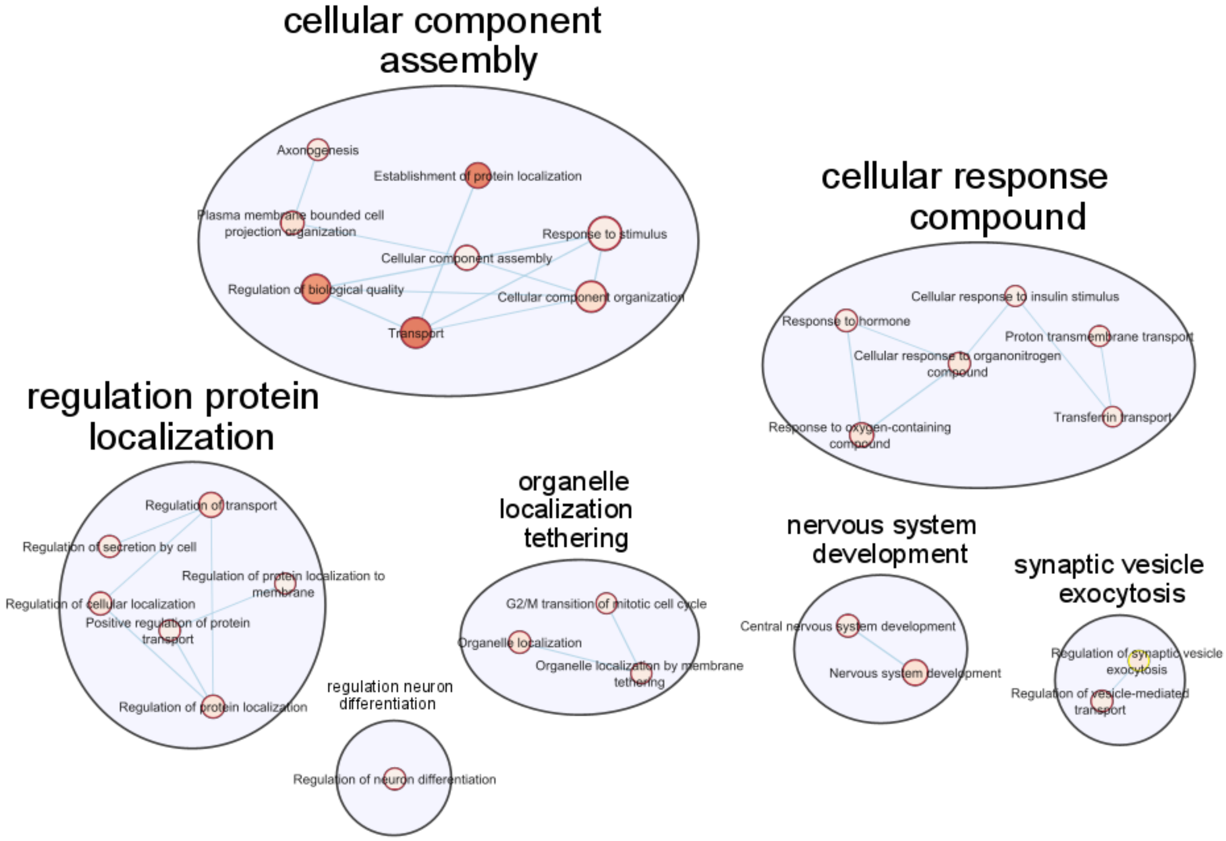
cellular component assembly
Axonogenesis
Establishment of protein localization
Plasma membrane bounded cell projection organization
Response to stimulus
Cellular component assembly
Regulation of biological quality
Cellular component organization
Transport
cellular response compound
Cellular response to insulin stimulus
Response to hormone
Proton transmembrane transport
Cellular response to organonitrogen compound
Response to oxygen-containing compound
Transferrin transport
regulation protein localization
Regulation of transport
Regulation of secretion by cell
Regulation of protein localization to membrane
Regulation of cellular localization
Positive regulation of protein transport
Regulation of protein localization
organelle localization tethering
G2/M transition of mitotic cell cycle
Organelle localization
Organelle localization by membrane tethering
nervous system development
Central nervous system development
Nervous system development
synaptic vesicle exocytosis
Regulation of synaptic vesicle exocytosis
Regulation of vesicle-mediated transport
regulation neuron differentiation
Regulation of neuron differentiation

Figure 23: Region of the network representing enriched pathways grouped by non-redundant functional modules from PPI analysis using overexpressed genes in VV2 from DGE analysis.

The results of the various parts of functional analysis will be thoroughly discussed in the discussion chapter.

EPIGENOMICS: HI-C

Four samples, two CJD cases and two healthy controls, were selected to undergo the Hi-C library preparation as described in the materials and methods section. Intermediate quality checks were passed for all samples, during the Hi-C protocol as well as in the library preparation for whole genome sequencing. Post alignment processing provided the metrics reported in table 16, that show the number of acquired reads, uniquely mapped and valid pairs for each biological replicate. After mapping, biological replicates of CJD and healthy controls were merged, leading to a controls group with 556’182127 valid interactions and a cases group with 609’179878 valid interactions.

Experiment	Reads	Uniquely Mapped	% Mapped	Valid int.	% Valid interactions
Control1 R1	480’457’513	366’629’918	80%	217’266’479	45%
Control1 R2	480’457’513	360’300’498	80%		
Control2 R1	446’163’413	382’716’166	86%	338’915’648	76%
Control2 R2	446’163’413	383’706’325	86%		
Case 1 R1	466’747’975	406’630’305	87%	331’791’787	71%
Case 1 R2	466’747’975	408’308’072	88%		
Case 2 R1	436’058’606	379’638’968	87%	277’388’091	64%
Case 2 R2	436’058’606	377’034’075	87%		

Table16: Summary of post alignment Hi-C metrics for each biological replicate.

Interaction matrices were produced at 100kb and 50kb resolution for each chromosome to visually inspect major structural features at a chromosome level in cases and controls. Literature research was performed to make a list of candidate genes known to play a role in the peripheral immune system in the early stages of CJD and other types of dementia, together with control genes in which no changes are expected (Table 17). The genomic locus corresponding to 1Mb ahead and after the open reading frame of each of these genes was analyzed at 10Kb resolution, to acquire a more detailed insights into possible structural differences. In these matrices, shown in figure 24, TADs are clearly visible as green-yellow triangles along the diagonal. These inter-TAD contact results in coordinated regulation of gene expression, usually associated to positive regulation of gene expression.

Gene	**Name**	**Gene locus**	**TAD search (± 1 Mb)**
IL1B	Interleukin 1 Beta	2:113587328-113594480	2:112587328-114594480
IL1R1	interleukin 1 receptor type 1	2:102681004-102796334	2:101681004-103796334
IL10	Interleukin-10	1:206940947-206945839	1:205940947-207945839
CCL2	C-C motif chemokine ligand 2	17:32582304-32584222	17:31582304-33584222
IL18	Interleukin-18	11:112013974-112034840	11:111013974-113034840
IL4	Interleukin-4	5:132009678-132018368	5:131009678-133018368
PRNP	prion protein	20:4666882-4682236	20:3666882-5682236
ACT	Actin 1	14:69340860-69446157	14:68340860-70446157

Table17: selection of genes involved in early immune response in peripheral blood in patients with CJD or other types of dementias, together with *PRNP* and *ACT*, as negative control.

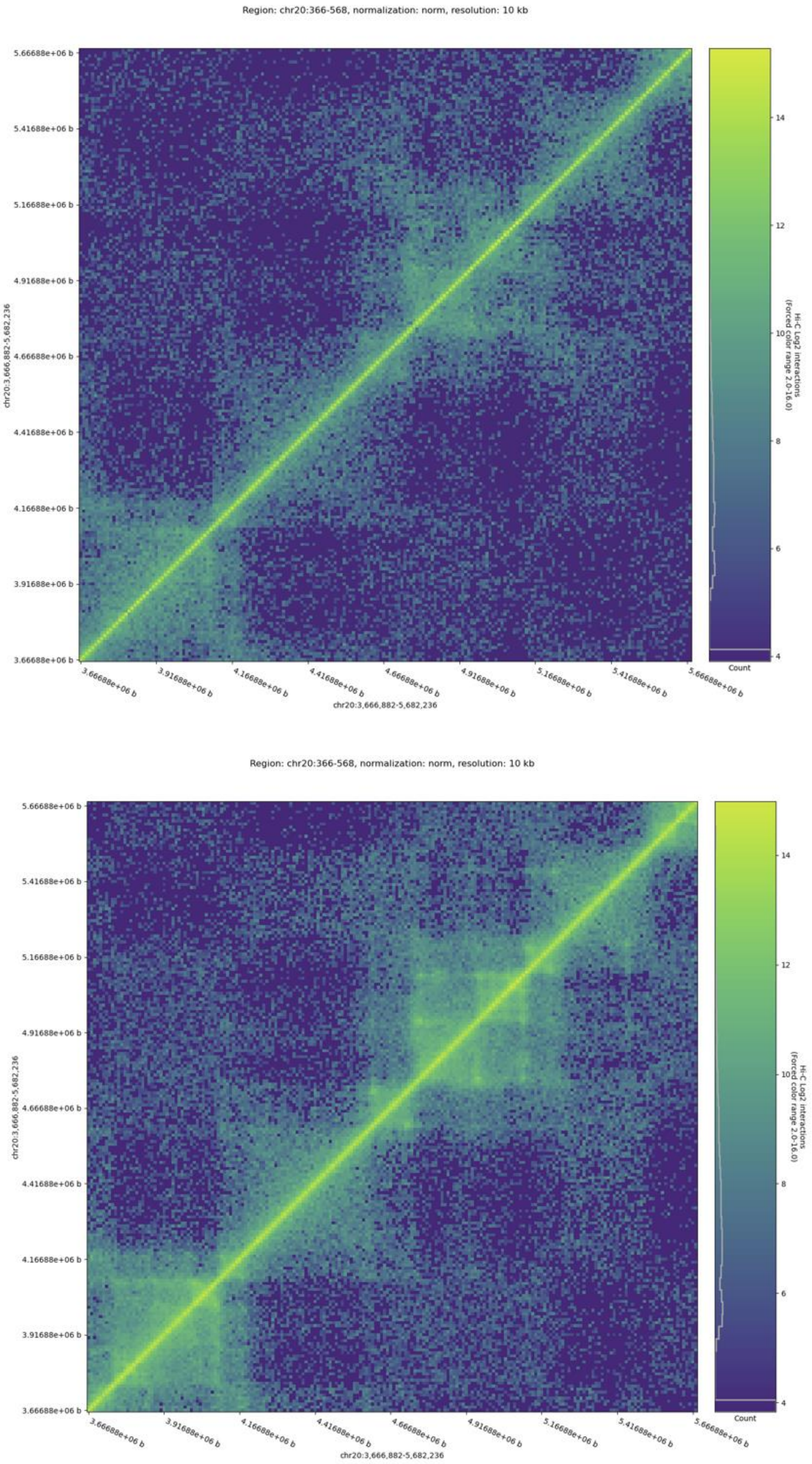

Figure 24: Hi-C matrices of the genomic locus 1Mg ahead and after *PRNP*. In the upper panel is reported the interaction matrix of the merged data of CJD patients, in the lower the interaction matrix of the merged data controls. The matrices are symmetrical along the diagonal, and interaction rates between genomic regions are marked by green-yellow areas, according to the legend on the right of the Hi-C map.

CHESS provided the genomic coordinates of the regions with significant changes in terms of number of physical interactions, together with the signal/noise ratio. In figure 25 are reported some examples of plots, in which a similarity score is reported on the y axis along the observed genomic regions.

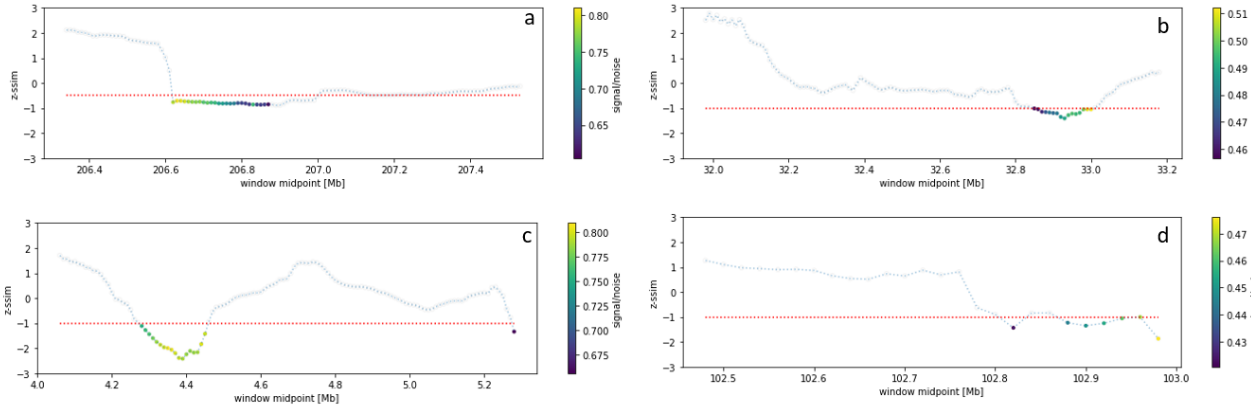

Figure 25: Plots of the similarity score along four of the selected genes. In the picture, ‘a’ represents *IL10*, ‘b’ *CCL2*, ‘c’ *PRNP* and ‘d’ *IL1R1*. Colored dots represent the signal/noise ratio value, according to the scale reported for each plot.

Among the considered genomic regions, *PRNP* locus shows the lowest similarity score and the highest signal/noise ratio. This suggests a possible significant difference in the interaction rates in cases and controls. In particular, a decreased rate of interaction seems to affect the CJD population compared to the control population, as shown in figure 26 and in the quantitative results of the feature extraction performed with CHESS. A further normalization for coverage on this specific region was added, with no changes in the result. Even though with lower intensity compared to the described 2Mb window around *PRNP*, a general increase in interaction rates and a higher signal/noise ratio was present in every genomic region in controls. This could be due to technical reasons and to the low number of analyzed samples and will be considered in the interpretation of these results.

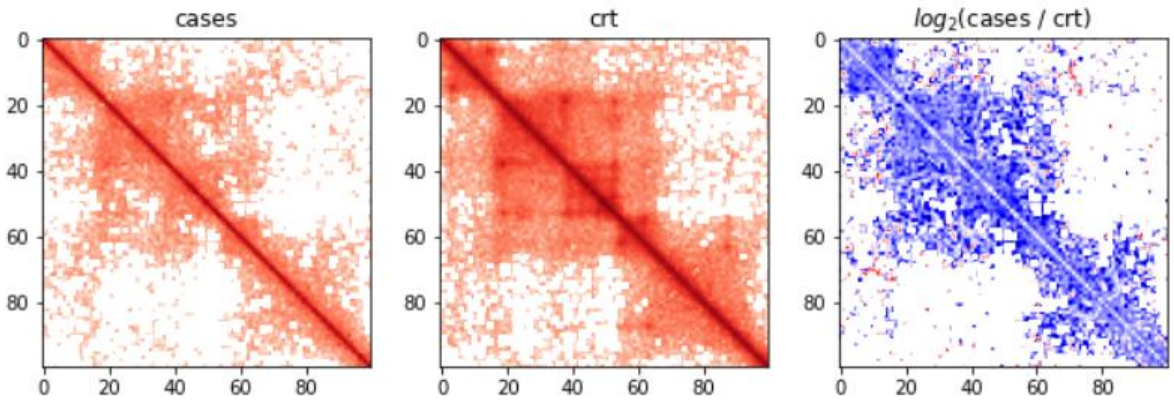

Figure 26: zoom into a sub-region of the 2Mb window around *PRNP*. A decreased rate of interactions is visible in the cases matrix, compared to the controls one, clarified by the third panel, showing the cases/control ratio of $\log_2$ Hi-C interactions.

Relevantly, the 1Mb ahead of *PRNP* is rich in regulatory regions in particular CTCF motifs, enhancers and promoters (figure 27). This previous knowledge together with the presented results open to the concrete possibility to an impairment of structural chromatin organization in patients with prion diseases.

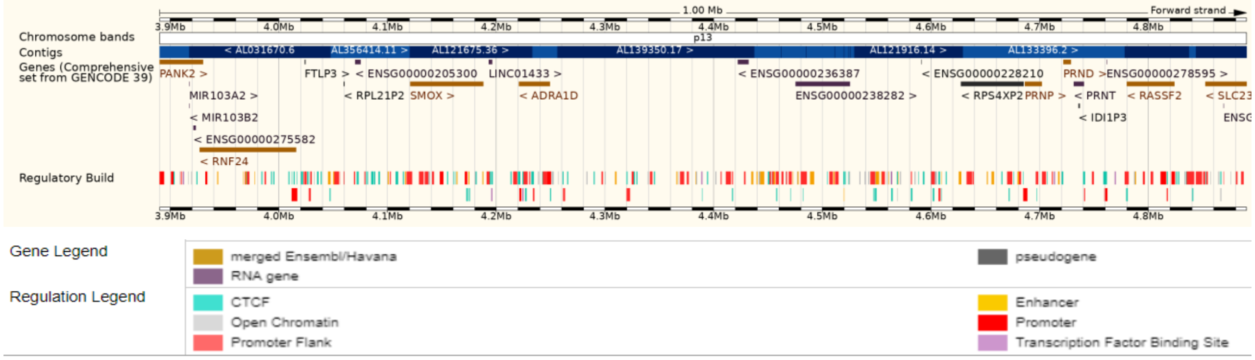

Figure 27: Snapshot of the genomic region covering *PRNP* and its upstream region, taken from Ensembl genome viewer. In the upper panel are visible the genomic coordinates together with the names and location of coding genes and regulatory transcripts. *PRNP* is visible on the up-right, together with its homologous genes *PRND* (Prion Like Protein Doppel, protein coding) and *PRNT* (Prion Locus LncRNA). In the lower panel, is visible that the upstream of these loci particularly rich in CTCF (light blue bars), enhancers (yellow bars) and promoter (red bars) regions.

DISCUSSION

In this project, a clinically and pathologically well characterized cohort of 24 samples of MM1 CJD and 24 of VV2 CJD was compared at a genomic and transcriptomic level to improve the understanding of the different molecular features between the two strains. The low numerosity of the dataset is compensated by the high characterization of the samples, aimed at reducing as much as possible intra-group variation. All samples showed pure MM1/VV2 phenotype with no/minimum co-pathology. This is reflected by the average disease duration of the two groups (3.1 months MM1 and 6.8 months VV2) that is perfectly in line with the average of 4 and 6.3 months respectively reported in literature[109,120]. Similarly, the average age of onset of our cohort fits the reported average in the VV2 group with 65.7 years in the cohort and 64.5 years as expected mean value, while in this MM1 group the observed average age of onset is slightly lower than expected (65.8 compared to the 70.1 years reported in literature)[109].

GENOMICS: DNA TARGET SEQUENCING

- VARIANT ANNOTATION

DNA sequencing results provided two main types of information, that will be discussed separately: a list of predicted damaging mutations and the overall genetic pattern of single nucleotide variants in each sample. In terms of the number of missense and likely pathogenic variants (table 5), MM1 and VV2 samples do not show significant differences, neither in terms of the distribution of the 62 predicted damaging mutations (31 per strain) nor in terms of number of samples carrying this type of variants (17 MM1 and 16 VV2). Nevertheless, of the 26 genes interested by this type of variants, seven are mutated both in MM1 and VV2 samples (*CR1, NME8, RIN3)*, while nine genes were affected only in MM1 samples *(ABCA7, ALS2, ATP13A2, CTSC, PANK2, SETX, SLC24A4, TH,* *VPS13C)* and other ten only in VV2 samples *(ATM, GRN, HTRA2, LMNB1, NECTIN2, NEK1, PINK1, PLA2G6, PRKN, ZCWPW1)*. Relevantly, *HTRA2, PINK1* and *PRKN* -affected only in the VV2 group-, are involved in Parkinson Disease (PD), where mutations in *PINK1* and *PRKN* are associated to early onset Parkinson Disease. The serine/threonine-protein kinase PINK1 is localized both in the cytosol and in outer and inner mitochondrial membranes, where exerts a protective role against mitochondrial dysfunction by activating mitochondrial quality controls mechanisms that mediate mitophagy and lysosomal function through phosphorylation of other mitochondrial proteins such as the E3 ubiquitin-protein ligase PRKN. In addition, PINK1 has been suggested to be also involved in the mitochondrial unfolded protein response (UPRmt) through its interactions with HTRA2[164]. Mitochondrial dysfunction has been associated to several neurodegenerative disorders, but very few studies are available in sCJD: recently Flønes et colleagues[165] showed a positive correlation between the level of impairment of the five respiratory complexes in neurons of both MM1 and VV2 strains with the severity of other neuropathological changes such as gliosis, vacuolation and PrP^{sc} accumulation. The missense variants *HTRA2* p.(Gly399Ser) and *PINK1* p.(Gly513Asp) were found only once in this dataset, while *PRKN* p.(Arg402Cys) was found in two different VV2 samples. *PINK1* p.(Gly513Asp) is not present in GnomAD and no clinical interpretation in available on ClinVar. *HTRA2* p.(Gly399Ser) in ClinVar is reported as likely benign or as a possible risk factor, since this variant has been found both in in healthy controls and in PD, essential tremor and cervical dystonia patients[166]. *PRKN* p.(Arg402Cys) is present in the GnomAD database, and is reported on ClinVar as a variant of uncertain significance since its pathogenicity has never been proven but it has been described in several works both in PD patients and healthy controls[167]. Despite the described variants are present only in four samples and this data lacks any statistical relevance, it is worth noticing that the putative pathogenic variants related to Parkinson Disease are found only in the VV2 strain, where movement impairment are more relevant than in MM1. This suggests that these variants might act as genetic modifier in the sporadic form of the disease, suggesting a possible overlapping mechanism regarding mitochondrial quality control dysfunction in Parkinson Disease and in VV2 sCJD. On the contrary, the genes with probably pathogenic variants in the MM1 group are involved in a more heterogeneous group of molecular processes and neurodegenerative diseases. In addition to genes involved in PD (*VPS13C* and *ATP13A2),* in this list appear also the ATP binding cassette subfamily A member 7 (*ABCA7*)[168], that is a known risk factor for Alzheimer Disease (AD) for its role in amyloid clearance and in decreasing the Aβ production through interference with APP processing and *ALS2*, which codes for the GTPase activator Alsin, associated to amyotrophic lateral sclerosis.

- STATISTICAL ANALYSIS

Statistical analysis of allele frequencies found 238 variants with a significantly (p.adj <0.05) altered allele frequency compared to the healthy European population, distributed in 37 genes. A genome wide association study (GWAS) on 5208 sCJD has been recently published[169] , identifying *STX6* rs3747957 and *GAL3ST1* rs2267161 as high risk loci, in addition to *PRNP* codon 129. Unfortunately, both *STX6* and *GAL3ST1* genes are not covered by the gene panel used in this study, therefore no data are available about these loci in the cohort analyzed. Nevertheless, coherently with the hypothesis of a polygenic contribution to the disease, functional enrichment analysis of the thirty-seven genes showing significantly over or under-represented alleles in the sCJD cohort highlights as most significantly affected, pathways already known to be altered in the disease, such as chaperon-mediated regulation of protein stability, alteration in synapse organization, Ca^{2+} transport and maintenance of cellular components. These results suggest a possible contribution of other genes in addition to SNV in the coding region of the prion protein gene, that creates a complex genetic background in which the reinforcing role of multiple variants increases the risk of developing the disease and possibly influence the prevailing strain. The statistical comparison between MM1 and VV2 samples did not find any significant over or under-represented allele in the comparison between the two groups. Although this may be due to a lack of relevantly different allele frequencies in one of the two groups, the low numerosity of this dataset could not allow to unveil such differences. Three of the overrepresented variants in the sCJD, *GRN*, *NME8* and *RIN3* genes, were also predicted as probably pathogenic by SIFT and PolyPhen2 variant predictors. Each of these three variants was found only once in this cohort; thus, this result should be confirmed in a larger population. This preliminary finding suggests a possible polygenic contribution to the development of the disease in sporadic cases. In addition, it is worth noticing that the variants in *GRN* and *RIN1* were found in VV2 cases, while a variant in *NME8* was found in MM1 CJD. Progranulin (*GRN*) and Ras and Rab interactor 3 (*RIN3*) play pivotal roles in the regulation of lysosomal function and in the activation of guanine nucleotide exchange (GEFs), respectively. Interestingly, both these biological processes turned out in the top ten significantly under expressed pathways in the gene expression analysis in the VV2 cohort (see following paragraph, transciptomics). The thioredoxin domain-containing protein 3 (*NME8*) still has an elusive function, but several studies have shown a significant association of SNPs in this gene with Alzheimer disease[170–172]. Therefore, these preliminary results are coherent with the affected pathways highlighted by the previous parts of this study and reinforce the hypothesis of a polygenic contribution of multiple mildly predisposing variants to the onset of sporadic CJD.

o DATA SCIENCE: SUPERVISED AND UNSUPERVISED MACHINE LEARING

Differently from previous analysis that focused only on exonic variants or on variants reported in reference databases, data science methods were applied on the complete genetic information acquired with NGS experiments. This approach, aimed at identifying possible recurrent genetic patterns that not necessarily involve only coding variants, consist in a recently proposed workflow[162] that involves supervised and unsupervised machine learning methods increasingly used in the analysis of biological data[173–175]. Here, unsupervised methods show quite a homogeneous genetic background between the analysed samples: the principal component analysis shows that most of the forty-eight samples tend to group around the centre of the PCA plot, and a high number of principal components (eighteen) are needed to explain 50% of the variance in the dataset. Both findings suggest an overall similar genetic background in the analysed cohort. Nevertheless, a stratification along the first two principal components is clearly visible specially in the 2D plot (figure 13). Here two clusters of samples identified by K-Means, mainly distributed along PC1, are represented with different colours. In this work, this clusterization could not be attributed to any specific feature since all the available biological, clinical and demographic labels did not match it. The main contributors of the two principal components are variants in *PRKN* and in *PICALM* genes, that therefore represent the strongest sources of variation in this dataset. *PRKN* already appeared in this work in the discussion of predicted-pathogenic variants as a gene exclusively mutated in the VV2 group. Here, interestingly, this gene represents the strongest contributor to the explained variance of the overall genetic background. The E3 ubiquitin-protein ligase PRKN (coded by *PRKN*, also known as *PARK2*) catalyses the ubiquitination of substrate proteins, playing a pivotal role in protein turnover, clearance of misfolded proteins and stress response. This protein has been described to participate in the removal or detoxification of misfolded or damaged protein by mediating their 'Lys-63'-linked-polyubiquitination, leading to their recruitment into aggresomes, followed by degradation[176]. Parkin is also involved in mono-ubiquitination of the apoptosis regulator Bcl-2 (*BCL2*), thus acting as a positive regulator of autophagy[177]. These are all molecular processes strongly impaired in various neurodegenerative disorders, and mutations in *PRKN* are associated to autosomal recessive inheritance of PD[178]. The Phosphatidylinositol-Binding Clathrin Assembly Protein coded by the gene *PICALM* is an adapter protein involved in clathrin-mediated endocytosis, that is pivotal for several biological processes such as the internalization of cell receptors, for synaptic transmission and removal of apoptotic cells. PICALM is mainly located in the cytosol, Golgi apparatus and cellular membrane, where it mediates the endocytosis of small R-SNARES (Soluble NSF Attachment Protein REceptors). It also modulates the turnover of autophagy substrates: through GWAS and functional studies, *PICALM* has been associated to AD for its role in the autophagy of substrates such as Tau or amyloid precursor protein[179]. The impairment of all the described biological functions have been described by previous literature in CJD[25] and in general in several neurodegenerative diseases. These results are in line with those previously presented for the statistical comparison of allele frequencies in the two strains: in this dataset, the genomic data acquired with the Illumina Neurodegeneration panel do not allow to highlight well defined differences in the genomic background between MM1 and VV2 sCJD samples, but allow to identify possible contributing genes to the disease that are coherent with previous knowledge about the most affected pathways in CJD.

The same data were used also as input for supervised machine learning algorithms. Supervised classifiers correctly identified the two strains using the overall genetic profile of 57005 SNVs identified with target sequencing. The classification was carried out with 100% accuracy based only on the codon 129, which shows that these tools are able to detect a single functional variant among several thousands and thus could be useful also in the clinical practice for advanced diagnostics or precision medicine purposes. This result anyway does not provide any new knowledge about the genetic contributors to the phenotypic heterogeneity of the two strains, given that the classification into strain uses the codon 129 as a criterion. To gain new insights on other genetic contributors, the codon 129 was removed from input data and the classification task repeated. The classification this time was carried out with 81% accuracy on the test set, based on two other intronic variants, one in the intronic region between exon 1 and 2 nearly 5kb upstream codon 129 and one in an intronic region of *FERMT2*. The intronic variant in *PRNP* used in the classification (rs6037932), was never reported as a functional non-coding variant, nevertheless it was used in a phylogenetic study about founder effect in Fatal Familiar Insomnia[163] in the European population. This SNP is not in complete linkage disequilibrium with the allele 129V, even though in this cohort as well as in the previously cited work it is more frequently associated to Valine. Both works comprehended small cohorts exclusively made by people affected by prion diseases and this variant is not present in neither in ClinVar nor in GnomAD, therefore it is difficult to draw conclusions about a possible functional role of this intronic variant or just phylogenetic association given the close proximity to the open reading frame of the gene. The intronic variant *FERMT2* chr14-53391236-T-G was never previously reported as functional intronic variant and is not reported in ClinVar or GnomAD. It is located in the non-coding region downstream the last exon of the *FERMT2* gene. In this cohort, it is always associated to the allele 129V. *FERMT2* has been associated to AD by GWAS[170], its corresponding protein is expressed in the brain where it is involved in axonal growth, synaptic connectivity, and long-term potentiation and directly interacting with the Amyloid Precursor Protein to modulate its metabolism.

TRANSCRIPTOMICS: RNA SEQUENCING

On a subset of the previously described cohort of MM1 and VV2 sCJD, a comparative analysis of transcriptomic profiles was carried out on post-mortem frontal cortex samples with RNA sequencing (cDNA capture). Since the analysis was performed on the terminal stage of the disease, transcriptomic data provide an overall picture of how the cells responded to stresses at the time of death. In neurodegenerative diseases this means that the observed pathogenic or adaptive gene expression changes refer only to the cells that survived the neurodegeneration and not those that were mostly affected by the disease. For this reason, frontal cortex was chosen as an appropriate region to carry out this analysis since it is partially spared in both strains. The current literature provides some insights into changes in gene expression in CJD, but several considerations must be taken into account. Most of the available studies have been carried out on murine models, that have the advantage to allow the acquisition of gene expression responses at multiple stages of the disease on big dataset but have the limit of studying only the acquired form of the disease. In addition, despite still being an important and necessary resource for biomedical studies, animal models do not always reproduce faithfully all the features of human diseases, especially when the central nervous system and cognition are involved. On the other hand, studies carried out on human samples also describe the sporadic forms of the disease but only at the terminal stage.

From a technological point of view, the most recent next generation sequencing technologies for transcriptomics have been applied only on animal models[131,132,180], while studies performed on human samples used previous methodologies, such as reverse transcribed real-time PCR or microarray[25,136,181]. This work represents the first application of full exome RNA sequencing on human samples of sporadic CJD, and also the first transcriptomic analysis comparing two subtypes of the disease, and not cases versus controls.

Quality metrics during sequencing and bioinformatic analysis showed a reliable outcome of the experimental part and a depth of sequencing adequate to have a complete picture of the coding transcriptome, covering also lowly abundant transcripts. Differential gene expression analysis compared transcript abundancies in the two strains providing a list of overexpressed genes in each condition. The results of this analysis showed a high number of differentially expressed genes, confirming previous findings about a strong impairment of gene expression, especially in the late stages of the disease. In particular, in the VV2 group a higher number of over expressed genes was found compared to the MM1 group (1196 vs 602). The hierarchical clustering and the heatmap associating the whole transcriptomic profile to sample classes identified two clusters, one clearly enriched of MM1 cases and one of VV2 cases. This finding suggest that the transcriptomic profile contains specific trends and signatures that are associated more frequently to each one of the two considered strains, and shared traits that lead to minor overlap in the clusterization. This overlap does not represent an unexpected result, given that the comparison involves two subtypes of the same disease, that are characterized by peculiar traits and a common driving mechanism towards neurodegeneration. Notably, previous transcriptomic technologies did not manage to detect a good clusterization of CJD subtypes in other cohorts[25,136].

Results of DGE analysis represented the starting point for functional analysis, that was performed with different methods. Despite their different assumptions and technical procedures, all three types of functional analysis provided coherent and complementary outcomes, that helped defining a detailed picture of the pathways most affected in the two subtypes. In this work, we observed in the VV2 strain an upregulation of genes involved in synaptic regulation affecting both pre and post synaptic terminals. Similarly, VV2 showed an overexpression of genes involved in vesicle transport and turnover compared to MM1. Both these biological processes have been described as strongly impaired in any subtype of CJD, especially in the mid and late stages of the disease[25,132,136]. Studies comparing CJD cases with non-affected controls, already observed that the intermediate phases of the disease are characterized by strong impairment of transcriptional response directed to pathways involving vesicular trafficking, with the activation of cholesterol synthesis and efflux, glycosaminoglycan metabolism, and sphingolipid synthesis and degradation[132] that affect both exocytosis and endocytosis as well as intra cellular vesicle trafficking. In later stages of disease the impairment increases as genes associated to synaptic transmission and axon guidance are down regulated and ultimately cellular processes associated with cell death are activated[132,136,182]. The observed overexpression in VV2 compared to MM1 was concordant between all functional enrichment methods, suggesting that in the VV2 strain the downregulation of these processes may take place with less intensity. Previous works already demonstrated that in CJD genes regulating these pathways are under expressed compared to controls[25,183]. Notably, tissue used for testing RNA expression were selected to have a similar level of cellular damage, therefore even if these results are acquired by bulk RNA sequencing, cell abundancy and survival should not be a confounder element. In the VV2 group an enrichment of pathways involving dopamine secretion, regulation of calcium release, GABA signaling, and mitochondrial permeability was observed. Both glutamatergic and GABAergic synapses have been described to be impaired in CJD[25,136] while dopamine secretion, transport and response have been less studied in CJD. This finding is coherent with the results of the genomic layer of this thesis, where an enrichment of SNPs and probably pathogenic missense variants in genes associated to Parkinson Disease were observed only in VV2. In VV2 CJD movement deficits are more relevant than in other subtypes, largely because of the prominent degeneration in the cerebellum. These results suggest that motor symptoms could be also tracked back to impairment of dopaminergic synapses and GABA signaling, possibly with shared mechanisms with Parkinson Disease. Interestingly, no significant differences in terms of immune response and neuroinflammation were found in this comparison. Neuroinflammation and activation of immune response are upregulated since the early stages of the disease[184], therefore the lack of significant differences could be explained by the fact that in the terminal part of the disease these processes are already fully activated and widespread in both subtypes.

In the MM1 group, the most affected pathways involved regulatory pathways mediated by guanosine triphosphatases (GTPases), regulation of catabolic processes and maintenance of proper cell morphology and matrix organization. Additionally, Ras transduction pathways showed and overexpression in MM1 compared to VV2, indicating a major impairment of key signal transduction pathway, consequently influencing a multitude of downstream pathways. Small guanosine triphosphatases (GTPases) of the Ras superfamily are important regulators of key cellular processes like cell cycle regulation, proliferation, intracellular trafficking, and apoptosis. GTPases act as intracellular molecular switches and are inactive when bound to guanosine diphosphate (GDP), and active when the GTPase is bound to guanosine triphosphate (GTP). Their involvement in neurodegenerative diseases is linked to many processes, most relevantly to the impairment of catabolic processes, vesicular trafficking and to regulation of apoptosis[185]. Here, results of the functional analysis confirm their dysregulation also in sCJD, similarly to other neurodegenerative diseases. Likewise to what observed in the genomic profiling study, also at the transcriptomic level the MM1 group shows a gene expression profile with several traits shared with different neurodegenerative, without a clear distinctive trait or similarities with a specific disease. On the contrary, in the VV2 strain transcriptomic analysis confirms the involvement of pathways dysregulated in Parkinson's disease, adding strength to the findings of genomic profiling.

EPIGENOMICS: HI-C

A comparative analysis of the three-dimensional chromatin organization with Hi-C was performed between white blood cells of CJD patients at the time of diagnosis and healthy controls of the same age. How chromatin organization changes with cellular senescence and how this influences healthy aging and neurodegenerative diseases is only beginning to be explored. This work represents the first application of Hi-C or any chromosome conformation capture technique to prion diseases, therefore the presented results should be considered as exploratory. There are nevertheless some very recent works that explored alterations in chromatin organization in other types of neurodegenerative diseases and age-related conditions[46]: in a Huntington Disease mouse model, it has been recently shown with 4C-seq that the CAG expansion responsible for the disease affects the correct insulation of the TAD upstream *HTT*[59]. In fact, it has been proposed as a general mechanism that Short Tandem Repeats (STRs) associated diseases (such as fragile X syndrome, Huntington's disease, Amyotrophic Lateral Sclerosis and Friedreich's ataxia) could cause the pathogenic condition through the topological disruption of higher-order chromatin folding[74]. Another example is Cornelia de Lange Syndrome (CdLS), a multisystemic disease with relevant neurological damage that is caused by the impairment of the Cohesin complex, which cooperate with CTCF proteins to establish and maintain the correct 3D chromatin structure. In a recent work[58] has been shown in post-mortem cerebral cortex of CdLS patients a prominent downregulation of hundreds of genes enriched for fundamental neuronal functions, including synaptic transmission and signaling processes. A concordant transcriptomic profile was obtained in an in vitro model of cortical neuron depleted for the cohesin complex, and interestingly gene expression alterations were often greatly recovered after restoration of cohesin function in the in vitro model, indicating that at least some of these changes may be reversible. Aging itself seems to be associated with a progressive loss of local interactions and weakening of TADs boundaries, associated to the establishment of Senescent-Associated Heterochromatin Foci (SAHFs)[46,186].

In this work, post-alignment metrics showed a number of valid pairs that allowed reliable analysis of chromatin interactions, also at high resolution (20-10Kb). Low resolution matrices at 700Kb on the whole genome identified chromosome compartments and, as expected, did not highlight any inter-chromosomal rearrangement. Low resolution matrices of single chromosomes at 100Kb did not highlight any intra-chromosomal rearrangement in neither CJD cases and controls, nevertheless, as confirmed also by the binning at 50Kb, a general higher rate of interactions was observed in controls. Being Hi-C a genome wide technique, the comparative analysis focused mainly on a selection of genes, known from previous literature to be involved in early response in immune cells to either CJD or other neurodegenerative diseases[30,187–190]. Also at this resolution, in the CJD samples was observed a significant decrease of interaction rates, that became even more evident in the genomic locus containing *PRNP*. This work shows that a general loss of genomic interactions is visible at disease onset in immune cells, and that this general trend seems to be exacerbated in the disease-causing gene. Relevantly, the genomic region upstream of *PRNP* harbors several regulation sites, like CTCF motifs, enhancers and promoter sites. These results should be considered preliminary since these observations were carried out in a small dataset, so we cannot completely exclude that some biases might still be present also after data normalization. Previous literature demonstrated that no changes in cell-type composition in white blood cells are present of CJD patients compared to healthy controls[191], thus the differences observed in this work are not due to different subpopulation abundancies. Even though carrying out this experiment in peripheral immune cells is not optimal since the disease affects mainly the brain, an approximation was necessary in order to compare CJD patients at disease onset and healthy controls, given the obvious unavailability of the best tissue type. Immune cells represent a good approximation, since *PRNP* is physiologically expressed, it is a noninvasive tissue to collect, and the immune cells have been described to play a critical role especially in the initial stages of prion diseases[188]. These results suggest that a possible structural impairment leading to a loss of interactions and weakening of TADs boundaries takes place in the early stages of the disease: in this context there are not enough data to make definitive claims, but these findings are in line with what has been observed in normal aging and also with the observations of an accelerated aging in Huntington Disease based on the integration of epigenomic and transcriptomic data[59]. In addition, it is worth noticing that in the transcriptomic analysis performed in this thesis several genes coding for the cohesin complex (such as *RAD51AP2*, *RAD51C, RAD23A* and *STAG3)* were significantly differentially expressed between the two analyzed strains, meaning that this complex is in some way involved in this disease. Similarly the previously cited work about CdLS[58], also here synaptic transmission and signaling were the most affected biological processes: it would be interesting to investigate more in detail the role of Cohesin in CJD given the recovery they observed in the in vitro model upon restoration of the cohesin complex. This work represents thus a starting point for further studies, that could provide additional insights in terms of the role played by specific protein complexes or through the acquisition of other -omic data, like transcriptomic profiles or other epigenomic modifications, such as histone modifications, DNA methylation or chromatin accessibility.

CONCLUDING REMARKS

In this work, we addressed the problem of phenotypic heterogeneity in prion diseases by producing, analyzing and comparing multiple omic layers, with the aim of characterizing the molecular differences and similarities between the two most common strains of sporadic Creutzfeldt-Jakob Disease. In addition, to further characterize the genetic background of the disease, a comparative analysis of the tridimensional chromatin organization between CJD patients and healthy controls was carried out, thus investigating also the epigenomic layer. The results of this project represent a novelty in the state of the art in this field, both from a biomedical and technological point of view.

At the genomic level, the presented results suggest that sCJD could have polygenic contributions able to influence the prevalent strain. Both restricted analysis of missense and probably damaging variants as well as data science approaches on overall genetic profiles were consistent in identifying a contribution of genes associated to Parkinson Disease particularly in VV2 patients, while MM1 samples showed contributions of genes associated to a more heterogeneous variety of neurodegenerative diseases.

The transcriptomic analysis presented in this work represents the first application of RNA sequencing on human sCJD samples and the first comparative analysis between sCJD strains using next generation sequencing technologies. On a subset of the genomic dataset, we identified nearly 1800 differentially expressed genes between the two strains. The subsequent functional analysis confirmed and expanded the finding of an impairment of some pivotal biological processes associated to Parkinson disease, such as dopamine secretion, regulation of calcium release, GABA signaling, and mitochondrial permeability in VV2 sCJD. The transcriptomic profile of MM1, coherently with the genomic findings, portrayed a gene expression profile where several biological processes shared by different neurodegenerative diseases were involved, without a clear distinctive trait.

This multi omics analysis provides robust evidence for the distinctive features of the two most common strains of sCJD. In both strains, multiple possible genetic contributors have been identified, suggesting at a polygenic predisposing background to the disease. In addition, given the pronounced motor impairment in the VV2 strain, these findings strongly suggest that motor symptoms could also be tracked to impairment of dopaminergic synapses and GABA signaling, possibly through mechanisms shared with Parkinson Disease.

On the epigenomic layer, the tridimensional folding of the genome in peripheral immune cells of CJD patients at onset and healthy controls was investigated with Hi-C. This work represents the first application of this cutting-edge technology in prion diseases, and one of the first in general in the broader field of neuroscience. The results show that no major chromosome reorganization, such as translocation or duplications, takes place during the disease. Nevertheless, a significant and diffuse loss of genomic interactions was detected at disease onset in immune cells, and this general trend seems to be exacerbated in *PRNP* locus.

All these results update and increase the current knowledge on the molecular features of phenotypic heterogeneity and genomic background of prion diseases, representing a promising starting point for future research projects.

APPENDIX

GENE NAME	**PROTEIN NAME**
ABCA7	Phospholipid-transporting ATPase ABCA7
ALS2	Alsin
ATM	Serine-protein kinase ATM
ATP13A2	Polyamine-transporting ATPase 13A2
CR1	Complement receptor type 1
CTSC	Dipeptidyl peptidase 1
EPHA1	Ephrin type-A receptor 1
FERMT2	Fermitin family homolog 2
GRN	Progranulin
HTRA2	Serine protease HTRA2, mitochondrial
LMNB1	Lamin-B1
NECTIN2	Nectin-2
NEK1	Serine/threonine-protein kinase Nek1
NME8	Thioredoxin domain-containing protein 3
NOTCH3	Neurogenic locus notch homolog protein 3
PANK2	Pantothenate kinase 2, mitochondrial
PINK1	Serine/threonine-protein kinase PINK1, mitochondrial
PLA2G6	85/88 kDa calcium-Independent phospholipase A2
POLG	DNA polymerase subunit gamma-1
PRKN	E3 ubiquitin-protein ligase parkin
RAD23A	UV excision repair protein RAD23 homolog A
RAD51AP2	RAD51-associated protein 2
RAD51C	DNA repair protein RAD51 homolog 3
RIN1	Ras and Rab interactor 1
RIN3	Ras and Rab interactor 3
SETX	Probable helicase senataxin
SLC24A4	Sodium/potassium/calcium exchanger 4
STAG3	Cohesin subunit SA-3
TH	Tyrosine Hydroxylase
TOR1A	Torsin-1A
VPS13C	Vacuolar protein sorting-associated protein 13C
ZCWPW1	Zinc finger CW-type PWWP domain protein 1

Table S1: Names of genes and corresponding extended protein names appearing in the text.

Gene	HGVSC	HGVSP	Sift Prediction	PolyPhen Prediction	strain	Sample
ABCA7	c.2026G>A	p.(Ala676Thr)	deleterious	probably damaging	MM1	#30
ABCA7	c.5168C>T	p.(Ser1723Leu)	deleterious	probably damaging	MM1	#30
ABCA7	c.1456C>G	p.(Pro486Ala)	deleterious	probably damaging	MM1	#43
ALS2	c.3685T>A	p.(Trp1229Arg)	deleterious	possibly damaging	MM1	#46
ATM	c.998C>T	p.(Ser333Phe)	deleterious	probably damaging	VV2	#10
ATP13A2	c.3472C>T	p.(Arg1158Cys)	deleterious	possibly damaging	MM1	#31
CR1	c.313C>T	p.(Arg105Cys)	deleterious	probably damaging	VV2	#1
CR1	c.2537C>T	p.(Ser846Phe)	deleterious	possibly damaging	MM1	#33
CTSC	c.1357A>G	p.(Ile453Val)	deleterious	possibly damaging	MM1	#27
CTSC	c.1357A>G	p.(Ile453Val)	deleterious	possibly damaging	MM1	#42
CTSC	c.1357A>G	p.(Ile453Val)	deleterious	possibly damaging	MM1	#46
EPHA1	c.2090C>T	p.(Pro697Leu)	deleterious	possibly damaging	MM1	#29
EPHA1	c.2090C>T	p.(Pro697Leu)	deleterious	possibly damaging	VV2	#14
EPHA1	c.2090C>T	p.(Pro697Leu)	deleterious	possibly damaging	MM1	#15
EPHA1	c.2897G>A	p.(Arg966His)	deleterious	probably damaging	MM1	#47
EPHA1	c.2090C>T	p.(Pro697Leu)	deleterious	possibly damaging	MM1	#47
EPHA1	c.2090C>T	p.(Pro697Leu)	deleterious	possibly damaging	VV2	#22
EPHA1	c.2090C>T	p.(Pro697Leu)	deleterious	possibly damaging	VV2	#24
GRN	c.325G>A	p.(Gly109Arg)	deleterious	probably damaging	VV2	#21
HTRA2	c.1195G>A	p.(Gly399Ser)	deleterious	probably damaging	VV2	#23
LMNB1	c.1474G>T	p.(Ala492Ser)	deleterious	possibly damaging	VV2	#4
NECTIN2	c.577C>T	p.(Arg193Trp)	deleterious	probably damaging	VV2	#9
NEK1	c.2235T>G	p.(Asn745Lys)	deleterious	probably damaging	VV2	#9
NME8	c.1013T>C	p.(Ile338Thr)	deleterious	probably damaging	MM1	#25
NME8	c.1013T>C	p.(Ile338Thr)	deleterious	probably damaging	VV2	#5
NME8	c.1013T>C	p.(Ile338Thr)	deleterious	probably damaging	VV2	#6
NME8	c.1013T>C	p.(Ile338Thr)	deleterious	probably damaging	VV2	#8
NME8	c.1013T>C	p.(Ile338Thr)	deleterious	probably damaging	MM1	#15
NME8	c.1247G>A	p.(Ser416Asn)	deleterious	probably damaging	MM1	#40
NME8	c.1013T>C	p.(Ile338Thr)	deleterious	probably damaging	MM1	#42
NME8	c.1630G>A	p.(Ala544Thr)	deleterious	probably damaging	VV2	#17
NME8	c.1013T>C	p.(Ile338Thr)	deleterious	probably damaging	VV2	#18
NOTCH3	c.3293C>T	p.(Thr1098Ile)	deleterious	possibly damaging	VV2	#3
NOTCH3	c.5854G>A	p.(Val1952Met)	deleterious	probably damaging	VV2	#10
NOTCH3	c.5854G>A	p.(Val1952Met)	deleterious	probably damaging	MM1	#38
NOTCH3	c.5854G>A	p.(Val1952Met)	deleterious	probably damaging	MM1	#39
NOTCH3	c.1487C>T	p.(Pro496Leu)	deleterious	possibly damaging	MM1	#15
NOTCH3	c.3691C>T	p.(Arg1231Cys)	deleterious	possibly damaging	VV2	#17
NOTCH3	c.1487C>T	p.(Pro496Leu)	deleterious	possibly damaging	VV2	#23
PANK2	c.775G>A	p.(Gly259Arg)	deleterious	probably damaging	MM1	#48
PINK1	c.1538G>A	p.(Gly513Asp)	deleterious	probably damaging	VV2	#5
PLA2G6	c.481C>T	p.(Arg161Cys)	deleterious	probably damaging	VV2	#6

POLG	c.2492A>G	p.(Tyr831Cys)	deleterious	possibly damaging	MM1	#29
POLG	c.2492A>G	p.(Tyr831Cys)	deleterious	possibly damaging	VV2	#17
PRKN	c.1204C>T	p.(Arg402Cys)	deleterious	probably damaging	VV2	#9
PRKN	c.1204C>T	p.(Arg402Cys)	deleterious	probably damaging	VV2	#10
RIN3	c.1742G>A	p.(Arg581Gln)	deleterious	probably damaging	VV2	#8
RIN3	c.2377T>C	p.(Tyr793His)	deleterious	possibly damaging	MM1	#43
SETX	c.4660T>G	p.(Cys1554Gly)	deleterious	probably damaging	MM1	#15
SLC24A4	c.377T>C	p.(Leu126Pro)	deleterious	probably damaging	MM1	#47
TH	c.484T>G	p.(Phe162Val)	deleterious	possibly damaging	MM1	#46
TOR1A	c.646G>C	p.(Asp216His)	deleterious	possibly damaging	VV2	#1
TOR1A	c.646G>C	p.(Asp216His)	deleterious	possibly damaging	MM1	#25
TOR1A	c.646G>C	p.(Asp216His)	deleterious	possibly damaging	VV2	#6
TOR1A	c.646G>C	p.(Asp216His)	deleterious	possibly damaging	VV2	#9
TOR1A	c.646G>C	p.(Asp216His)	deleterious	possibly damaging	MM1	#15
TOR1A	c.646G>C	p.(Asp216His)	deleterious	possibly damaging	MM1	#41
TOR1A	c.646G>C	p.(Asp216His)	deleterious	possibly damaging	VV2	#17
TOR1A	c.646G>C	p.(Asp216His)	deleterious	possibly damaging	MM1	#48
TOR1A	c.646G>C	p.(Asp216His)	deleterious	possibly damaging	VV2	#24
VPS13C	c.2342T>G	p.(Leu781Trp)	deleterious	probably damaging	MM1	#41
ZCWPW1	c.314A>G	p.(Glu105Gly)	deleterious	probably damaging	VV2	#9

Table S2: complete list of missense variants identified with target sequencing predicted as deleterious.

baseMean	log2FoldChange	pvalue	padj	transcript_id	gene_name
9,2522	-3,0190196	0,00036	0,013699	ENSG00000124107.5	*SLPI*
15,0865	-2,831184	0,00183	0,028777	ENSG00000136689.14	*IL1RN*
34,3051	-2,5516605	0,00025	0,011418	ENSG00000106258.9	*CYP3A5*
331,9027	-2,3293563	0,00015	0,008828	ENSG00000273429.1	*AC253572.2*
62,6188	-2,2242763	0	0,001217	ENSG00000157653.7	*C9orf43*
15,2572	-2,199948	0,00094	0,020892	ENSG00000188162.6	*OTOG*
6,9398	-2,1861505	0,00092	0,020712	ENSG00000188060.6	*RAB42*
16,4217	-2,170038	0,00244	0,033023	ENSG00000135476.7	*ESPL1*
54,5661	-2,1537979	0,00054	0,01619	ENSG00000085265.6	*FCN1*
48,1299	-2,1339709	0,00025	0,011578	ENSG00000185860.9	*C1orf110*

Table S3: Ten overexpressed genes in the MM1 cohort, compared to VV2. Results are in decreasing order based on intensity of differential gene expression (log2FoldChange) and not in terms of statistical significance (padj).

baseMean	log2FoldChange	pvalue	padj	transcript_id	gene_name
48,116	5,3974197	1E-05	0,002565	ENSG00000186081.7	*KRT5*
33,3155	4,6097328	0	0,000655	ENSG00000198601.2	*OR2M2*
24,8391	4,0013478	4E-04	0,014472	ENSG00000120586.4	*MRC1*
10,8484	3,9841836	0,002	0,027512	ENSG00000113889.7	*KNG1*
15,1611	3,7129433	0	0,001225	ENSG00000163898.5	*LIPH*
17,0479	3,4083771	6E-05	0,00568	ENSG00000161905.8	*ALOX15*
162,7731	3,0622715	2E-05	0,003182	ENSG00000091664.7	*SLC17A6*
21,2658	3,0419085	1E-05	0,002067	ENSG00000253457.1	*SMIM18*
6,277	3,0096604	2E-04	0,010293	ENSG00000184544.7	*DHRS7C*
39,7363	2,9737858	5E-05	0,005394	ENSG00000091482.5	*SMPX*

Table S4: Ten overexpressed genes in the MM1 cohort, compared to VV2. Results are in decreasing order based on intensity of differential gene expression (log2FoldChange) and not in terms of statistical significance (padj).

REFERENCES

1. WATSON, J. D. & CRICK, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nat. 1953 1714356* **171**, 737–738 (1953).

2. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463 (1977).

3. Heather, J., Chain, B., Heather, J. M. & Chain, B. The Sequence of Sequencers : The History of Sequencing DNA Genomics The sequence of sequencers : The history of sequencing DNA. *Genomics* **107**, 1–8 (2015).

4. Mullis, K. *et al.* Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. *Cold Spring Harb. Symp. Quant. Biol.* **51**, 263–273 (1986).

5. Cohen, S. N., Chang, A. C. Y., Boyer, H. W. & Helling, R. B. Construction of Biologically Functional Bacterial Plasmids In Vitro. *Proc. Natl. Acad. Sci. U. S. A.* **70**, 3240 (1973).

6. JC, V. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).

7. Pereira, R. & Oliveira, J. Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics.

8. SJ, E., WB, B., L, L. & PS, S. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res.* **17**, 69–73 (2007).

9. U, N. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).

10. Friedman, N. & Rando, O. J. Epigenomics and the structure of the living genome. *Genome Res.* **25**, 1482–1490 (2015).

11. Jerkovic´, I. & Cavalli, G. Understanding 3D genome organization by multidisciplinary methods. *Nat. Rev. Mol. Cell Biol.* **0123456789**, (2021).

12. Wang, D. & Bodovitz, S. Single cell analysis: the new frontier in 'Omics'. *Trends Biotechnol.* **28**, 281 (2010).

13. R, L., N, S., M, B., S, D. & T, S. Transcriptomics technologies. *PLoS Comput. Biol.* **13**, (2017).

14. A, M., BA, W., K, M., L, S. & B, W. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).

15. Lister, R. *et al.* Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell* **133**, 523 (2008).

16. Nagalakshmi, U. *et al.* The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* **320**, 1344 (2008).

17. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57 (2009).

18. Mehmood, A. *et al.* Systematic evaluation of differential splicing tools for RNA-seq studies. *Brief. Bioinform.* **21**, 2052–2065 (2020).

19. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* **16**, 1–13 (2015).

20. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* (2018) doi:10.1038/s41592-018-0051-x.

21. Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283 (2009).

22. Hutchins, A. P., Poulain, S., Fujii, H. & Miranda-Saavedra, D. Discovery and characterization of new transcripts from RNA-seq data in mouse CD4+ T cells. *Genomics* **100**, 303–313 (2012).

23. Haas, B. J. *et al.* STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv* 120295 (2017) doi:10.1101/120295.

24. Sampaio-Silva, F., Magalhães, T., Carvalho, F., Dinis-Oliveira, R. J. & Silvestre, R. Profiling of RNA Degradation for Estimation of Post Morterm Interval. *PLoS One* **8**, 56507 (2013).

25. Bartoletti-Stella, A. *et al.* Analysis of RNA Expression Profiles Identifies Dysregulated Vesicle Trafficking Pathways in Creutzfeldt-Jakob Disease. *Mol. Neurobiol.* **56**, 5009–5024 (2019).

26. Marti-Renom, M. A. & Mirny, L. A. Bridging the resolution gap in structural modeling of 3D genome organization. *PLoS Comput. Biol.* **7**, 1002125 (2011).

27. Chen, Z., Li, S., Subramaniam, S., Shyy, J. Y. J. & Chien, S. Epigenetic Regulation: A New Frontier for Biomedical Engineers. *Annu. Rev. Biomed. Eng.* **19**, 195–219 (2017).

28. Landgrave-Gómez, J., Mercado-Gómez, O. & Guevara-Guzmán, R. Epigenetic mechanisms in neurological and neurodegenerative diseases. *Front. Cell. Neurosci.* **9**, 58 (2015).

29. Berson, A., Nativio, R., Berger, S. L. & Bonini, N. M. Epigenetic Regulation in Neurodegenerative Diseases. *Trends Neurosci.* **41**, 587–598 (2018).

30. Evans, S. A., Horrell, J. & Neretti, N. The three-dimensional organization of the genome in cellular senescence and age-associated diseases. *Semin. Cell Dev. Biol.* **90**, 154–160 (2019).

31. Chakraborty, A. & Ay, F. The role of 3D genome organization in disease: From compartments to single nucleotides. *Semin. Cell Dev. Biol.* **90**, 104–113 (2019).

32. Berger, F. Emil Heitz, a true epigenetics pioneer. *Nature Reviews Molecular Cell Biology* vol. 20 572 (2019).

33. Passarge, E. Emil Heitz and the concept of heterochromatin: Longitudinal chromosome differentiation was recognized fifty years ago. *Am. J. Hum. Genet.* **31**, 106–115 (1979).

34. Manuelidis, L. Individual interphase chromosome domains revealed by in situ hybridization. *Hum. Genet.* **71**, 288–293 (1985).

35. Kornberg, R. D. Chromatin structure: A repeating unit of histones and DNA. *Science (80-.).* **184**, 868–871 (1974).

36. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).

37. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **19**, 371–384 (2018).

38. Sala, C. *et al.* Evaluation of pre-processing on the meta-analysis of DNA methylation data from the Illumina HumanMethylation450 BeadChip platform. *PLoS One* **15**, 1–15 (2020).

39. Freeman, D. M. & Wang, Z. Epigenetic Vulnerability of Insulator CTCF Motifs at Parkinson's Disease-Associated Genes in Response to Neurotoxicant Rotenone. *Front. Genet.* **11**, 1–15 (2020).

40. Croft, J. A. *et al.* Differences in the localization and morphology of chromosomes in the human nucleus. *J. Cell Biol.* **145**, 1119–1131 (1999).

41. Serra, F. *et al.* Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. **13**, e1005665 (2017).

42. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (80-.).* **326**, 289–293 (2009).

43. Tavares-Cadete, F., Norouzi, D., Dekker, B., Liu, Y. & Dekker, J. Multi-contact 3C reveals that the human genome during interphase is largely not entangled. *Nat. Struct. Mol. Biol.* **27**, 1105–1114 (2020).

44. Yu, M. & Ren, B. The three-dimensional organization of mammalian genomes. *Annu. Rev. Cell Dev. Biol.* **33**, 265–289 (2017).

45. Tanabe, H. *et al.* Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 4424–4429 (2002).

46. Sun, L., Yu, R. & Dang, W. Chromatin architectural changes during cellular senescence and aging. *Genes (Basel).* **9**, (2018).

47. Steensel, B. van & Belmont, A. S. Lamina-associated domains: links with chromosome architecture, heterochromatin and gene repression. *Cell* **169**, 780 (2017).

48. Pombo, A. & Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol. 2015 164* **16**, 245–257 (2015).

49. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nat. 2012 4857398* **485**, 376–380 (2012).

50. Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: Interpreting chromatin interaction data. *Nat. Rev. Genet.* **14**, 390–403 (2013).

51. Szabo, Q. *et al.* Regulation of single-cell genome organization into TADs and chromatin nanodomains. *Nat. Genet. 2020 5211* **52**, 1151–1157 (2020).

52. P, V. *et al.* High-Resolution Mapping of Multiway Enhancer-Promoter Interactions Regulating Pathogen Detection. *Mol. Cell* **80**, 359-373.e8 (2020).

53. EP, N. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).

54. G, F. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* **15**, 2038–2049 (2016).

55. G, W. *et al.* Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J.* **36**, 3573–3599 (2017).

56. SSP, R. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320.e24 (2017).

57. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. doi:10.1073/pnas.1518552112.

58. Weiss, F. D. *et al.* Neuronal genes deregulated in Cornelia de Lange Syndrome respond to removal and re-expression of cohesin. *Nat. Commun.* **12**, 1–13 (2021).

59. Alcalá-Vida, R. *et al.* Age-related and disease locus-specific mechanisms contribute to early remodelling of chromatin structure in Huntington’s disease mice. *Nat. Commun.* **12**, 1–16 (2021).

60. Valton, A.-L. & Dekker, J. TAD disruption as oncogenic driver. *Curr. Opin. Genet. Dev.* **36**, 34 (2016).

61. Babu, D. & Fullwood, M. J. 3D genome organization in health and disease: emerging opportunities in cancer translational medicine. (2015) doi:10.1080/19491034.2015.1106676.

62. Akdemir, K. C. *et al.* Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nat. Genet.* doi:10.1038/s41588-019-0564-y.

63. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters ENCODE Encyclopedia of DNA Elements. *Nature* **489**, (2012).

64. Li, G. *et al.* Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. (2012) doi:10.1016/j.cell.2011.12.014.

65. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371 (2014).

66. Schoenfelder, S. & Fraser, P. Long-range enhancer–promoter contacts in gene expression control. *Nat. Rev. Genet. 2019 208* **20**, 437–455 (2019).

67. J, D., K, R., M, D. & N, K. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).

68. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nat. Genet. 2006 3811* **38**, 1348–1354 (2006).

69. J, D. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16**, 1299–1309 (2006).

70. Fraser, J. *et al.* Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol. Syst. Biol.* **11**, 852 (2015).

71. S, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25**, 582–597 (2015).

72. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

73. JM, B. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).

74. Sun, J. H. *et al.* Disease-Associated Short Tandem Repeats Co-localize with Chromatin Domain Boundaries. *Cell* **175**, 224–238 (2018).

75. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).

76. Grüning, B. *et al.* Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods 2018 157* **15**, 475–476 (2018).

77. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Research 2021 1033* **10**, 33 (2021).

78. Dall'Olio, D. *et al.* Impact of concurrency on the performance of a whole exome sequencing pipeline. *BMC Bioinforma. 2021 221* **22**, 1–15 (2021).

79. Roy, S. *et al.* Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J. Mol. Diagnostics* **20**, 4–27 (2018).

80. Burrows, M. & Wheeler, D. J. A Block-sorting Lossless Data Compression Algorithm. (1994).

81. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* (2009) doi:10.1093/bioinformatics/btp324.

82. NK, R. *et al.* Chromosome 21q21 sublocalisation of gene encoding beta-amyloid peptide in cerebral vessels and neuritic (senile) plaques of people with Alzheimer disease and Down syndrome. *Lancet (London, England)* **1**, 384–385 (1987).

83. Ho, D. S. W., Schierding, W., Wake, M., Saffery, R. & O'Sullivan, J. Machine learning SNP based prediction for precision medicine. *Front. Genet.* **10**, 267 (2019).

84. Álvarez-Machancoses, Ó., Galiana, E. J. D., Cernea, A., de la Viña, J. F. & Fernández-Martínez, J. L. On the role of artificial intelligence in genomics to enhance precision medicine. *Pharmgenomics. Pers. Med.* **13**, 105–119 (2020).

85. Dias, R. & Torkamani, A. Artificial intelligence in clinical and genomic diagnostics. *Genome Med.* **11**, 1–12 (2019).

86. Diaz-Papkovich, A., Anderson-Trocmé, L. & Gravel, S. A review of UMAP in population genetics. *J. Hum. Genet.* **66**, 85–91 (2021).

87. Li, W., Cerise, J. E., Yang, Y. & Han, H. Application of t-SNE to human genetic data. *J. Bioinform. Comput. Biol.* **15**, 1–14 (2017).

88. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).

89. The Gene Ontology, C. *et al.* The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* (2019) doi:10.17863/CAM.36439.

90. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).

91. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).

92. Ay, F. & Noble, W. S. Analysis methods for studying the 3D architecture of the genome. **16**, 183 (2015).

93. SB, P., Prusiner, S. B. & SB, P. Novel proteinaceous infectious particles cause scrapie. *Science (80-.).* **216**, 136–144 (1982).

94. Chen, C. & Dong, X.-P. Epidemiological characteristics of human prion diseases. *Infect. Dis. Poverty* **5**, (2016).

95. Baker, H. F. & Ridley, R. M. What Went Wrong in BSE ? From Prion Disease to Public Disaster. *Brain Res. Bull.* **40**, 237–244 (1996).

96. WHO | Bovine spongiform encephalopathy. *WHO* (2011).

97. Minikel, E. V. *et al.* Quantifying penetrance in a dominant disease gene using large population control cohorts. *Sci. Transl. Med.* **8**, 322ra9-322ra9 (2016).

98. Capellari, S., Strammiello, R., Saverioni, D., Kretzschmar, H. & Parchi, P. Genetic Creutzfeldt-Jakob disease and fatal familial insomnia: Insights into phenotypic variability and disease pathogenesis. *Acta Neuropathol.* **121**, 21–37 (2011).

99. Mead, S., Lloyd, S. & Collinge, J. Genetic Factors in Mammalian Prion Diseases. *Annu. Rev. Genet.* **53**, 117–147 (2019).

100. WQ, Z. *et al.* Identification of novel proteinase K-resistant C-terminal fragments of PrP in Creutzfeldt-Jakob disease. *J. Biol. Chem.* **278**, 40429–40436 (2003).

101. G, S., MA, P., I, D., B, O. & JR, R. Scrapie prion protein structural constraints obtained by limited proteolysis and mass spectrometry. *J. Mol. Biol.* **382**, 88–98 (2008).

102. SB, P. *et al.* Scrapie prions aggregate to form amyloid-like birefringent rods. *Cell* **35**, 349–358 (1983).

103. Sanz-Hernández, M. *et al.* Mechanism of misfolding of the human prion protein revealed by a pathological mutation. doi:10.1073/pnas.2019631118/-/DCSupplemental.

104. Pilla, E., Schneider, K. & Bertolotti, A. Coping with Protein Quality Control Failure. *Annu. Rev. Cell Dev. Biol.* **33**, 439–465 (2017).

105. Levavasseur, E., Privat, N. & Haïk, S. In vitro Modeling of Prion Strain Tropism. *Viruses* (2019) doi:10.3390/v11030236.

106. Poggiolini, I., Saverioni, D. & Parchi, P. Prion protein misfolding, strains, and neurotoxicity: An update

from studies on mammalian prions. *Int. J. Cell Biol.* **2013**, (2013).

107. Walker, L. C. & Jucker, M. Neurodegenerative Diseases: Expanding the Prion Concept. *Annu. Rev. Neurosci.* **38**, 87–103 (2015).

108. Collinge, J. Mammalian prions and their wider relevance in neurodegenerative diseases. *Nat. 2016 5397628* **539**, 217–226 (2016).

109. Parchi, P. *et al.* Incidence and spectrum of sporadic Creutzfeldt-Jakob disease variants with mixed phenotype and co-occurrence of PrP Sc types: an updated classification. *Acta Neuropathol* **118**, 659–671 (2009).

110. Vaquer-Alicea, J. & Diamond, M. I. Propagation of protein aggregation in neurodegenerative diseases. *Annu. Rev. Biochem.* **88**, 785–810 (2019).

111. Spillantini, M. G. *et al.* α-Synuclein in Lewy bodies. *Nat. 1997 3886645* **388**, 839–840 (1997).

112. LI, B. *et al.* Aggregation and motor neuron toxicity of an ALS-linked SOD1 mutant independent from wild-type SOD1. *Science* **281**, 1851–1854 (1998).

113. G, S. *et al.* TDP-43 in skeletal muscle of patients affected with amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler.* **11**, 240–243 (2010).

114. Tian, Y., Meng, L. & Zhang, Z. What is strain in neurodegenerative diseases? *Cell. Mol. Life Sci.* **77**, 665–676 (2020).

115. Morales, R., Abid, K. & Soto, C. The prion strain phenomenon: Molecular basis and unprecedented features. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1772**, 681–691 (2007).

116. Telling, G. C. *et al.* Evidence for the Conformation of the Pathologic Isoform of the Prion Protein Enciphering and Propagating Prion Diversity. *Science (80-.).* **274**, 2079–2082 (1996).

117. Parchi, P. *et al.* Molecular basis of phenotypic variability in sporadc creudeldt-jakob disease. *Ann. Neurol.* **39**, 767–778 (1996).

118. Solforosi, L., Milani, M., Mancini, N., Clementi, M. & Burioni, R. A closer look at prion strains: Characterization and important implications. *Prion* **7**, 99–108 (2013).

119. RL, C. Encephalopathy in mice produced by inoculation with scrapie brain material. *Lancet (London, England)* **1**, 1378–1379 (1961).

120. Rossi, M., Baiardi, S. & Parchi, P. Understanding prion strains: Evidence from studies of the disease forms affecting humans. *Viruses* **11**, 1–27 (2019).

121. Parchi, P. *et al.* Classification of sporadic Creutzfeldt-Jakob disease based on molecular and phenotypic analysis of 300 subjects. *Ann. Neurol.* **46**, 224–233 (1999).

122. Baiardi, S., Rossi, M., Capellari, S. & Parchi, P. Recent advances in the histo-molecular pathology of human prion disease. *Brain Pathol.* **29**, 278–300 (2019).

123. Zerr, I. & Parchi, P. *Sporadic Creutzfeldt–Jakob disease*. *Handbook of Clinical Neurology* vol. 153 (Elsevier B.V., 2018).

124. Hirsch, T. Z., Martin-Lannerée, S. & Mouillet-Richard, S. Functions of the Prion Protein. *Prog. Mol. Biol. Transl. Sci.* **150**, 1–34 (2017).

125. van Rheede, T., Smolenaars, M. M. W., Madsen, O. & De Jong, W. W. Molecular Evolution of the Mammalian Prion Protein. *Mol. Biol. Evol.* **20**, 111–121 (2003).

126. H, B. *et al.* Mice devoid of PrP are resistant to scrapie. *Cell* **73**, 1339–1347 (1993).

127. A, P. *et al.* The prion gene is associated with human long-term memory. *Hum. Mol. Genet.* **14**, 2241–2246

(2005).

128. Tobler, I. *et al.* Altered circadian activity rhythms and sleep in mice devoid of prion protein. *Nat. 1996 3806575* **380**, 639–642 (1996).

129. Jones, E. *et al.* Identification of novel risk loci and causal insights for sporadic Creutzfeldt-Jakob disease: a genome-wide association study. *Artic. Lancet Neurol* **19**, 840–888 (2020).

130. Erginel-Unaltuna, N. *et al.* Distribution of the M129V polymorphism of the prion protein gene in a Turkish population suggests a high risk for Creutzfeldt-Jakob disease.

131. Sorce, S. *et al.* Genome-wide transcriptomics identifies an early preclinical signature of prion infection. *bioRxiv* 2020.01.10.901637 (2020) doi:10.1101/2020.01.10.901637.

132. Kim, T. K. *et al.* Core transcriptional regulatory circuits in prion diseases. *Mol. Brain* **13**, 1–14 (2020).

133. Hwang, D. *et al.* A systems approach to prion disease. *Mol. Syst. Biol.* **5**, (2009).

134. López González, I., Garcia-Esparcia, P., Llorens, F. & Ferrer, I. Genetic and Transcriptomic Profiles of Inflammation in Neurodegenerative Diseases: Alzheimer, Parkinson, Creutzfeldt-Jakob and Tauopathies. *Int. J. Mol. Sci. 2016, Vol. 17, Page 206* **17**, 206 (2016).

135. Llorens, F. *et al.* PrP mRNa and protein expression in brain and PrPc in CSF in Creutzfeldt-Jakob disease MM1 and VV2. *Prion* **7**, 383–393 (2013).

136. Xiang, W. *et al.* Cerebral gene expression profiles in sporadic Creutzfeldt-Jakob disease. *Ann. Neurol.* **58**, 242–257 (2005).

137. Zerr, I. *et al.* Updated clinical diagnostic criteria for sporadic Creutzfeldt-Jakob disease. *Brain* **132**, 2659 (2009).

138. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

139. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).

140. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **0 7**, Unit7.20 (2013).

141. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

142. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).

143. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* (2011).

144. Waskom, M. Seaborn: Statistical Data Visualization. *Seaborn* (2012).

145. Plotly Technologies Inc. Collaborative data science, https://plot.ly. *Plotly Technologies Inc.* (2015).

146. Harris, C. R. *et al.* Array programming with NumPy. *Nat. 2020 5857825* **585**, 357–362 (2020).

147. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods 2020 173* **17**, 261–272 (2020).

148. Rydbirk, R. *et al.* Assessment of brain reference genes for RT-qPCR studies in neurodegenerative diseases. *Sci. Reports 2016 61* **6**, 1–11 (2016).

149. Durrenberger, P. F. *et al.* Selection of novel reference genes for use in the human central nervous system: a BrainNet Europe Study. *Acta Neuropathol.* **124**, 893–903 (2012).

150. A, D. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

151. Kluyver, T. *et al.* Jupyter Notebooks—a publishing format for reproducible computational workflows. in *Positioning and Power in Academic Publishing: Players, Agents and Agendas - Proceedings of the 20th International Conference on Electronic Publishing, ELPUB 2016* (2016). doi:10.3233/978-1-61499-649-1-87.

152. Zhu, A., Ibrahim, J. G. & Love, M. I. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics* **35**, 2084–2092 (2019).

153. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *https://home.liebertpub.com/omi* **16**, 284–287 (2012).

154. Yu, G., Wang, L.-G., Yan, G.-R. & He, Q.-Y. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* **31**, 608–609 (2015).

155. Luo, W. & Brouwer, C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* **29**, 1830–1831 (2013).

156. Troll, C. J. *et al.* Structural Variation Detection by Proximity Ligation from Formalin-Fixed, Paraffin-Embedded Tumor Tissue. *J. Mol. Diagnostics* **21**, 375–383 (2019).

157. Di Stefano, M. *et al.* Analysis, Modeling, and Visualization of Chromosome Conformation Capture Experiments. *Methods Mol. Biol.* **2157**, 35–63 (2021).

158. S, G. *et al.* CHESS enables quantitative comparison of chromatin contact data and automatic feature extraction. *Nat. Genet.* **52**, 1247–1255 (2020).

159. Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods 2012 912* **9**, 1185–1188 (2012).

160. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods 2012 910* **9**, 999–1003 (2012).

161. Machado do Nascimento, P., Gomes Medeiros, I., Maia Falcão, R., Stransky, B. & Estefano Santana de Souza, J. A decision tree to improve identification of pathogenic mutations in clinical practice. doi:10.1186/s12911-020-1060-0.

162. M, T. *et al.* Identification of Recurrent Genetic Patterns From Targeted Sequencing Panels With Advanced Data Science: A Case-Study On Sporadic And Genetic Neurodegenerative Diseases. (2021) doi:10.21203/RS.3.RS-930395/V1.

163. Rodríguez-Martínez, A. B. *et al.* Molecular evidence of founder effects of fatal familial insomnia through SNP haplotypes around the D178N mutation. *Neurogenetics* **9**, 109–118 (2008).

164. Chen, C., Turnbull, D. M. & Reeve, A. K. biology Mitochondrial Dysfunction in Parkinson’s Disease-Cause or Consequence? (2019) doi:10.3390/biology8020038.

165. Flønes, I. H. *et al.* Mitochondrial respiratory chain deficiency correlates with the severity of neuropathology in sporadic Creutzfeldt-Jakob disease. *Acta Neuropathol. Commun.* **8**, 50 (2020).

166. Tzoulis, C. *et al.* HTRA2 p.G399S in Parkinson disease, essential tremor, and tremulous cervical dystonia. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E2268 (2015).

167. Olgiati, S. *et al.* PARK20 caused by SYNJ1 homozygous Arg258Gln mutation in a new Italian family. *Neurogenetics* **15**, 183–188 (2014).

168. Satoh, K., Abe-Dohmae, S., Yokoyama, S., St George-Hyslop, P. & Fraser, P. E. ABCA7 Loss of Function Alters Alzheimer Amyloid Processing. *J. Biol. Chem.* (2015) doi:10.1074/jbc.M115.655076.

169. Jones, E. *et al.* Identification of novel risk loci and causal insights for sporadic Creutzfeldt-Jakob disease: a genome-wide association study. *Lancet Neurol.* **19**, 840–848 (2020).

170. Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).

171. Liu, Y. *et al.* Association between NME8 Locus Polymorphism and Cognitive Decline, Cerebrospinal Fluid and Neuroimaging Biomarkers in Alzheimer's Disease. *PLoS One* **9**, e114777 (2014).

172. Karch, C. M. & Goate, A. M. Alzheimer's disease risk genes and mechanisms of disease pathogenesis. *Biological Psychiatry* (2015) doi:10.1016/j.biopsych.2014.05.006.

173. Laing, C. *et al.* The Application of Unsupervised Clustering Methods to Alzheimer's Disease. *Front. Comput. Neurosci. | www.frontiersin.org* **1**, 31 (2019).

174. Navarro, F. C. P. *et al.* Genomics and data science: an application within an umbrella. *Genome Biol. 2019 201* **20**, 1–11 (2019).

175. Xu, C. & Jackson, S. A. Machine learning and complex biological data. *Genome Biology* vol. 20 76 (2019).

176. Olzmann, J. A. *et al.* Parkin-mediated K63-linked polyubiquitination targets misfolded DJ-1 to aggresomes via binding to HDAC6. *J. Cell Biol.* **178**, 1025–1038 (2007).

177. Chen, D. *et al.* Parkin mono-ubiquitinates Bcl-2 and regulates autophagy. *J. Biol. Chem.* **285**, 38214–38223 (2010).

178. Klein, C. & Westenberger, A. Genetics of Parkinson's Disease. doi:10.1101/cshperspect.a008888.

179. Moreau, K. *et al.* PICALM modulates autophagy activity and tau accumulation. *Nat. Commun. 2014 51* **5**, 1–20 (2014).

180. Kanata, E. *et al.* RNA editing alterations define manifestation of prion diseases. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 19727–19735 (2019).

181. Llorens, F. *et al.* PrP mRNA and protein expression in brain and PrP c in CSF in Creutzfeldt-Jakob disease MM1 and VV2. (2013) doi:10.4161/pri.26416.

182. Hwang, D. *et al.* A systems approach to prion disease. *Mol. Syst. Biol.* **5**, 252 (2009).

183. Andres Benito, P., Dominguez Gonzalez, M. & Ferrer, I. Altered gene transcription linked to astrocytes and oligodendrocytes in frontal cortex in Creutzfeldt-Jakob disease. *Prion* **12**, 216–225 (2018).

184. Sorce, S. *et al.* Genome-wide transcriptomics identifies an early preclinical signature of prion infection. *PLOS Pathog.* **16**, e1008653 (2020).

185. Sastre, A. A. *et al.* Molecular Sciences Small GTPases of the Ras and Rho Families Switch on/off Signaling Pathways in Neurodegenerative Diseases. doi:10.3390/ijms21176312.

186. Chandra, T., Andrew, P., Fraser, P. & Correspondence, W. R. Global Reorganization of the Nuclear Landscape in Senescent Cells. *CellReports* **10**, 471–483 (2015).

187. Mabbott, N. A., Bradford, B. M., Pal, R., Young, R. & Donaldson, D. S. The effects of immune system modulation on prion disease susceptibility and pathogenesis. *Int. J. Mol. Sci.* **21**, 1–39 (2020).

188. Bradford, B. M. & Mabbott, N. A. Prion Disease and the Innate Immune System. *Viruses* **4**, 3389–3419 (2012).

189. Crespo, I., Roomp, K., Jurkowski, W., Kitano, H. & Del Sol, A. Gene regulatory network analysis supports inflammation as a key neurodegeneration process in prion disease. (2012).

190. Labzin, L. I., Heneka, M. T. & Latz, E. Innate Immunity and Neurodegeneration. *Annu. Rev. Med.* **69**, 437–449 (2018).

191. Choi, E. M. *et al.* Prion proteins in subpopulations of white blood cells from patients with sporadic Creutzfeldt-Jakob disease. *Lab. Invest.* **89**, 624 (2009).

ACKNOLEDGMENTS

The author wants to thank the families of the people affected by CJD who, despite the very painful time they were experiencing, chose to allow the use of the biologic material of their loved ones in research studies.

The author wishes to thank the all the institutions involved in this work; University of Bologna, Institute of Neurological Sciences of Bologna (ISBN), Center for Genomic Regulation (CRG) and National center for genomic analysis (CNAG) and in particular the professors who supervised this work: Professor Sabina Capellari, Professor Gastone Castellani and Professor Marc A. Marti-Renom and all the amazing members of their research teams. It has been a truly enriching experience to work with so many different colleagues and friends, a large amount of the things I learnt during these years would not have been possible without this environment.

Thanks to my friends, old and new, for easing this journey with our laughs, chats and wine. Special thanks to Matteo and Giulia for putting up with my overthinking and endless doubts of the last year.

Thanks to all my extended family and in particular my parents, Rita and Claudio, and my brother Andrea for their endless love and support.

Last but not least, I want to thank Paolo, for the love, support, laughs, courage of these last years and for being the best part of my days.