Trajectory-based differential expression analysis of genetic subtypes of Parkinson's Disease throughout differentiation to dopaminergic neurons

Master of Systems Biology

Luca Farinola
Student Number: i6326662
Email : l.farinola@student.maastrichtuniversity.nl
Supervisors : Dr. Lars Eijssen; Dr. Ehsan Pishva, MD
University : Maastricht University

Abstract

Commonly when dealing with disease we study biological processes in which cells are subject to change. A typical scenario is a differentiation process that occurs throughout development. In this project, data generated from human induced pluripotent cells (iPSCs) undergoing differentiation into neurons will be used to investigate genetic variations in Parkinson's Disease (PD). Those data have been collected from PD patients expressing different genotypes. The GBA N370S genetic variant, the LRRK2 G2019S variant, and healthy controls are of interest for comparison. Data have been obtained at three time points thus allowing analysis of gene expression patterns on a temporal spectrum. Various computational approaches have been developed for this purpose. A relatively traditional approach to analyzing this kind of data is using mixed effect linear models (MLMs). Still, MLMs do not fully represent a continuous progression of cells along a developmental trajectory. For this reason, Trajectory Inference methods (TI) are often used in single-cell sequencing to determine the pattern of a dynamic process experienced by cells. TI can be employed to establish more accurately how far in the developmental process is each data point, this is established through a value also known as 'pseudotime'. The advantage is that pseudotemporal ordering can be used as a dependent variable in our model to account for non-linear trends reflecting subtle changes in gene expression across time. Another possibility is represented by ANOVA-Simultaneous Component Analysis (ASCA), an extension of traditional mixed linear models. ASCA allows to decomposition of variance components associated with different experimental factors and their interactions providing a clear picture of how each factor (Time and PD subgroups) and the interaction contributes to the observed variation. The main goal of this thesis is to spot disparities in expression patterns of genes and pathways in distinct PD subgroups using computational and statistical tools including regression, dimensional reduction, and statistical inference. the limited number of time points and overall lack of statistical power in the data make it difficult to rely solely on hypothesis testing. ASCA modeling on the other side seems to be an interesting tool to examine such complicated multivariate datasets.

Contents

1 Introduction

1.1 Genetics mutations in PD

Parkinson's disease (PD) affects millions worldwide, causing progressive motor and various non-motor symptoms [14]. PD is a complex neuro-degenerative disease and it is impossible to establish a single direct cause. It is thought that a mix of hereditary and environmental variables influence how PD develops. After more than a decade of genetic studies, many monogenic variants of PD and various genetic risk factors that increase the likelihood of developing PD have been identified. With the advent of high throughput technologies like next-generation sequencing throughout the past 20 years, a lot of research has concentrated on genetic architectures to comprehend the pathophysiology of Parkinson's disease. At the molecular level, PD is characterized by impairment in intracellular trafficking followed by degeneration of the dopaminergic neurons [27]. Those neurons are very fragile due to their high energy demand. The increase in cellular stress is exacerbated by the formation of protein aggregation and inclusion bodies, a common phenomenon in many neurodegenerative disorders. Fibrillar aggregates called Lewy bodies (LBs) are a common signature of PD [57]. Recent studies suggest that several crucial events, for instance, abnormal accumulation of a protein called alpha-synuclein leading to the formation of LB and neuronal degeneration, may have a neurodevelopmental component [1].

From a clinical point of view, it is still very challenging to correctly characterize many neurological disorders due to their complexity and variability. In fact, PD shows distinct clinical manifestations. Motor features can vary significantly between individuals, for example, some patients present with tremors as the dominant and persistent motor feature while in some cases that is never experienced. Patients are also afflicted by a variety of non-motor characteristics, including cognitive impairment and sleep disruption [23]. Given this heterogeneity, efforts have been made to subtype PD patients into meaningful groupings. Genetics is an important piece of the puzzle and it is definitely an aspect of a better understanding of PD [45].

Studies have shown that individuals with GBA mutations have an increased risk of developing PD. The GBA gene encodes for the lysosomal enzyme glucocerebrosidase (GCase), which maintains glycosphingolipid homeostasis. Mutations in this gene can lead to an activation of stress response and contribute to neurodegeneration. GBA mutations are in fact frequently identified in Parkinson's disease and related LB diseases [54]. 5–15% of PD patients are linked to the GBA gene, considered the most important genetic risk factor for PD [50].

Another common risk gene for PD is the LRRK2 gene, which encodes the dardarin protein, a multi-functional protein with expression in dopaminoreceptive and other areas of the brain. Dardarin phosphorylates a broad range of proteins involved in multiple processes such as neuronal plasticity and vesicle trafficking [36].

Often when dealing with such a complex disease it is difficult to categorize and distinguish different phenotypic outcomes and PD makes no exception; on the other side studying subtle differences in those genetic variants can be very useful for a more precise understanding of the disease at a molecular level [23].

1.2 Brief Overview of the Project

Following a preliminary examination of the data, the first task will be to become acquainted with computational methods. Familiarizing with the main features of the different packages to perform TI, for instance, is critical to standardizing the results and to compare the methodologies. Different packages for some trajectory inference (TI) methods of choice will be used to investigate the temporal development of each genotype. TI are designed for single-cell data, therefore it is important to take into consideration that the different samples are treated as cells. Mixed effect linear models (MLMs) will be employed both using pseudotime values to identify genes that show differential expression along the different inferred trajectories as well as with the three time points as categorical variables. Through this it is possible to evaluate whether TI offer any added value. Lastly ANOVA-Simultaneous Component Analysis (ASCA) model is used to obtain results from an additional alternative approach. The main goal of this study is to investgate witch features (genes) are more relevant in explaining the interaction of time and genotype, in other words how genes can explain differences among the given genotypes (GBA, LRRK2 amd healthy controls) across time. The thesis will discuss the results obtained in the three different approaches employed for this analysis shown in Figure 1.

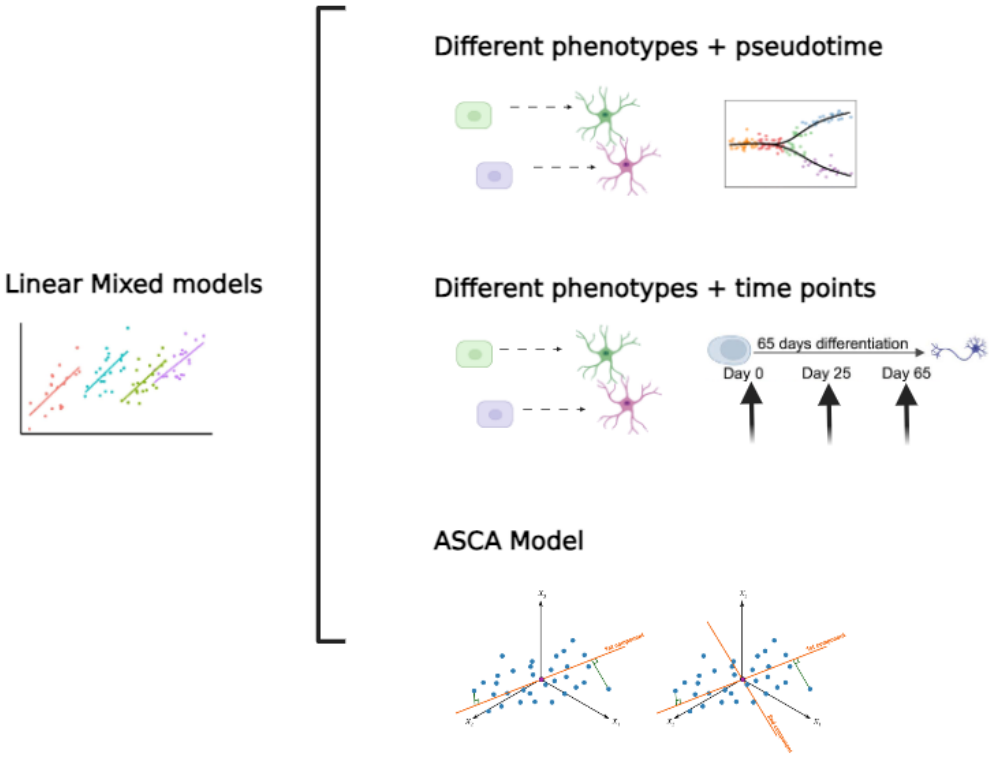


Figure 1: Overview of the Project plan. Main approaches explored during this thesis project, image created with Biorender

1.3 Mixed linear models

MLMs are an advanced and flexible method for evaluating data that have nested or hierarchical structures, which are frequently seen in a variety of situations, including, of course, experimental designs in biomedical studies. The fundamental concept underlying mixed effects models is the recognition that data often exhibit correlation and dependence structures due to hierarchical or repeated measures designs [11]. For instance, in this case, as we are dealing with of time course data it is necessary to consider multiple measurements per subject, meaning observations within the same subject tend to be more similar than those across different subjects. Mixed effects models can address this correlation by partitioning the variability into fixed effects, which explain systematic trends, and random effects, which account for subject-specific variations. By adding both fixed and random effects this technique expands on the basic linear regression framework [21].

1.4 trajectory inference analysis

TI methods are employed in the context of omics data processing to position data points along a continuous temporal space, whether they be samples or cells in the case of bulk or single-cell RNA sequencing. The big merit of this kind of analysis is the possibility of identifying genes according to their dynamic changes in expression over time. To do so it is necessary to sample at multiple time-points and obtain snapshots of the gene expression profiles. This has a huge relevance when taking into consideration complex processes as in the case of neuronal differentiation. Through TI it is possible to use statistical methods to draw one or more trajectories on a lower dimensional space which represents the underlying developmental process [49]. Samples are ordered on a continuous timeline, this is referred to as "pseudotime". Different kinds of TI methods have been developed and there are some key aspects to distinguish those from one another. These methods may differ in terms of the underlying mathematical models, assumptions about the nature of trajectories, and strategies for capturing the complexity of cellular processes and may follow different computational solution in each step of the analysis (Figure 2). For instance, the method may force the trajectory in a specific topology (linear, bifurcating, circular) or it may not. The kind of algorithmic approach it uses (tree, graph, cluster) should be taken into consideration when choosing what kind of method to pick [48].

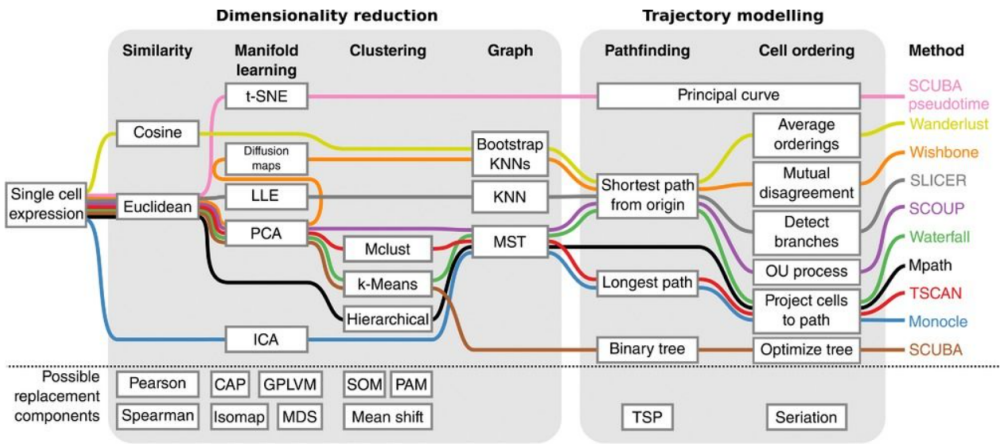


Figure 2: Comparison between TI Methods Various methods and tools for each step in trajectory analysis (TI) workflows are highlighted, showing the flexibility and interchangeability of components in the analysis pipeline.Source

1.5 ASCA

ANOVA-Simultaneous Component Analysis, or ASCA, is a multivariate statistical technique that combines the advantages of Principal Component Analysis (PCA) and Analysis of Variance (ANOVA). It is designed for analyzing data from complex experimental designs. This method is particularly powerful to identify and characterize patterns of variation across multiple factors or conditions (Figure 3) [28], as in our case, time points and (genetic/disease) subgroups. ASCA represents a sophisticated yet accessible approach for dimensionality reduction and pattern recognition particularly suited for focusing on developmental trajectories, making it a valuable addition to the analytical toolkit for this thesis.

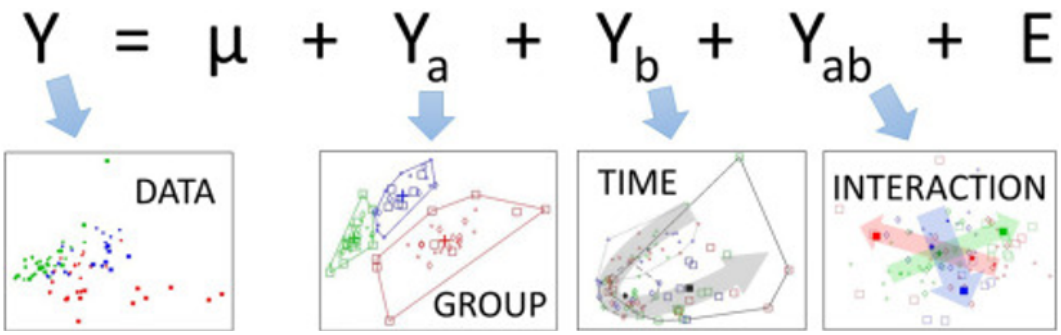


Figure 3: ANOVA-Simultaneous Component Analysis (ASCA) main gpourpose is to separate the different effects into separate matrices to study the effects [5]

1.6 Research Question(s)

This thesis' primary goals, which came from methodological and biological perspectives, respectively, can be summarized as follows :

- Do the iPSCs cells that are affected by different PD mutations (LRRK2, GBA) follow distinct developmental trajectories upon neuronal differentiation ?
- Is it possible to identify genes to characterize those different mutations ?
- Is continuous temporal ordering a valuable addition to linear modelling ?
- Is ASCA modeling a possible solution to explore the effect of the interaction of time and genotype?

1.7 Significance

It is estimated that by 2030, there will be twice as many people (8.7-9.3 milion) with Parkinson's disease (PD) in Western Europe's five most and the world's ten most populous nations as there were in 2005 (4.1–4.6 million) [16].To fight the burden caused by this trend, in the era of big data molecular biologists, bioinformaticians and neuroscientists are getting new insights into complex diseases, such as PD. The relevance of this project relies on the ability to characterize the different genetic sub-genotypes of PD. By unraveling the specific genetic factors contributing to an individual's Parkinson's phenotypic variant, we would get a better understanding of the underlying pathophysiological mechanism. Furthermore, based on this type of results tailored therapeutic approaches can be designed, optimizing treatment strategies and potentially both for improving outcomes of treatments on patients based on their unique genetic profiles and detecting the desease at an earlier stage [23] [45] (Figure 4).

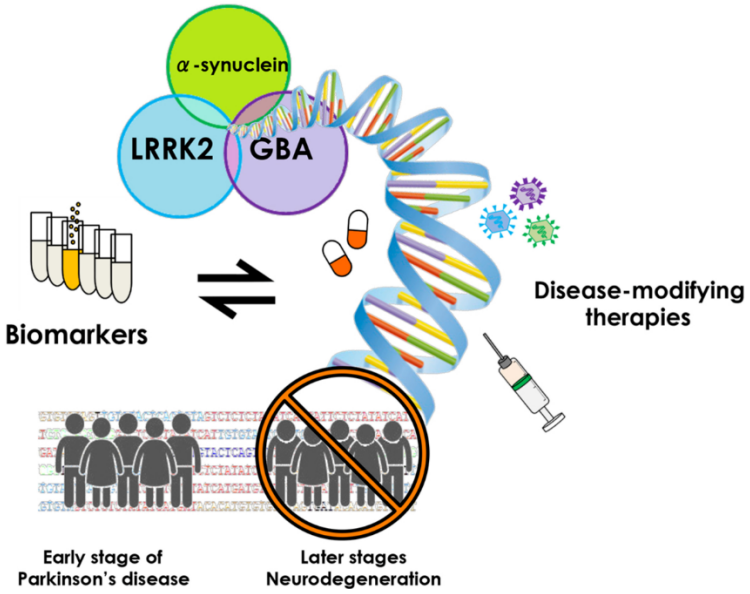


Figure 4: There is ample evidence that each of the protein products of these genes may also play a different role in multiple aspects of disease pathogenesis and/or progression, a personalized medicine approach to target those specific biomarkers is needed [45]

2 Methods

2.1 Data Acquisition

Collection and analysis of omics data play a huge role in this process and many projects provide access to those kinds of data. That is the case for the Foundational Data Initiative for Parkinson's Disease (FOUNDIN-PD) [8], an international, collaborative, multi-year project to create different kinds of omics data for Parkinson's Disease (PD). FOUNDIN-PD project provides genetic, epigenetic, regulatory, transcriptomic, and longitudinal cellular imaging data from iPSC-derived differentiating dopaminergic (DA) neurons to understand molecular mechanisms and disease-associated genetic variation (Figure 5). For this thesis, already processed bulk RNA-seq data will be used. Those data were generated on day 0, day 25, and day 65 with each time point including five technical replicates of the control line. For library preparation, RiboGone and SMART were employed to remove ribosomal RNA (rRNA) and synthesize cDNA. Illumina-compatible libraries were then obtained through SYBR Fast qPCR amplification and AMPure Bead Purification. Using a NovaSeq 6000, 100 bp $\times$ 100 bp paired-end libraries were sequenced and measurements were performed at the gene and transcript levels detecting Protein-coding RNA, lncRNA, and other non-coding RNA. The reads counts were then finally generated by using cutadapt for the removal of adapter sequences, primers, and poly-A tails. STAR [15] and salmon[44] were respectively used for alignment and quantification.

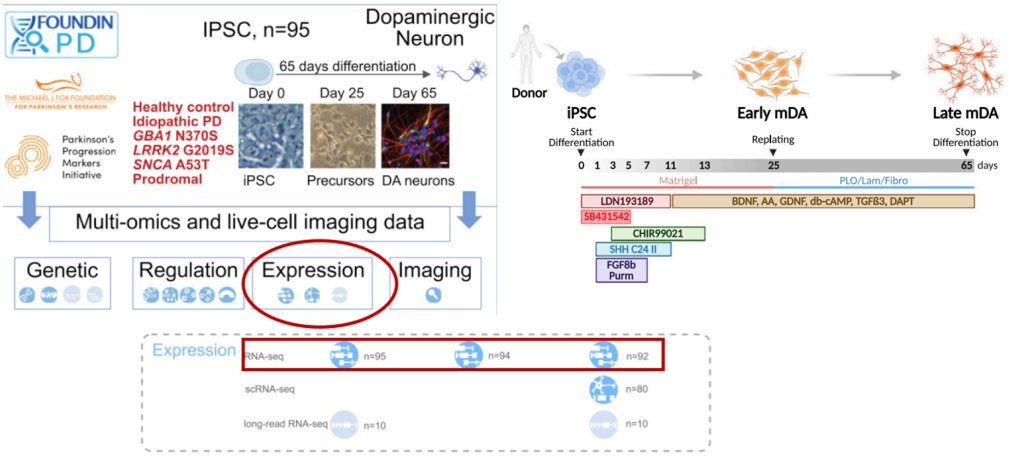


Figure 5: General overview of the types of data collected and the PD genotypes (upper left). A schematic overview of the differentiation protocol (on the right). Types of omics data collected to study gene regulation, note bulk rnaseq is used in this thesis source :[8]

2.2 Data Pre Processing

Once RNA-seq derived count are obtained the first step is to perform some basic pre-processing on the count table. Genes were filtered using the filterByExpr() function from EdgeR [47] which was set to retain only genes with a total count of 10 or more across all samples and having a minimum proportion of samples in which it must be expressed of 80 %. After this filtering, data were normalized and log 2 transformed using counts per million (CPM). When performing CPM number of raw reads mapped to a transcript is multiplied by the number of sequencing reads in that sample and then multiplied by one million, therefore normalizing for sequencing depth, allowing comparison across samples [10]. To adjust for unwanted variability surrogate variable analysis (sva) [34] is performed on the dataset. This analysis allows to compute several surrogate variables that are used to fit linear models, calculate and retain residuals to remove all sources of latent variation.

2.3 Trajectory Inference

The main goal of using TI methods is to account for developmental trends observed in gene expression values for the cell differentiation process. This is achieved as the trajectory inferred when connecting the clusters, which generally should represent different time stages, is curved to fit the data points. Samples will then be mapped into this fitted trajectory to calculate the pseudotime. This value can then be used as an alternative (continuous) time variable to be taken into account when performing statistical analysis.

The first step when performing this analysis is to define the space onto which the trajectory is calculated. For this purpose, a PCA model was constructed using only Healthy controls. Other conditions are projected in the same space, in such a way emphasize a comparison to a 'reference' represented by the healthy condition. PCA was chosen for the calculation of pseudotime for all the methods that are mentioned in this thesis even though PCA as base for trajectory modelling is not necessarily the only option when dealing with TI analysis.

The main TI methods chosen are Slingshot, TSCAN, and Scorpius. Those are particularly suited for the extraction of a linear (non-bifurcating) trajectory in different kinds of datasets (including of course bulk-RNA) and are flexible when it comes to choosing what kind of dimensional reduction can be used (PCA is picked for all of them). The main two algorithmic approaches explored to run that analysis are Minimum spanning trees (MST) (TISCAN, Scorpius) and space-filling curve (Slingshot) run in combination with different clustering methods including k-mean (Scorpius) and mclust (TSCAN, Slingshot). Pseudodtime values are calculated for each condition separately using the two-dimensional space obtained from the projection to the first two principal components of the healthy controls

PCA model. In principle, those methods could be used either within the dynverse package or using the original packages associated with each method. Those options have benefits and downsides therefore both approaches are followed. To assess potential differences among the methods pseudotime values are compared. A linear model was fitted using the pseudotimes values for each combination of methods (slingshot vs scorpius, TSCAN vs slingshot etc...).

2.4 Mixed Effect Models

For linear models, each gene is extracted from the processed count matrix to fit separately. gene expression values are treated as the dependent variable and combined with metadata given for each sample (ID of the patient and genotype) and pseudotime values. The lmer() function, from lme4 [3], is then employed to fit the linear mixed-effects model specified by

$$y \sim \text{time} + \text{genotype} + \text{time:genotype} + (1 \mid ID) \tag{1}$$

y is therefore modeled as a function of the interaction between time and genotype. The model was run twice, with time being either the continuous pseudotime value for each sample or the timepoint labels given in the metadata. The ID of the patients represents a random effect so that the model can correct for repeated measures. Healthy controls are given as reference group. Additionally, post-hoc multiple comparisons using Tukey's method [4] is performed using the multicomp library [25]. This is done to rigorously examine pairwise differences between the subgroups. The necessity for such a post-hoc test arises from the need to identify differences genes not only deviating from the healthy controll but also showing different patterns when comparing the specific conditions (GBA and LRRK2).

Given the multiple comparison test there are mainly three values that are useful to characterize the different conditions. p-values of the interaction term from the lme4() linear model. Given healthy control are taken as a reference p-values are calculated for both LRRK2 and GBA and those will give an idea whether each condition differ from the reference or not. On the other side the multiple comparison done with the glht() function from the multcomp package outputs p-values to compare the two genetic mutations. Once the values are sorted for multiple comparison testing, p-values of the interaction term from the LMMs indicates to witch extent the expression profile differ from the referece, thus allowing to understand wether genes appear to be ore or less expressd in the different conditions.

2.5 ASCA

The main two options to run the ASCA method in R are limpca [53] and ALASCA [29]. Despite those having both the main underline principles some key differences are important to consider. Firstly ALASCA can include random effects, as the algorithmic approach used in ALASCA is the repeated measures-ASCA+ (RM-ASCA+) [38] as opposed to the ASCA/APCA family of methods implemented in limpca. Additionally, ALASCA has the important advantage of Incorporating validations through bootstrap, allowing for more robust results.

Firstly RM-ASCA+ computes linear regression, producing regression coefficients. ALASCA includes three main options to run estimate regression coefficients from linear models; Rfast [43], lm, and lme4 (the last two respectively used depending on whether there is the need to include random effects). lme4 was chosen using the same formula reported above 1. The second step in RM-ASCA+ is to decompose the X matrix into effect matrices representing specific parts of the regression model as follows.

$$\mathbf{X} = \mathbf{X}_{\text{time}} + \mathbf{X}_{\text{genotype}} + \mathbf{X}_{\text{time:genotype}} + \mathbf{E} \tag{2}$$

The last stage is applying PCA to individual or combined effect matrices, depending on the research objective, and extracting scores and loadings. Plotting the generated scores and loadings allows one to see how the chosen effects impact the various effects.

Time and group effects are analyzed as separate units, by selecting separate_effects = TRUE. The effect matrix for time is separated from the interaction matrices (time + genotype + time:genotype Eq. 2) and analyzed separately and two sets of scores and loadings are extracted. The first one describing the development of the reference group (healthy control), the second, instead, how the other groups diverge from the reference by avaraging out the first set.

As we are working with transcriptomic data, bootstrapping can be extremely computationally expensive. ALASCA provides the possibility to work with a lower number of dimensions by applying an initial PCA to reduce the number of variables before performing linear regression and automatically transform back loadings to the original variable space for interpretations as shown in Figure 6.

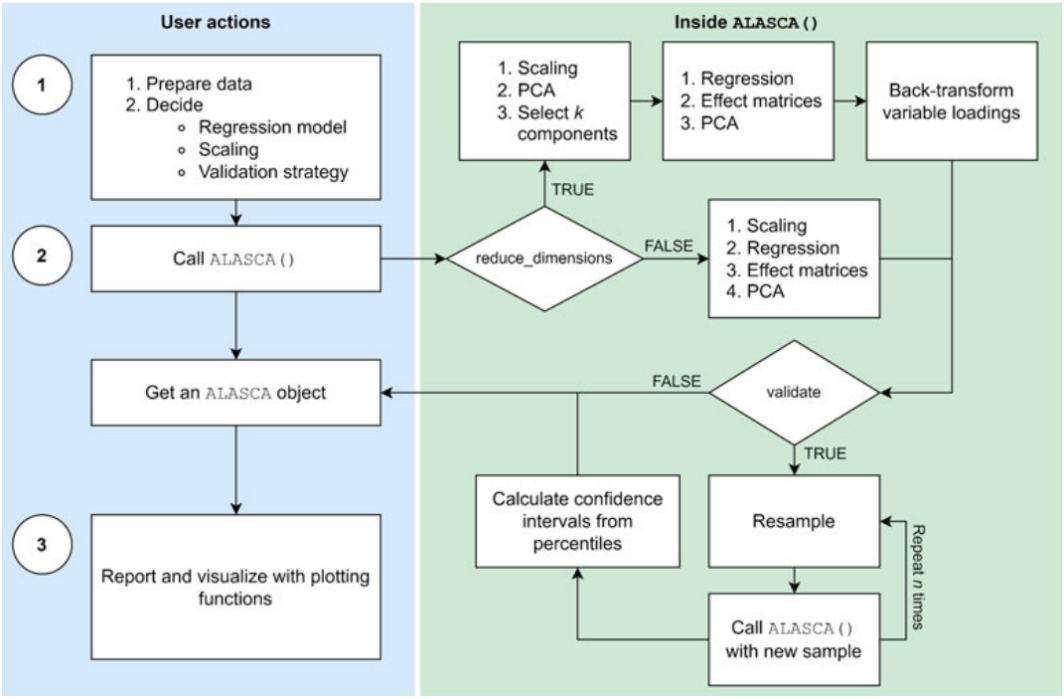


Figure 6: Three main staiges of the ALASCA package are described : (1) Preparation, (2) execution, and (3) visualization. Data need to be prepared in the long format. The ALASCA() function will then scale the data, perform MLM, apply principal component analysis (PCA) on the decomposed effect matrices, and extract loadings and scores. The option reduce_dimensions = TRUE will use PCA to reduce the number of variables, and loadings are automatically transformed back to the original variable space. Validation through bootstrap is performed with validate = TRUE . percentiles from the repeated validation rounds are used to calculate confidence intervals for loadings and scores.

3 Results

3.1 Data Pre Processing

After filtering, out of the 94946 transcripts (more than 50% of witch are lnc-RNA as shown in Figure 24) only 18683 are retained, representing the 20 % of all genes that were initially measured. As expected The resulting PCA plot displays a clear separation of the three time points. However, some unexplained sources of variance are also observed horizontally along PC2 throughout all time points (Fig 7 A). With surrogate variable analysis it was possible to estimate 24 significant surrogate variables out of which only 5 are used to remove the undesired variability. The reason to choose only 5 surrogate variables is that five are the number of cell types estimated from the single cell dataset analysis performed in the FOUND-IN PD original paper [8]. Additionally by performing this analysis multiple time changing the number of svs that are regressed out each time, five appeared to be the lowest possible number of svs to remove artifacts from the data and simultaneously avoid manipulating the data more than necessary. To fit the linear model also sex and age are included together with the first 5 surrogate variables. Finally residuals are kept so that the effects of the surrogate variables and known covariates are removed (Figure 7)

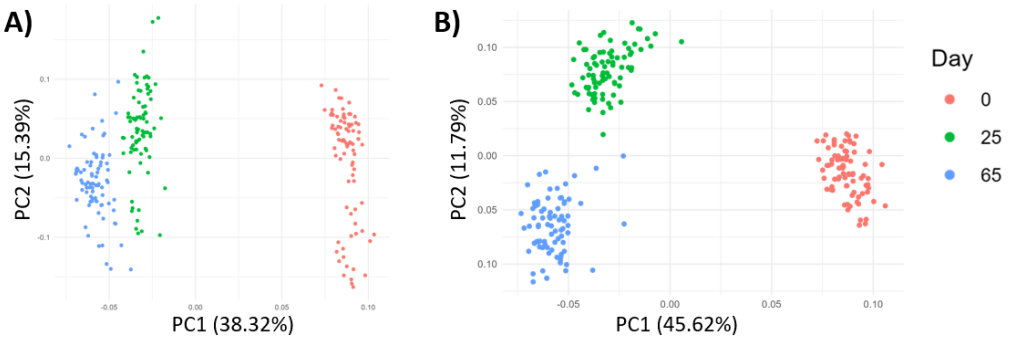


Figure 7: Effect of pre processing on the bulk-RNA sequencing filtered data seen through principal component analysis (PCA). Data displayed before (A) and after (B) surrogate variable analysis (sva)

The next step before moving to Trajectory Inference (TI) analysis is to use pre-processed data to generate the healthy control (reference) PCA space and project onto it the data from the samples of the other conditions as well. The pre processed count matrix is therefore subsetted to run PCA only with the control data points and project each specific PD mutation genotype afterwards. The projected samples do not show any obvious difference to their reference (Figure 8).

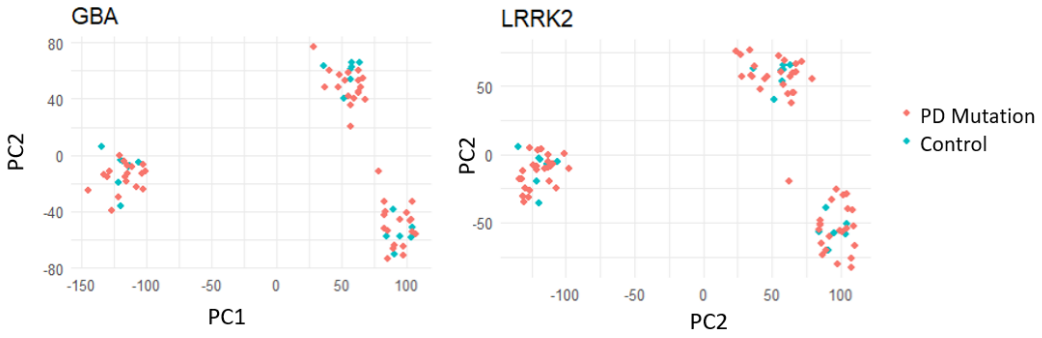


Figure 8: PCA plot after projecting onto healthy control PCA space respectively GBA (left) and LRRK2 (right) samples

3.2 Trajectory Inference

Despite being particularly handy for comparing TI methods using a single pipeline, the dynverse seemed to not provide consistent results for the various methods, in some cases providing output that were different from the one obtained with the original packages. The ease of use in dynverse comes with the main disatvantage of being extremely complicated to fully understand and debug in case of errors, as it is often the case for wrap up methods. The idea of using the methods within the dynverse was therefore abbandoned in favour of the three original packages of scorpis, TSCAN, and slingshot.

As already mentioned the three methods differ in terms of algoritmic approach for the inference of the trajectory and clustering. Given the strong separation observed across time, the clustering algorithm should not have an effect on the end result. One difference is the pseudotime being a continuous value ranging from 0 to 1 in Scorpius and 0 to 400 in Slingshot and a rank-based value ranging from 1 up to the number of samples for TSCAN. However, when comparing the outputs of the different packages for the calculation of pseudotime the three methods provide consistent pseudotemporal ordering (Figure 9). The main difference is given by the way TSCAN orders by ranking the samples rather than assigning a continuous value.

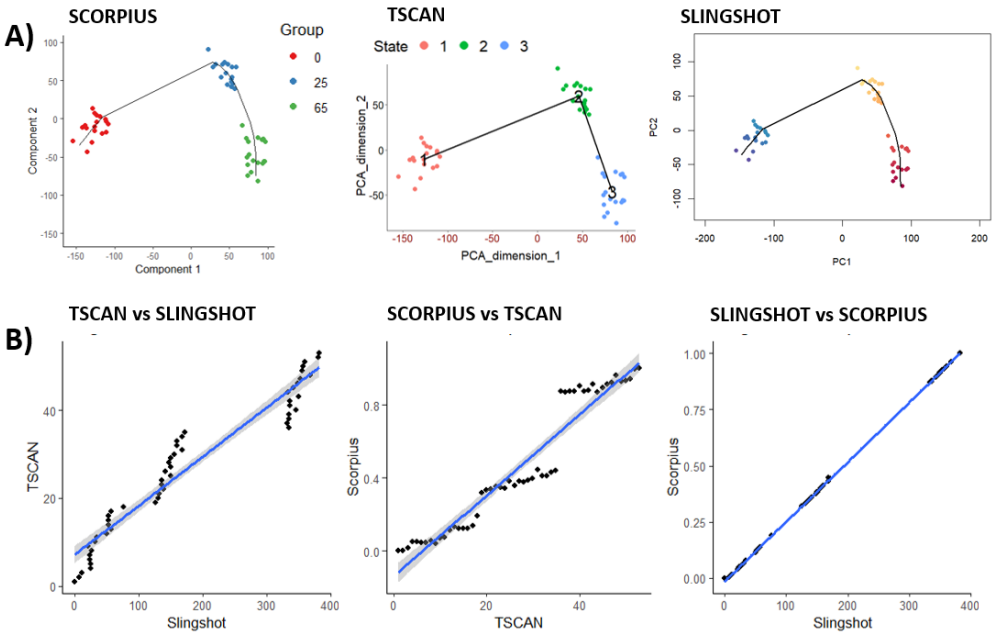


Figure 9: Visual representation of pseudotemporal trajectories identified by three different tools (A) Pairwise comparison of pseudotime orderings (B). Each scatter plot shows a strong correlation between the pseudotime values assigned by each pair of tools, with a fitted regression line (blue) and confidence interval (shaded area)

The ordering provided by TSCAN is not ideal for the pourpose of the thesis. Slingshot and Scorpius appear to be (besides scaling) very similar and would very likely provide the same results. Compared to Slinghsot, Scorpius result to be more flexible, therefore the choice falls on the latter. The trajectories inferred by Scorpius for the different conditions follow the same overall trend (Figure 10).

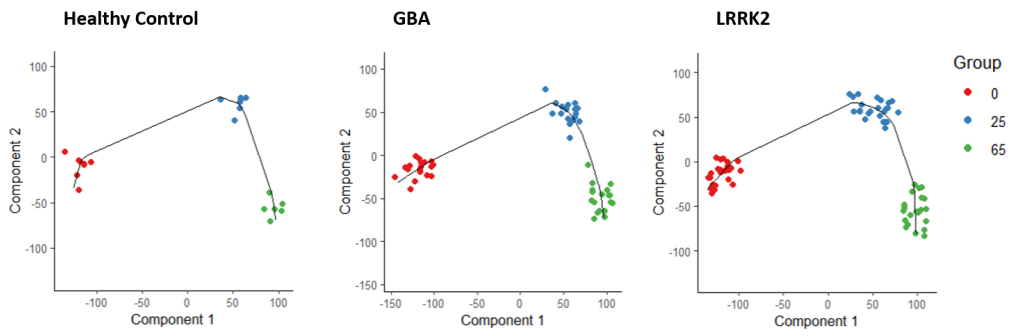


Figure 10: Trajectories inferred through Scorpius of the three different conditions healthy controls, GBA and LRRK2. Note that LRRK2 and GBA data points are the same those displayed figure 8

3.3 Mixed Effects Models

As anticipated, linear modelling was performed using both numerical pseudotime values and original time point labels used as categorical variables. After performing hypothesis testing for multiple comparison it results that, in both cases there is not enough statistical power to adjust p-values using FDR correction. Few genes are selected by simply setting a threshold to p-values obtained from the Tukey multiple comparison test. Time groups are very separate and pseudotime seem to appear in three main different shades of blue, just as the categorical values of the time points. Interestingly results from the two models do not overlap completely and there are genes found with a p-value lower than 0.05 in only one of the two models and vice versa. In order to visualize result clustered heatmaps using the pheatmap [32] library are used (Figure 11).

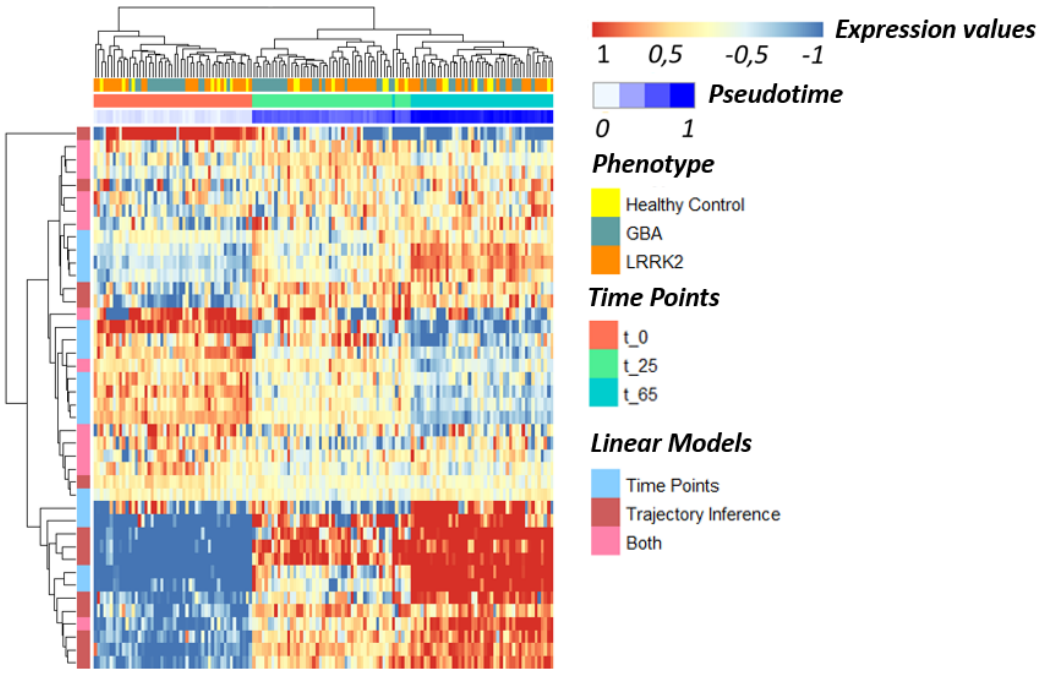


Figure 11: Heatmap representing genes with p-values < 0.05 in at least one model across different pseudotime points and subgroups. The columns are annotated by pseudotime values, reflecting the inferred trajectory of cell differentiation, time points, and genotype indicating different experimental conditions. Rows are annotated according to genes that are selected using p-values of linear models that uses time points, pseudo-temporal values or both

Despite the obvious effect of time, it is impossible to establish any clear difference between the phenotypic conditions from Figure 11. A closer examination of the horizontal annotation in the heatmap above reveals that genes present in both models (violet) are changing less dramatically over time. Those are, in fact, primarily located in the heatmap's top region. To examine various gene expression patterns in more detail separate heatmaps are generated (Figures 12, 13, 14). Both conditions (Healthy Control, GBA, and LRRK2) and Linear model results (genes relevant in LMMS using Trajectory Inference, Time Points labels or found in both) are isolated. The columns are arranged to match the pseudo-time values as the goal is to examine patterns across time. For readability, gene with ow p-value only in Time Points will be called TP, those with a low p-value only in Trajectory inference will be called TI and those with a low p-value in both models as common genes

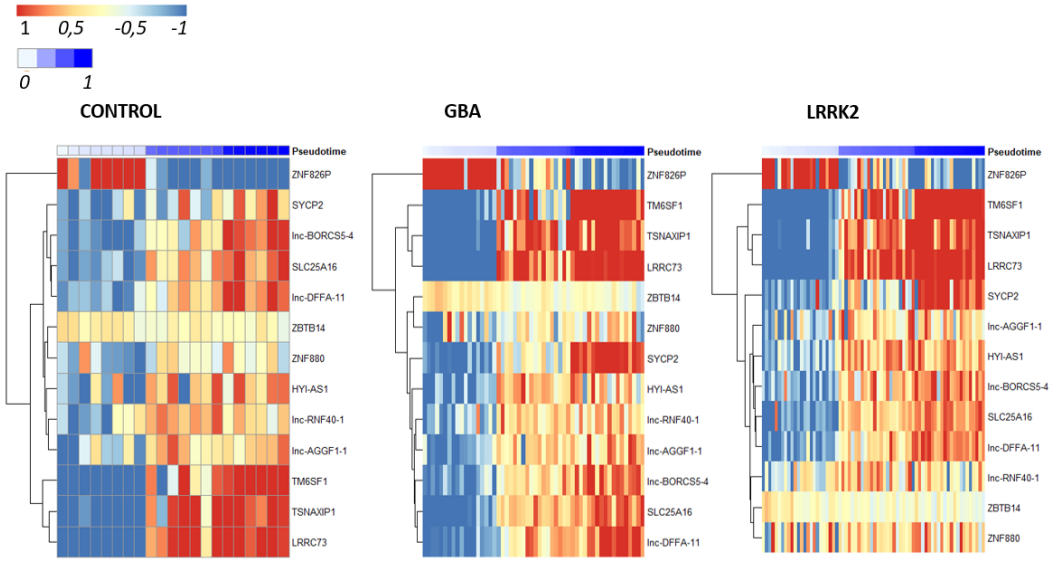


Figure 12: Heatmap representing genes with p-values < 0.05 only in the linear model that uses trajectory inference pseudo-time values

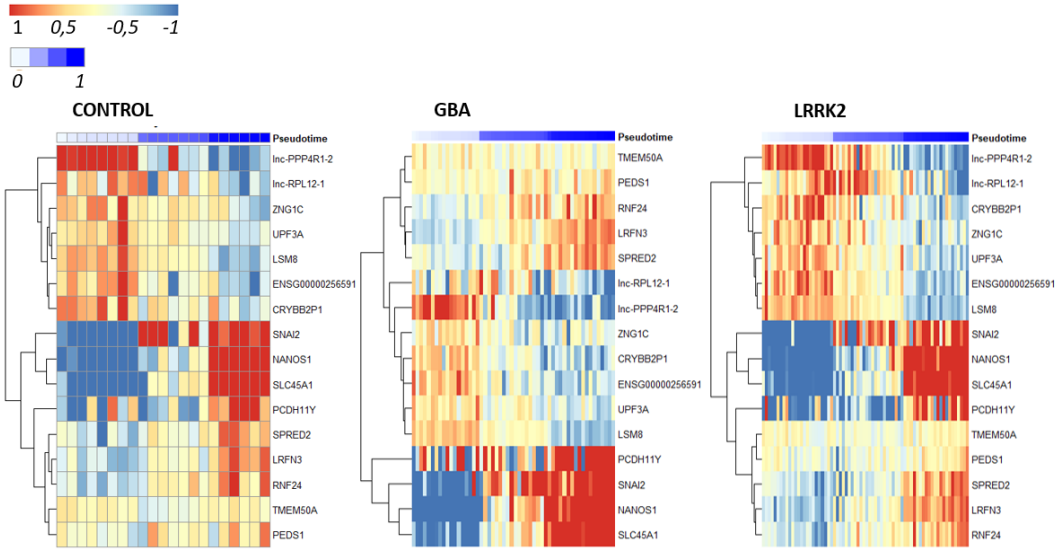


Figure 13: Heatmap representing genes with p-values < 0.05 only in the linear model that uses time point categorical labels

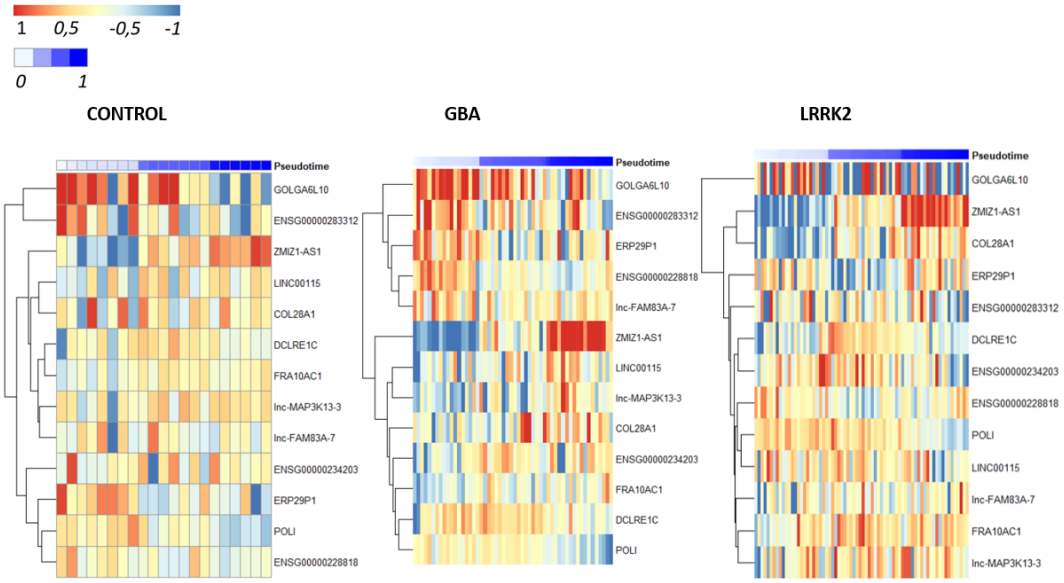


Figure 14: Heatmap representing genes with p-values < 0.05 in both models

Some common patterns are observed for genes relevant in exclusively one of the two LMMs (TI and TP genes). While for TP genes there is a roughly 50-50 ratio of genes increasing and decreasing over time, TI genes seem to be mostly increasing with the only exception of ZNF826P.By looking at the dendrogram the main differences observed across conditions in the first two sets of heatmaps are mainly how those genes cluster together

Nevertheless, for TI and TP genes it is possible to pintpoint at areas of the heatmap showing these different trends. For instance TI genes (Figure 12), TM6SF1 TSNAXIP1 LRRC73 are strongly increasing and occupy a different 'position' in healthy control with respect to GBA ans LRRK2. For TP genes (Figure 13) this is the case for genes SNA2, NANOS1, SLC45A1. and PCHD11Y as well as, with the opposite trends, for lnc-RPL12-1 and lnc-PPP4R1-2. On the other hand there does not seem to be any specific clustering according to time in common genes (Figure 14). Those, with the only exception of GOLGA6L10, do not seem to follow the same clustering patterns.

Differentially expressed genes were identified for every pairwise comparison using the fitted linear model and a p-value < 0.05 that was adjusted for multiple group comparisons (Table 26). The p-value was instead not corrected for multiple testing due to insufficient statistical power. By plotting the count values as boxplots it is interesting to see how those genes behave across the three time points (Figure 15, 16, 17), those figures report differentially expressed genes filtered for multiple comparison in the following contrast combinations : LRRK2 vs Healthy Control, GBA vs Healthy Control, and GBA vs LRRK2. Main differentially expressed genes in the first contrast are COL28A1, ERP29P1 , LINC00115 and lnc-RNF40-1. similarly ZMZ1-AS1, SYNCP2 and NANOs1 are differentially expressed between GBA and Healthy Controls. Genes FRA10AC1 and lnc-MAP3K13-3 do not appeat to be relevant in none of the conditions if compared to the controls despite beeing the two genes with the lowest p-value when it comes to the Tukey corrected multiple comparison p-value (LRRK2-GBA).

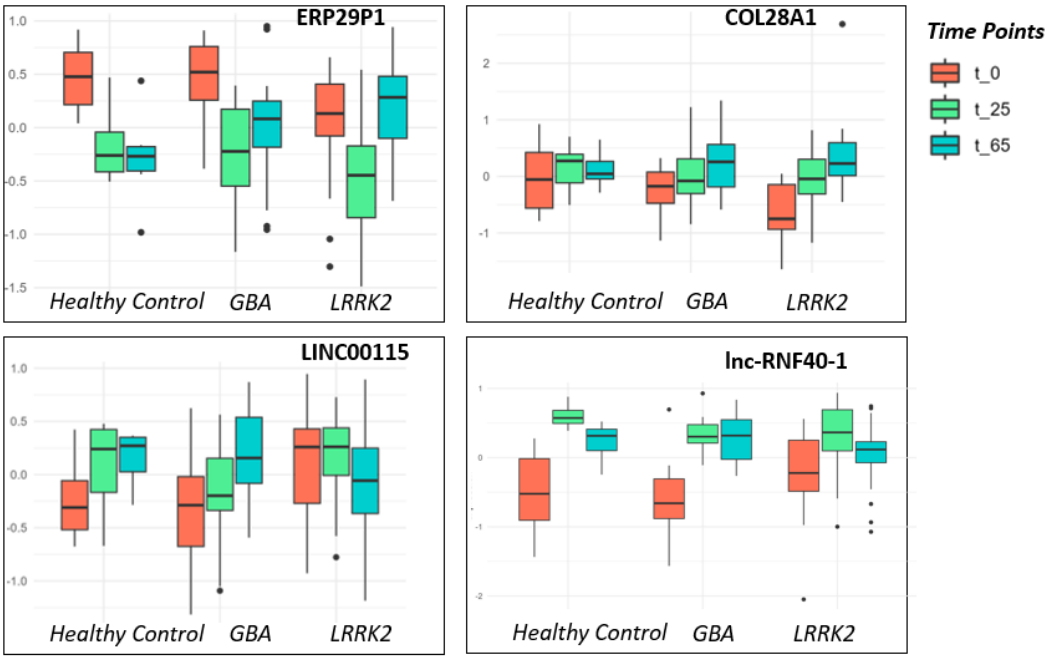


Figure 15: Boxplot of genes with a p-value < 0.05 in mixed effect linear models (GBA vs Healthy Control)

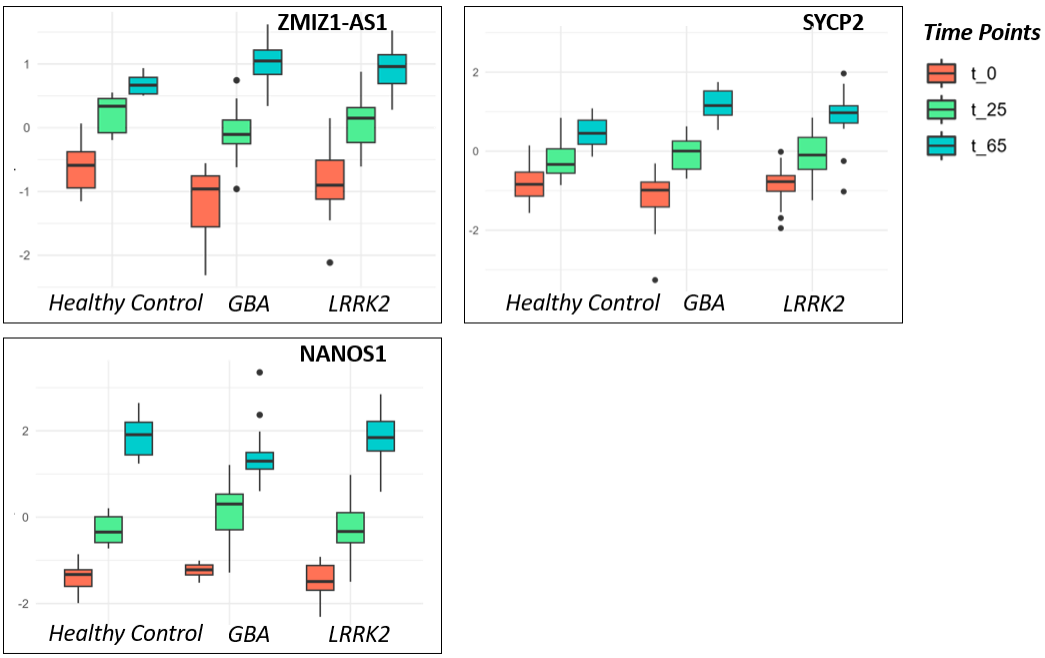


Figure 16: Boxplot of genes with a p-value < 0.05 in mixed effect linear models (LRRK2 vs Healthy Control)

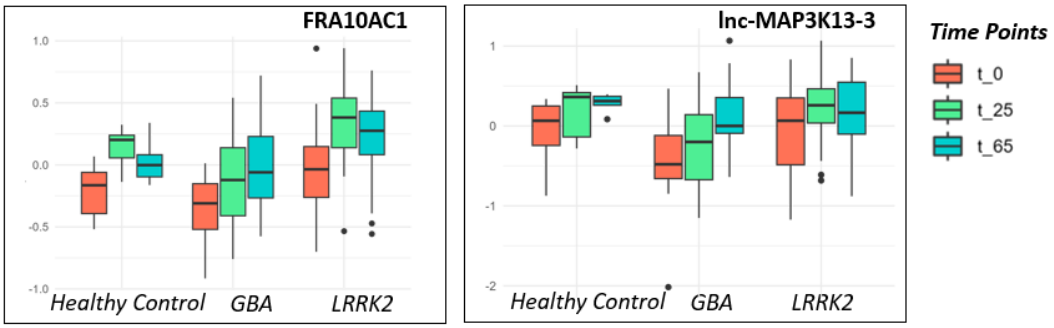


Figure 17: Boxplot of genes with p-value < 0.05 in multiple comparison (LRRK2 vs GBA)

3.4 ASCA

ASCA can represent an additional method to explore patterns in this dataset from a slightly different angle. The main point of ALASCA [29] is to separate the effect of time, genotype, and their interaction. Firstly the effect of time for the healthy control reference is isolated. To functionally characterize the main contributing genes, Gene ontology is performed on high loadings from the first principal component (Figure 18).

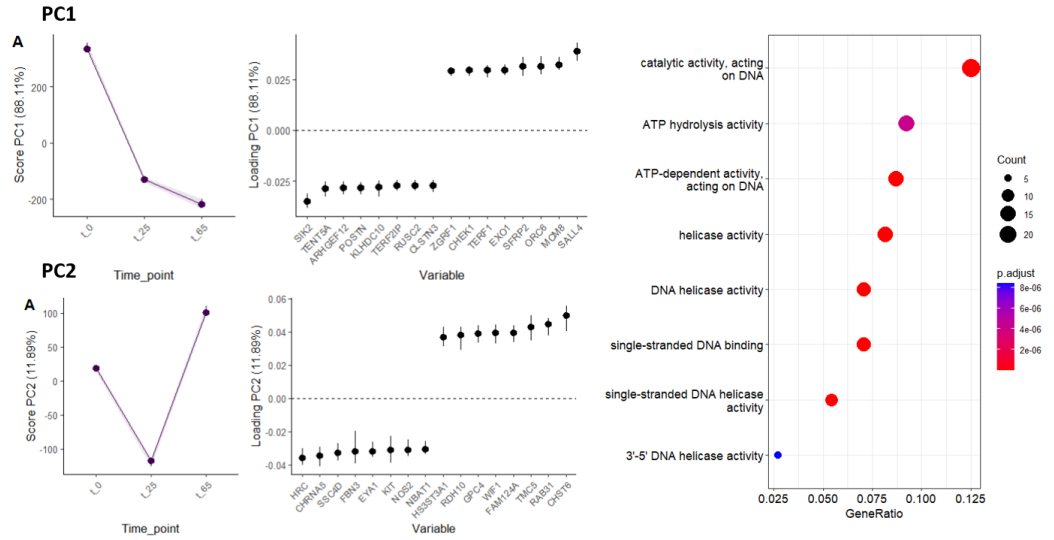


Figure 18: Time development of healthy controls is isolated, general trend of PC scores is shown on the left, higher loadings in the middle and gene ontology of those high loadings genes on the right

Interestingly, once the effect of time over the reference is taken out, by studying the phenotype and interaction effects, it is possible to sketch some differences between the groups. The first two principal components explain 82 % of the variation, respectively 46.65 % and 36.29 % (Figure 19).

By inspecting the scores of the first principal component there seems to be a separation between the two conditions and healthy controls while GBA and LRRK2 start to deviate from one another in the second principal component, especially at the later stages.

When a gene exhibits substantial changes in score across a given PC by having strong loading, it is helpful to see the marginal means derived from the underlying regression models (Figure 20). From those selected in the second principal component, it is clear how genes having positive loadings will display the same, or a similar, pattern as shown in Figure 19 C (PC2) while those having negative loadings will tend to have an opposite trend.

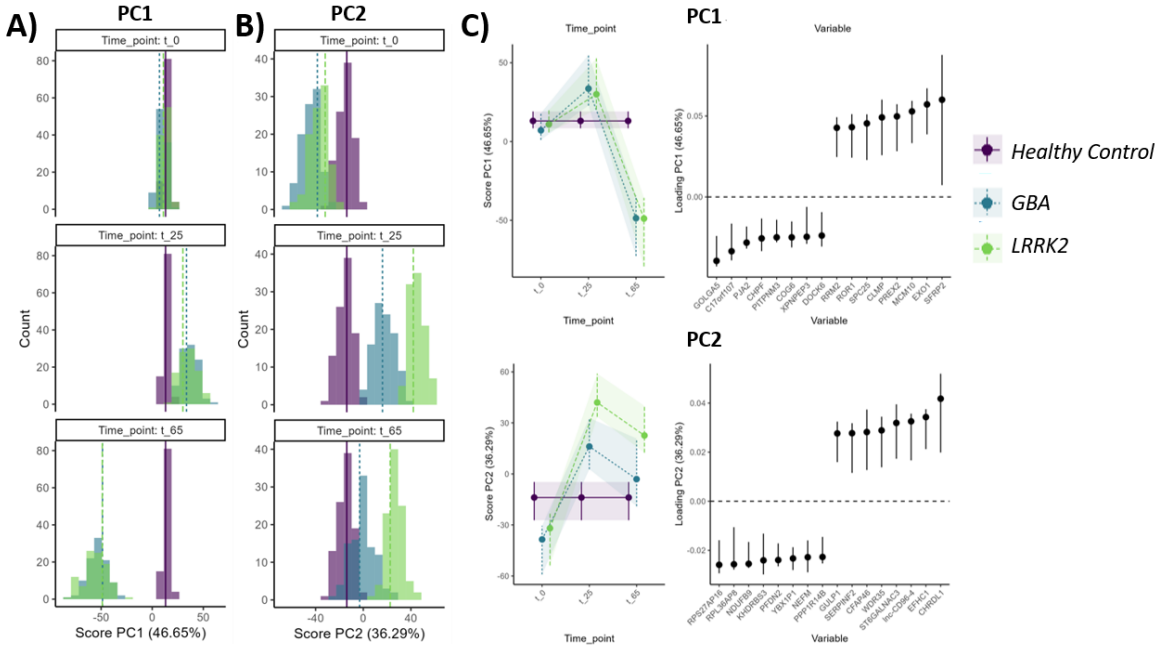


Figure 19: Scores and loadings from the PCA performed on time:genotype interaction term effect matrix form ALASCA. General scores trends with histograms (A-B) and a scatterplot (C). Genes with highest loadings are also shown (C)

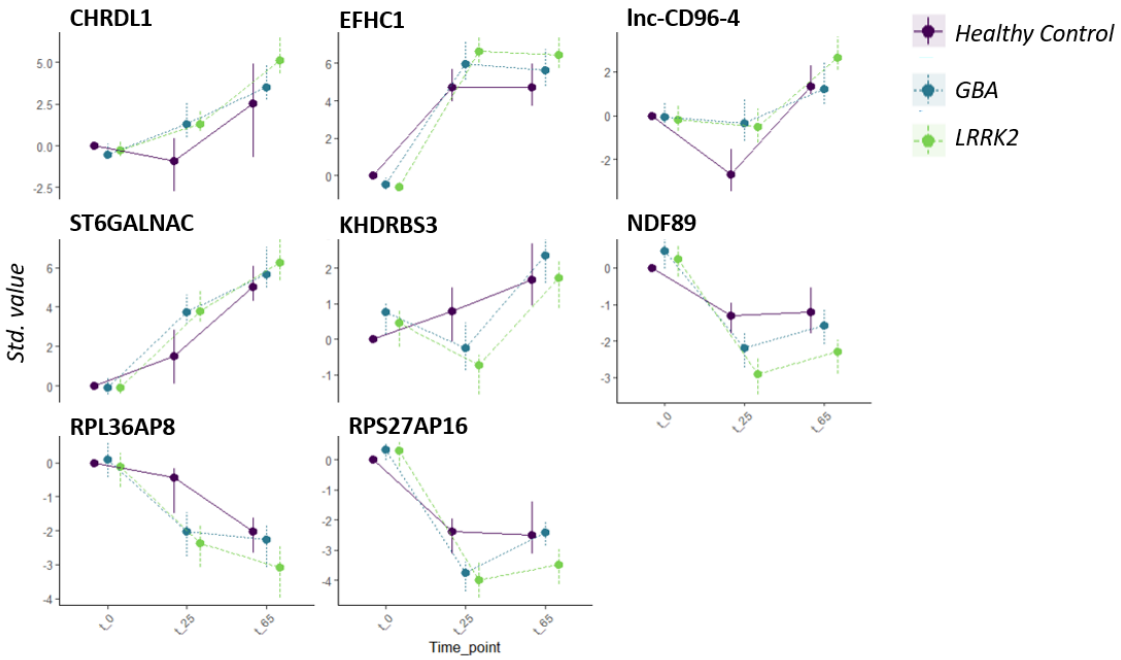


Figure 20: Marginal means from underlying regression models. Note that positive loadings increase when scores increase, negative loadins decrease when scores increase

To grasp to which extent those genes differ between conditions, PCA is performed on the counts subsetted only for the genes with highest loading (absolute values) for both PC1 and PC2 (genes shown on the right side of Figure 19 C). In Figure 21, it is possible to see how the temporal effect is less pronounced in the second two plots (Figure 21 C and D) especially for days 25 and 65. Those plots confirm the observations drawn before. in fact by looking at the day 65 cloud of points in the plot generated with high loadings in PC1 (Figure 21 B) Healthy control samples tend to separate slightly from the others, for the one generated with high ladings from PC2 (Figure 21 D) Healthy control tend to cluster together more in the middle and the two conditions separate more. Naturally, the separation is not perfect, but we can safely state that the positioning of the samples is not completely random and there might be some interesting patterns worth exploring within those genes. As shown for the result of the linear models it is of interest to explore the expression values of those genes in the different sections and across time (Figure 22, 23).

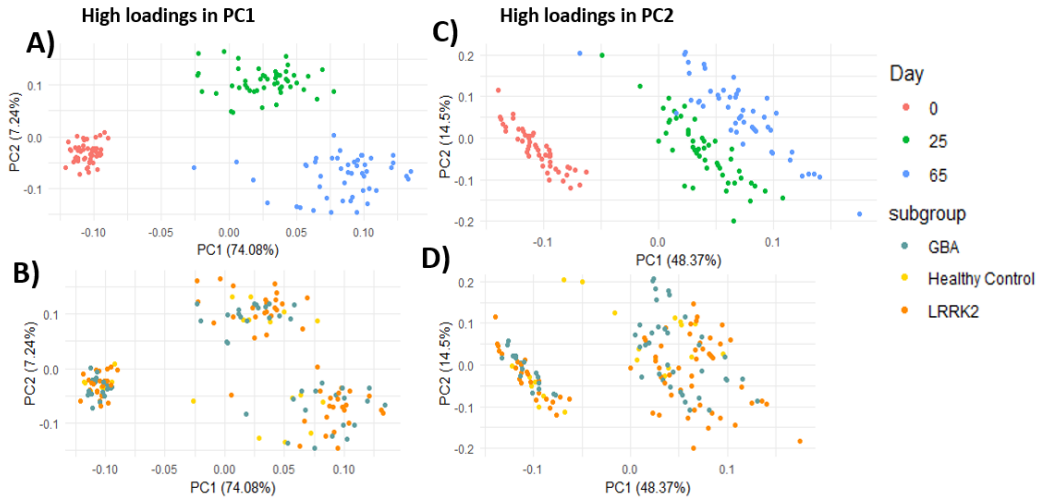


Figure 21: PCA analysis on top 10 higher loadings (absolute value) genes from alasca interaction effect PC1 (A and B) and PC2 (C and D) results. Colour coding shown for both time points(A and C) and subgroup(B and D)

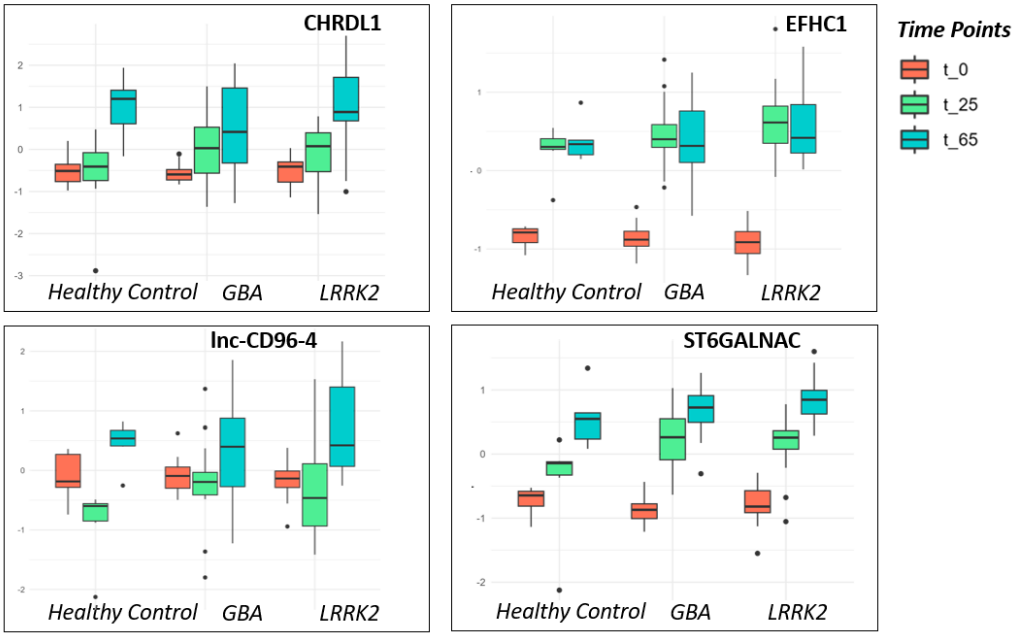


Figure 22: Box plot of genes with high positive loadings in the second pincipal component for the interaction effect matrix (time:genotype)

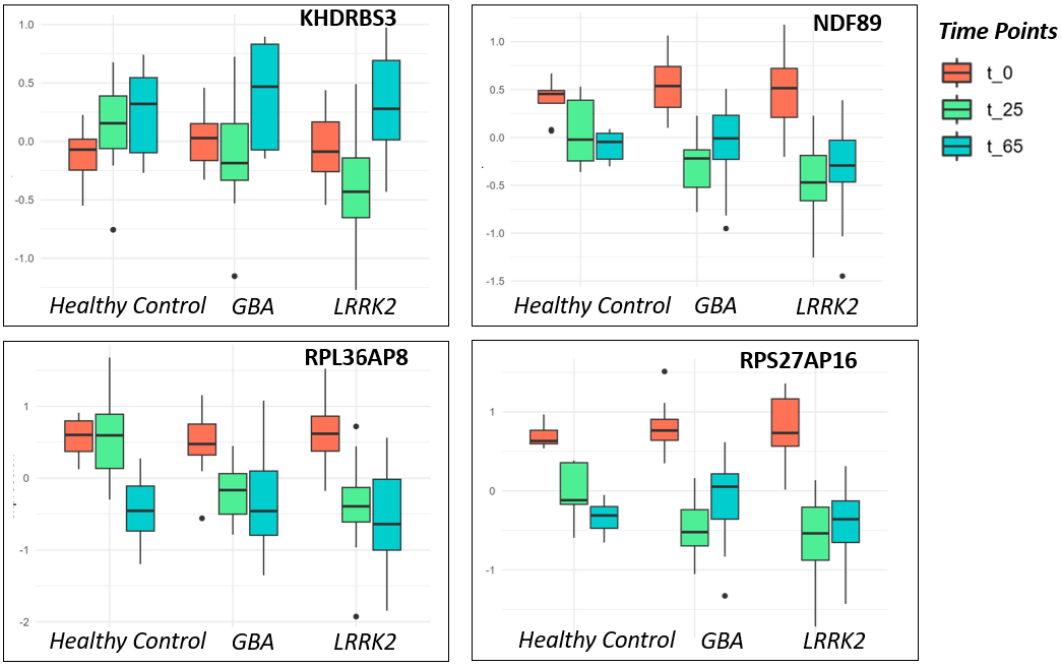


Figure 23: Box plot of genes with high negative loadings in the second pincipal component for the interaction effect matrix (time:genotype)

4 Discussion

4.1 Trajectory Inference

Trajectory inference methods are only partially used in this thesis as pseudotemporal ordering is not the ultimate goal of this method. The main point when using TI methods is to establish genes that are relevant in explaining gene expression changes outlining the inferred trajectory. Most of the packages have a function to establish gene importance or gene significance (depending on the algorithmic approach used for this step of the analysis). Some methods may be able to differentiate different lineages for more complex trajectories [55], when there is bifurcation, for example. To compare lineages there needs to be a massive change in expression at some point along the trajectory, as it is seen when stem cells differentiate into different cell types. However,this clearly does not reflect the situation investigated in this thesis.. In fact none of these TI methods is able to compare multiple trajectories inferred from the same basic developmental process in different conditions (disease vs healthy, or different mutations, for instance).

In addition to this, pseudo-time is probably not the most relevant method to study this dataset due to the limited number of time points and the sharply distinct clustering of those in PCA space. It is therefore not a surprise to see that pseudo-time values reflect perfectly the time points of three main clusters (Figure 11). It is still interesting to see that results in linear modeling are not completely identicalv when considering time or pseudotime, this will be further discussed in the next session.

A method to perform a Trajectory-based differential expression analysis within such conditions, as different mutations leading to the same disease, would be an interesting analytical tool, yet very challenging as it is not possible to rely on a massive divergence.

4.2 Mixed Effects Models

The main bottleneck for relying solely on MLMs is that the dataset does not provide enough statistical power to assess significance. Therefore it is not possible to draw strong conclusions but only make some considerations on the genes that are thought to be differentially expressed in the various conditions. Generally from these results, it is not possible to say whether LRRK2 and GBA samples follow distinct developmental trajectories but some genes seem to behave differently across conditions, unfortunately, that can still only be hypothesized due to the lack of statistical power.

One interesting observation was that a consistent number of differentially genes are long non-coding RNAs (lnc-RNAs). Those are a crucial component in gene expression processes due to their interactions with microRNAs. lnc-RNAs are fine regulators of expression and can affect chromatin remodeling and have the ability to upregulate translation without altering

mRNA levels[37]. This might explain partially why it is not so trivial to spot differences in expression levels for most of the protein coding genes.

Among those lnc-RNF40-1 (Figure 15) seems to be lower in expression for both LRRk2 and GBA at day 25 and particularly downregulated in LRRK2 at day 65. Not much is known about this specific noncoding. It is noteworthy to consider that this nocoding is related to the RNF40 gene. This gene can create a stable heterodimer complex that can monoubiquitylate histone H2B at lysine 120 and various nonhistone proteins [20]. aadditionally H2B ubiquitin-protein ligase RNF40 is needed for somatic cell reprogramming [58].

other non coding entities are found to be relevant in the LRRK2 condition, fore instance ERP29P1 seems to drop down in expression in the first 25 days and stabilize at the same level at day 65 in the control samples. for the other two conditions, the expression increases. Interestingly the pattern for LRRK2 seems to be particularly different as it is lower in expression at day 0 but it increases and seems to be upregulated at day 65. This is a pseudogene associated with Endoplasmic reticulum protein 29 (ERP29), a protein that might have a potential role in neuroprotection in retinal and neurodegenerative diseases [40]. Additionally, ERP29 is involved in the production and transport of several transmembrane and secretory proteins [7]. This might be interesting considering that, as stated in the introduction, proteins involved in vesicle trafficking are commonly phosphorylated by dardarin[36] (protein encoded by LRRK2).

Conversely COL28A1 (Collagen Type XXVIII Alpha 1 Chain), is a protein-coding gene that seems to be downregulated at day 0. This gene seem higher in expression both in GBA and LRRK2 with respect to control.Gene Ontology (GO) annotations related to this gene include serine-type endopeptidase inhibitor activity [42]. Difference in expression for ths gene was mainly seen in day 0, so iPSCs, and not in later time points.
Particularly interesting is the boxplot showing the expression patterns of LINC00115. This lnc-RNA shows a big variability in the two conditions. Surprisingly, LRRK2 seems to be upregulated in the first two time points and to be declining in expression at day 65, whereas GBA also shows a continuous rise throughout time and the healthy control group shows an increase at day 25 and stays stable at the same level at day 65. LINC00115 appears to be a critical modulator in GSC (glioma stem cells) self-renewal and tumorigenicity by interacting with miR-200s [52]. Tumor suppressive miR-200s, have an inverse relationship with the onset of Epithelial Mesenchymal-Transition (EMT) [2].

On the other hand, three are the main genes found by looking at GBA p-values. At day 65 ZMZ1-AS1 and SYNCP seem to be more expressed in GBA while NANOS1 expression is lower except for two outliers. ZMZ1-AS1 is another lnc-RNA. The respective ZMZ1 gene is a brain development regulator linked to intellectual impairment and autism [30]. This non

coding seems to control Prox1 [46], which during adult neuronal destiny determination is at the junction of several pathways [51]. Similarly, NANOS1 is found to control hippocampal synaptogenesis [39]. SYCP2 is linked to the loss of both H3K4me3 and H3K27me3 or the resolution of H3K27me3 in neural stem cells in the mesoderm, endoderm, and non-neural ectoderm [9].

Lastly FRA10AC1 and lnc-MAP3K13-3 are differentially expressed between GBA ans LRRK2. What is also surprising is that those genes, to some extent, seem to follow a similar path as the one observed in LINC00115. There is no extensive literature on lnc-MAP3K13-3, another variant of this noncoding is lnc-MAP3K13-7 is thought to inhibit ovarian granulosa cell Proliferation [22]. FRA10AC1 was found to be correlated with PAK4, a gene associated with DNA repair genes in ovarian cancer patients [33]. The link between PD and Ovarian cancer is still not clear, yer there is some evidence for some link between the two [35] [24]. Furthermore, FRA10AC1 is also associated with neurodevelopmental disorders [56].

4.3 ASCA

ASCA represent an alternative analysis to study this dataset to explore developmental trajectories in different conditions. The tool is developed to perform more in general multivariate analysis and is not specifically designed for time course analysis in biological settings, however it provides more insight with respect to trajectory inferences. Firstly there is a clear pattern emerging from Figure 19 that gives some evidences to start reasoning around the first research question : 'Do the iPSCs cells that are affected by different PD mutations (LRRK2, GBA) follow distinct developmental trajectories upon neuronal differentiation ?'.

Relevant genes would be those having high (both positive and negative) loadings that, in the case of the second PC would be those that differ in their trajectory not just from the reference but also between the two conditions. It is anyway not clear what would be a good threshold or computational approach to filter genes that are truly relevant. Simply choosing the top number 'n' of genes by taking the absolute value of the loadings can lead to wrong conclusions as it might include genes that are not relevant or, loose information leaving out relevant genes. To some extent that is the case also when a hard threshold approach is applied but in that scenario, at least, the criteria is not a fixed number of genes but rather a value, as for hypothesis testing, to assess significance. An overall observation is that by looking at the boxplots genes found to have high (positive or negative) loadings appear to have a higher number of outliers suggesting a possible methodological limitation. It is important to take into account also that the bootstrapping was performed for only 100 iterations that might not be sufficient for truly robust results. Additionally, if the model is run with the three conditions as in this case the main problem is that it is not possble to methodologically categorize genes as specific to one or the other condition.

Anyway to explore ALASCA results, genes retrieved from selecting the highest loadings shown in Figure 19 are analyzed (Figure 27). Those represent 4 genes with highest positive loadings (CHRDL1, EFHC1, lnc-CD96-4, ST6GALNAC) (Figure 22)and 4 genes with lowest positive loadings (KHDRBS3, NDF89, RPL36AP8, RPS27AP16)(Figure 23) .

The first set of four of genes (higher positive loadings) show an increase in expression over time. CHRDL1 is a structural glycoprotein. This gene appears to be downregulated in GBA at day 65. Through the release of Chrdl1, astrocytes promote GluA2-dependent synapse maturation and thereby limit synaptic plasticity [6]. EFHC1 expression is tight around the same value in healthy control samples and has a wider range of values in both GBA and LRRK2. EFHC1 is a microtubule-associated protein that has a role in the regulation of cell division and is involved in cortical development [12].

lnc-CD96-4 similarly to CHRDL1 has a wider range of expression levels in the two conditions at day 65. The main molecular function of lnc-CD96-4 is not well known, except for a suggested interaction with CD96 that regulates NK cell effector function and metastasis [41]. ST6GALNAC is instead upregulated on day 65 in both conditions. Interestingly it appears that ST6GALNAC5 expression encourage the development of brain metastases. [17].

The other four genes generally showed a decrease in expression over time. KHDRBS3 represents an exception to this as the pattern is different across conditions. While always increasing in healthy controls this its expression decreased at day 25 for the two conditions and then increase again. LRRK2 samples seem to be downregulated with respect to GBA at day 25. Similarly to most of the noncoding found also this gene is linked to regulation as is found to control splicing patterns of Neurexin pre-mRNAs in the Brain [19]. Despite being encoded by only three genes, Neurexin is among the most varied protein types in the body (Nrxn1-3) due to the presence of five alternatively spliced regions [19]. Those genes are relevant for the pre-synaptic terminal formation [13]. NDF89 seems also lower in expression in LRRK2, but in a later stage, at day 65 and shows a different path from healthy control to the two conditions as in the first case it stabilizes at the same level from day 25 to 65 while it increase in the two conditions. In recent studies NDUFB9 was found to be part of a robust gene coexpression network associated with cocaine use disorder and involved in neurotransmission (GABA, acetylcholine, serotonin, and dopamine) and drug addiction [26] [31].

RPL36AP8 and RPS27AP16 are both slightly lower in expression in LRRK2 and are both pseudogenes the second one has a role in ubiquitination, it has been detected in tissues alongside Ub RNAs [18].

4.4 Future Prospective

Once acknowledged the current limitations discussed in the section above, also linked to the fact that this method was not explored extensively throughout the overall period of the internship but only toward the end, ASCA modeling results are still promising. It is critical to emphasize again how crucial it is to understand how to methodologically select genes are more relevant using ALASCA. In fact this thesis provides only an introductory screening of the 8 genes that were found to be more relevant by filtering on scores loadings. Future work should look into a more statistically driven solid method for selection of relevant genes. To address the characterization of the single mutations it might be interesting to look at pairwise comparison of the different conditions. One possibility is to compare the outcomes ASCA models taking each condition separately with Healthy control as well as explore how LRRK2 developmental processes differs from GBA independently and vice versa by removing the effect of time on each donition in separate runs. Additionally, it would be interesting to explore the expression of the genes found in the different conditions for the single-cell RNA sequencing dataset as it is available within the FOUNDIN-PD project.

5 Conclusion

To wrap up, this thesis suggests some possibilities to perform Trajectory-based differential expression analysis exploring genetic subtypes of PD throughout differentiation to dopaminergic neurons. Two different genetic mutations observed in Parkinson's disease are investigated using data obtained from human induced pluripotent cells (iPSCs) during development into neurons. The main goal is to study how those mutations affect the developmental process and to sketch differences. To do so two main approaches are suggested: Mixed effect linear modeling (MLM) combined with pseudo-temporal ordering to better characterize the samples on a developmental trajectory, and ANOVA-Simultaneous Component Analysis (ASCA). In the first case, the dataset does not provide enough statistical power to provide any strong conclusion. However, some genes are still found to be differentially expressed across time and within conditions. Most of the genes found have a regulatory role and are involved in different biological processes including ubiquitination, protein transport, and as well as a developmental-related process including synaptogenesis and adult neuronal destiny determination. Given the different contrasts explored in linear models, those genes might help in better characterizing the given mutations. Outcomes from ASCA seem to suggest the presence of some overall systematic differences between the three genotypes. The best strategy to select relevant genes for this method more accurately and systematically is still up for debate and further investigation. The few genes investigated from the results for this model are also linked to synapse maturation, cortical development, and different kinds of regulatory processes including pre-mRNA splicing and ubiquitination. At this stage it is useful to recall the main research questions this thesis is trying to address :

1. Do the iPSCs cells that are affected by different PD mutations (LRRK2, GBA) follow distinct developmental trajectories?
2. Is it possible to spot genes to characterize those different mutations?
3. Is continuous temporal ordering a valuable addition to linear modeling?
4. Is ASCA modeling a possible solution to explore the effect of the interaction of time and genotype?

1) The main purpose of using Trajectory inference methods was to be able to answer this first research question. Pseudotemporal ordering was supposed to provide a reference space on which these conditions are compared. Naturally, as often in research, things are not that easy. As discussed in the previous section TI is not sensible enough to compare conditions. ALASCA on the other side appears to be more suited for getting an overview of temporal patterns. If we analyze Figures 19 and 20 there is a pattern emerging that might indicate differences across conditions along the developmental trajectories.

2) Once again, from the results obtained in the linear models in the first section of the thesis it is not possible to provide a definitive answer given the lack of statistical power. It would be reasonable to assume that with more data available this approach might lead to more accurate results. ALASCA seem also to provide some differentially expressed genes along the trajectories that can be specific to one of the two mutations. Naturally, to fully answer, those genes would need to be validated and a proper statistically driven solid method for selection of relevant genes needs to be used. Nevertheless, there is potential for interesting findings down this path.

3) This point was partly addressed already in the discussion, and the short answer is: not in this case. Pseudo-temporal ordering would be an interesting addition in a situation where the transition from one point to another is more subtle and the clustering less obvious. it is in this case not a surprise that pseudo-times values and time points are completely overlapping in Figure 11 .

4) ALASCA is an interesting tool to analyze multivariate data of various kinds. Despite not being designed specifically for temporal analysis, unlike trajectory inference, it seems to be quite extensively used for those cases. In the paper introducing ALASCA, there are some examples of such usage. In this case, it appears to correctly isolate the massive temporal effect in Figure 18 and mine the hidden pattern that reflects the interaction of time and genotype in Figure 19. Despite some more investigation being required the results obtained so far are encouraging.

REFERENCES

[1] Elissavet Akrioti, Timokratis Karamitros, Panagiotis Gkaravelas, Georgia Kouroupi, Rebecca Matsas, and Era Taoufik. Early Signs of Molecular Defects in iPSC-Derived Neural Stems Cells from Patients with Familial Parkinson's Disease. *Biomolecules*, 12(7):876, June 2022.

[2] Ghader Babaei, Negin Raei, Attabak Toofani Milani, Shiva Gholizadeh-Ghaleh Aziz, Nima Pourjabbar, and Faezeh Geravand. The emerging role of mir-200 family in metastasis: focus on emt, cscs, angiogenesis, and anoikis. *Molecular Biology Reports*, pages 1–13, 2021.

[3] Douglas M Bates. lme4: Mixed-effects modeling with r, 2010.

[4] Yoav Benjamini and Henry Braun. John w. tukey's contributions to multiple comparisons. *Annals of Statistics*, pages 1576–1594, 2002.

[5] Carlo Bertinetto, Jasper Engel, and Jeroen Jansen. Anova simultaneous component analysis: A tutorial review. *Analytica Chimica Acta: X*, 6:100061, 2020.

[6] Elena Blanco-Suarez, Tong-Fei Liu, Alex Kopelevich, and Nicola J Allen. Astrocyte-secreted chordin-like 1 drives synapse maturation and limits plasticity by increasing synaptic glua2 ampa receptors. *Neuron*, 100(5):1116–1132, 2018.

[7] Margaret Brecker, Svetlana Khakhina, Tyler J Schubert, Zachary Thompson, and Ronald C Rubenstein. The probable, possible, and novel functions of erp29. *Frontiers in Physiology*, 11:574339, 2020.

[8] Elisangela Bressan, Xylena Reed, Vikas Bansal, Elizabeth Hutchins, Melanie M. Cobb, Michelle G. Webb, Eric Alsop, Francis P. Grenn, Anastasia Illarionova, Natalia Savytska, Ivo Violich, Stefanie Broeer, Noémia Fernandes, Ramiyapriya Sivakumar, Alexandra Beilina, Kimberley J. Billingsley, Joos Berghausen, Caroline B. Pantazis, Vanessa Pitz, Dhairya Patel, Kensuke Daida, Bessie Meechoovet, Rebecca Reiman, Amanda Courtright-Lim, Amber Logemann, Jerry Antone, Mariya Barch, Robert Kitchen, Yan Li, Clifton L. Dalgard, Patrizia Rizzu, Dena G. Hernandez, Brooke E. Hjelm, Mike Nalls, J. Raphael Gibbs, Steven Finkbeiner, Mark R. Cookson, Kendall Van Keuren-Jensen, David W. Craig, Andrew B. Singleton, Peter Heutink, and Cornelis Blauwendraat. The Foundational Data Initiative for Parkinson Disease: Enabling efficient translation from genetic maps to mechanism. *Cell Genomics*, 3(3):100261, March 2023.

[9] Matthew J Burney, Caroline Johnston, Kee-Yew Wong, Siaw-Wei Teng, Vassilios Beglopoulos, Lawrence W Stanton, Brenda P Williams, Angela Bithell, and Noel J Buckley. An epigenetic signature of developmental potential in neural stem cells and early neurons. *Stem cells*, 31(9):1868–1880, 2013.

[10] Pierre R Bushel, Stephen S Ferguson, Sreenivasa C Ramaiahgari, Richard S Paules, and Scott S Auerbach. Comparison of normalization methods for analysis of tempo-seq targeted rna sequencing data. *Frontiers in Genetics*, 11:594, 2020.

[11] Avital Cnaan, Nan M Laird, and Peter Slasor. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in medicine*, 16(20):2349–2380, 1997.

[12] Laurence De Nijs, Christine Léon, Laurent Nguyen, Joseph J LoTurco, Antonio V Delgado-Escueta, Thierry Grisar, and Bernard Lakaye. Efhc1 interacts with microtubules to regulate cell division and cortical development. *Nature neuroscience*, 12(10):1266–1274, 2009.

[13] Camin Dean, Francisco G Scholl, Jenny Choih, Shannon DeMaria, James Berger, Ehud Isacoff, and Peter Scheiffele. Neurexin mediates the assembly of presynaptic terminals. *Nature neuroscience*, 6(7):708–716, 2003.

[14] George DeMaagd and Ashok Philip. Parkinson's Disease and Its Management: Part 1: Disease Entity, Risk Factors, Pathophysiology, Clinical Presentation, and Diagnosis. *P & T: A Peer-Reviewed Journal for Formulary Management*, 40(8):504–532, August 2015.

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

[16] E. R. Dorsey, R. Constantinescu, J. P. Thompson, K. M. Biglan, R. G. Holloway, K. Kieburtz, F. J. Marshall, B. M. Ravina, G. Schifitto, A. Siderowf, and C. M. Tanner. Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. *Neurology*, 68(5):384–386, January 2007.

[17] Aurore Drolez, Elodie Vandenhaute, Clément Philippe Delannoy, Justine Hélène Dewald, Fabien Gosselet, Romeo Cecchelli, Sylvain Julien, Marie-Pierre Dehouck, Philippe Delannoy, and Caroline Mysiorek. St6galnac5 expression decreases the interactions between breast cancer cells and the human blood-brain barrier. *International journal of molecular sciences*, 17(8):1309, 2016.

[18] Marie-Line Dubois, Anna Meller, Sondos Samandi, Mylène Brunelle, Julie Frion, Marie A Brunet, Amanda Toupin, Maxime C Beaudoin, Jean-François Jacques, Dominique Lévesque, et al. Ubb pseudogene 4 encodes functional ubiquitin variants. *Nature communications*, 11(1):1306, 2020.

[19] Ingrid Ehrmann, Caroline Dalgliesh, Yilei Liu, Marina Danilenko, Moira Crosier, Lynn Overman, Helen M Arthur, Susan Lindsay, Gavin J Clowry, Julian P Venables, et al.

The tissue-specific rna binding protein t-star controls regional splicing patterns of neurexin pre-mrnas in the brain. *PLoS genetics*, 9(4):e1003474, 2013.

[20] Junjiang Fu, Li Liao, Kyathegowdanadoddi Srinivasa Balaji, Chunli Wei, Jaehoon Kim, and Jiangzhou Peng. Epigenetic modification and a role for the e3 ligase rnf40 in cancer development and metastasis. *Oncogene*, 40(3):465–474, 2021.

[21] Joseph C Gardiner, Zhehui Luo, and Lee Anne Roman. Fixed effects, random effects and gee: What are the differences? *Statistics in medicine*, 28(2):221–239, 2009.

[22] Xueying Geng, Jun Zhao, Jiayu Huang, Shang Li, Weiwei Chu, Wang-sheng Wang, Zi-Jiang Chen, and Yanzhi Du. lnc-map3k13-7: 1 inhibits ovarian gc proliferation in pcos via dnmt1 downregulation-mediated cdkn1a promoter hypomethylation. *Molecular Therapy*, 29(3):1279–1293, 2021.

[23] Julia C. Greenland, Caroline H. Williams [U+2010] Gray, and Roger A. Barker. The clinical heterogeneity of Parkinson's disease and its therapeutic implications. *European Journal of Neuroscience*, 49(3):328–338, February 2019.

[24] Jian-Zeng Guo, Qian Xiao, Lang Wu, Fa Chen, Jia-Li Yin, Xue Qin, Ting-Ting Gong, and Qi-Jun Wu. Ovarian cancer and parkinson's disease: A bidirectional mendelian randomization study. *Journal of Clinical Medicine*, 12(8):2961, 2023.

[25] Torsten Hothorn, Frank Bretz, Peter Westfall, Richard M Heiberger, Andre Schuetzenmeister, Susan Scheibe, and Maintainer Torsten Hothorn. Package 'multcomp'. *Simultaneous inference in general parametric models. Project for Statistical Computing, Vienna, Austria*, pages 1–36, 2016.

[26] Spencer B Huggett and Michael C Stallings. Genetic architecture and molecular neuropathology of human cocaine addiction. *Journal of Neuroscience*, 40(27):5300–5313, 2020.

[27] Benjamin H.M. Hunn, Stephanie J. Cragg, J. Paul Bolam, Maria-Grazia Spillantini, and Richard Wade-Martins. Impaired intracellular trafficking defines early Parkinson's disease. *Trends in Neurosciences*, 38(3):178–188, March 2015.

[28] Jeroen J Jansen, Huub CJ Hoefsloot, Jan van der Greef, Marieke E Timmerman, Johan A Westerhuis, and Age K Smilde. Asca: analysis of multivariate data obtained from an experimental design. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 19(9):469–481, 2005.

[29] Anders Hagen Jarmund, Torfinn Støve Madssen, and Guro F Giskeødegård. Alasca: an r package for longitudinal and cross-sectional analysis of multivariate data by asca-based methods. *Frontiers in Molecular Biosciences*, 9:962431, 2022.

[30] Rajan KC, Alina S Tiemroth, Abbigail N Thurmon, Stryder M Meadows, and Maria J Galazo. Zmiz1 is a novel regulator of brain development associated with autism and intellectual disability. *Frontiers in Psychiatry*, 15:1375492, 2024.

[31] Arshad H Khan, Jared R Bagley, Nathan LaPierre, Carlos Gonzalez-Figueroa, Tadeo C Spencer, Mudra Choudhury, Xinshu Xiao, Eleazar Eskin, James D Jentsch, and Desmond J Smith. Genetic pathways regulating the longitudinal acquisition of cocaine self-administration in a panel of inbred and recombinant inbred mice. *Cell reports*, 42(8), 2023.

[32] Raivo Kolde and Maintainer Raivo Kolde. Package 'pheatmap'. *R package*, 1(7):790, 2015.

[33] Kei Kudo, Yoshimi Endo Greer, Teruhiko Yoshida, Brittney S Harrington, Soumya Korrapati, Yusuke Shibuya, Leah Henegar, Jeffrey B Kopp, Takeo Fujii, Stanley Lipkowitz, et al. Dual-inhibition of nampt and pak4 induces anti-tumor effects in 3d-spheroids model of platinum-resistant ovarian cancer. *Cancer Gene Therapy*, pages 1–15, 2024.

[34] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161, 2007.

[35] Yong Qi Leong, Shaun Wen Huey Lee, and Khuen Yen Ng. Cancer risk in parkinson disease: An updated systematic review and meta-analysis. *European Journal of Neurology*, 28(12):4219–4237, 2021.

[36] Jie-Qiong Li, Lan Tan, and Jin-Tai Yu. The role of the LRRK2 gene in Parkinsonism. *Molecular Neurodegeneration*, 9(1):47, December 2014.

[37] Elena López-Jiménez and Eduardo Andrés-León. The implications of ncrnas in the development of human diseases. *Non-coding RNA*, 7(1):17, 2021.

[38] Torfinn S Madssen, Guro F Giskeødegård, Age K Smilde, and Johan A Westerhuis. Repeated measures asca+ for analysis of longitudinal intervention studies with multivariate outcome data. *PLoS Computational Biology*, 17(11):e1009585, 2021.

[39] Darío Maschi, Ana J Fernández-Alvarez, and Graciela Lidia Boccaccio. The rna-binding protein nanos1 controls hippocampal synaptogenesis. *PloS one*, 18(4):e0284589, 2023.

[40] Todd McLaughlin, Marek Falkowski, Joshua J Wang, and Sarah X Zhang. Molecular chaperone erp29: a potential target for cellular protection in retinal and neurodegenerative diseases. *Retinal Degenerative Diseases: Mechanisms and Experimental Therapy*, pages 421–427, 2018.

[41] Deepak Mittal, Ailin Lepletier, Jason Madore, Amelia Roman Aguilera, Kimberley Stannard, Stephen J Blake, Vicki LJ Whitehall, Cheng Liu, Mark L Bettington,

Kazuyoshi Takeda, et al. Cd96 is an immune checkpoint that regulates cd8+ t-cell antitumor function. *Cancer immunology research*, 7(4):559–571, 2019.

[42] Ana M Moreira, Rui M Ferreira, Patrícia Carneiro, Joana Figueiredo, Hugo Osório, José Barbosa, John Preto, Perpétua Pinto-do Ó, Fátima Carneiro, and Raquel Seruca. Proteomic identification of a gastric tumor ecm signature associated with cancer progression. *Frontiers in Molecular Biosciences*, 9:818552, 2022.

[43] Manos Papadakis, Michail Tsagris, Marios Dimitriadis, Stefanos Fafalios, Ioannis Tsamardinos, Matteo Fasiolo, Giorgos Borboudakis, John Burkardt, C Zou, and K Lakiotaki. Package 'rfast', 2018.

[44] Rob Patro, Geet Duggal, and Carl Kingsford. Salmon: accurate, versatile and ultrafast quantification from rna-seq data using lightweight-alignment. *BioRxiv*, 10:021592, 2015.

[45] Alexia Polissidis, Lilian Petropoulou-Vathi, Modestos Nakos-Bimpos, and Hardy J Rideout. The future of targeted gene-based treatment strategies and biomarkers in parkinson's disease. *Biomolecules*, 10(6):912, 2020.

[46] KC Rajan, Nehal R Patel, Anoushka Shenoy, Joshua P Scallan, Mark Y Chiang, Maria J Galazo, and Stryder M Meadows. Zmiz1 is a novel regulator of lymphatic endothelial cell gene expression and function. *bioRxiv*, 2023.

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

[48] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5):547–554, May 2019.

[49] Eldad David Shulman and Ran Elkon. Genetic mapping of developmental trajectories for complex traits and diseases. *Computational and Structural Biotechnology Journal*, 19:3458–3469, 2021.

[50] Laura Smith and Anthony H. V. Schapira. Gba variants and parkinson disease: Mechanisms and treatments. *Cells*, 11(8):1261, April 2022.

[51] Athanasios Stergiopoulos, Maximilianos Elkouris, and Panagiotis K Politis. Prospero-related homeobox 1 (prox1) at the crossroads of diverse pathways during adult neural fate specification. *Frontiers in cellular neuroscience*, 8:454, 2015.

[52] Jianming Tang, Bo Yu, Yanxin Li, Weiwei Zhang, Angel A Alvarez, Bo Hu, Shi-Yuan Cheng, and Haizhong Feng. Tgf-β-activated lncrna linc00115 is a critical regulator of glioma stem-like cell tumorigenicity. *EMBO reports*, 20(12):e48170, 2019.

[53] Michel Thiel, Nadia Benaiche, Manon Martin, Sébastien Franceschini, Robin Van Oirbeek, and Bernadette Govaerts. limpca: An r package for the linear modeling of high-dimensional designed data based on asca/apca family of methods. *Journal of Chemometrics*, 37(7):e3482, 2023.

[54] Debby Tsuang, James B. Leverenz, Oscar L. Lopez, Ronald L. Hamilton, David A. Bennett, Julie A. Schneider, Aron S. Buchman, Eric B. Larson, Paul K. Crane, Jeffrey A. Kaye, Patricia Kramer, Randy Woltjer, Walter Kukull, Peter T. Nelson, Gregory A. Jicha, Janna H. Neltner, Doug Galasko, Eliezer Masliah, John Q. Trojanowski, Gerard D. Schellenberg, Dora Yearout, Haley Huston, Allison Fritts-Penniman, Ignacio F. Mata, Jia Y. Wan, Karen L. Edwards, Thomas J. Montine, and Cyrus P. Zabetian. GBA mutations increase risk for Lewy body disease with and without Alzheimer disease pathology. *Neurology*, 79(19):1944–1950, November 2012.

[55] Koen Van Den Berge, Hector Roux De Bézieux, Kelly Street, Wouter Saelens, Robrecht Cannoodt, Yvan Saeys, Sandrine Dudoit, and Lieven Clement. Trajectory-based differential expression analysis for single-cell sequencing data. *Nature Communications*, 11(1):1201, March 2020.

[56] Leonie von Elsner, Guoliang Chai, Pauline E Schneeberger, Frederike L Harms, Christian Casar, Minyue Qi, Malik Alawi, Ghada MH Abdel-Salam, Maha S Zaki, Florian Arndt, et al. Biallelic fra10ac1 variants cause a neurodevelopmental disorder with growth retardation. *Brain*, 145(4):1551–1563, 2022.

[57] Koichi Wakabayashi, Kunikazu Tanji, Fumiaki Mori, and Hitoshi Takahashi. The lewy body in parkinson's disease: Molecules implicated in the formation and degradation of [U+2010]synuclein aggregates. *Neuropathology*, 27(5):494–506, September 2007.

[58] Wanhua Xie, Michaela Miehe, Sandra Laufer, and Steven A Johnsen. The h2b ubiquitin-protein ligase rnf40 is required for somatic cell reprogramming. *Cell Death & Disease*, 11(4):287, 2020.

Appendix

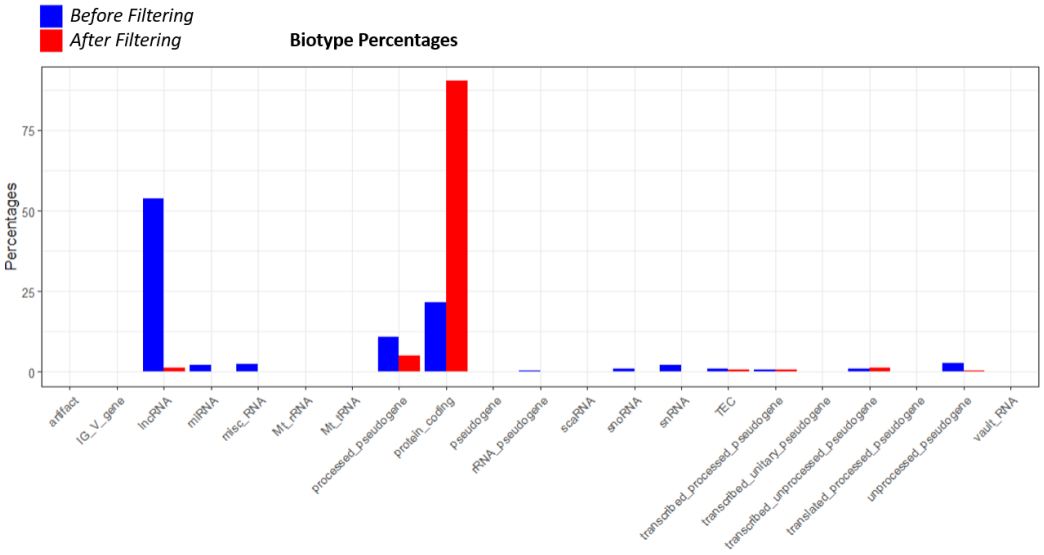


Figure 24: Percentages if different biotyped before and after filtering on the bulk RNA-seq dataset

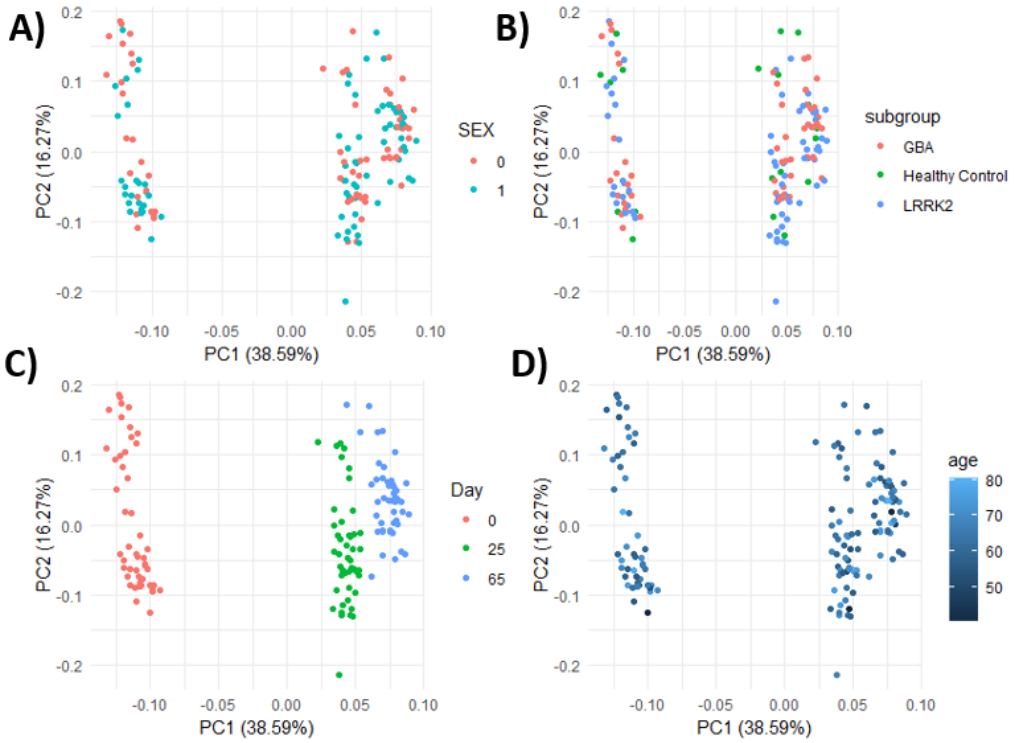


Figure 25: PCA performed on unprocessed RNA-seq data, colour coded for gender (A), phenotype (B), Time point (C) and Age (D). Note none of the variables associate / can explain the undesired variation observed along PC2

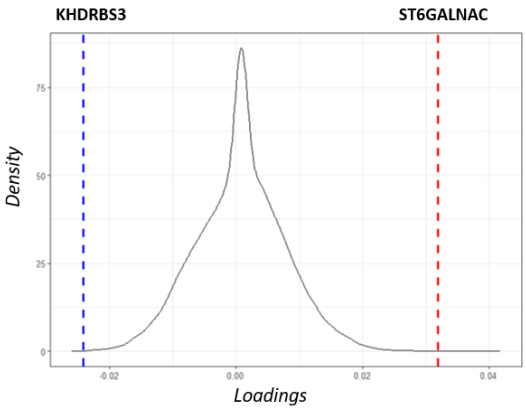
A)

Genes	pval_GBA_vs_Control	pval_LRRK2_vs_Control	pval_GBA_vs_LRRK2
LINC00115	0.8364219071	0.028614477	0.001343883
FRA10AC1	0.7147205662	0.997018236	0.001807181
ERP29P1	0.4242267179	0.009646909	0.009263927
lnc-MAP3K13-3	0.5164399595	0.803028411	0.009265626
ZMIZ1-AS1	0.0005234326	0.041301632	0.013057153
COL28A1	0.3102205033	0.004876314	0.025485123
SYCP2	0.0018958891	0.206984667	0.034097946
lnc-RNF40-1	0.9315955815	0.037638866	0.034525677

B)

Genes	pval_GBA_vs_Control	pval_LRRK2_vs_Control	pval_GBA_vs_LRRK2
FRA10AC1	0.545840545	0.963460712	0.0001806829
lnc-MAP3K13-3	0.434409165	0.670579731	0.0010842592
LINC00115	0.774577469	0.027987072	0.0012831912
ERP29P1	0.254799763	0.003825135	0.0104543672
NANOS1	0.035780034	0.968791053	0.0198996510
COL28A1	0.380668157	0.009105526	0.0212310065
ZMIZ1-AS1	0.001399851	0.089660931	0.0296026884

Figure 26: Table of p-values for each contrast in mixed effect linear models used with (A) Trajectory Inference and (B) Time Points



covars	PC	loading	low	high
RPL36AP8	2	-0.02568679	-0.02785181	-0.01057141
RPS27AP16	2	-0.02590548	-0.02935832	-0.01588524
NDUFB9	2	-0.02549318	-0.02717752	-0.01658859
KHDRBS3	2	-0.02408386	-0.02982072	-0.01318603
EFHC1	2	0.03431204	0.02126353	0.03753668
CHRDL1	2	0.04181141	0.01988227	0.05198491
lnc-CD96-4	2	0.03257920	0.01668011	0.03570000
ST6GALNAC3	2	0.03192799	0.01735168	0.03951440

Figure 27: Density plots of loadings from the 2000 genes with highest loadings. On the two dashed lines are the lowest loadings (in absolute value) genes among the 8 with the highest loadings selected by the ASCA model