2016

Structured clustering representations and methods

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

AND

COLLEGE OF ENGINEERING

Dissertation

STRUCTURED CLUSTERING REPRESENTATIONS AND METHODS

by

ADRIAN HEILBUT

B.Sc.(Hon), University of Toronto, 2001

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2016

Approved by

First Reader

Eric Kolaczyk, PhD
Professor of Mathematics and Statistics
Boston University

Second Reader

Myriam Heiman, PhD
Assistant Professor of Neuroscience
Massachusetts Institute of Technology and Broad Institute

Acknowledgments

Thanks to Dr. Eric Kolaczyk and Dr. Myriam Heiman for all of their advice, wisdom, and patience.

Thanks to Dr. Joseph Lehár for his mentorship and advice over many years. I'm also very grateful to Dr. Richard H. Myers and Dr. Luis Carvalho who graciously joined my committee at the bottom of the 11th inning.

Thanks to Dr. Robert Fenster for collaborating and sharing all of his data and expertise on the Huntington project.

Thanks to Dr. Scott Mohr and Dr. Tom Tullius of the Bioinformatics Program for advice and guidance through some tricky situations. I would also like to thank Dave King, Caroline Lyman, and Johanna Vasquez for helping me to deal with the BU bureaucracy.

Thanks to the BU Departments of Computer Science, Biology, and Math for the many opportunities to teach, and to Dr. Bennett Goldberg for the chance to participate in the NSF GK-12 program.

Thanks to Dr. Gary Benson and Dr. Yu (Brandon) Xia for their valuable courses, and to Dr. Pablo Tamayo for the interesting discussions and statistical advice.

Thanks to Dr. Michael Talkowski and Dr. James Gusella for the opportunity to spend some time working on human genetics in the CHGR at MGH.

From my past scientific lives, I also want to thank Dr. Gregg Morin, Dr. Mike Tyers, Dr. Yuen Ho, Dr. Daniel Figeys, and Dr. Mike Moran for their invaluable mentorship and support along the way. Thank you especially to Dr. Christopher Hogue for my first opportunities to do research in bioinformatics and the excellent training. I'm also grateful to the Computer Science and Biochemistry departments at UofT for all that they taught me.

Finally, thanks to my parents, Dr. Michele Heilbut and Harold Heilbut for all their love and support.

STRUCTURED CLUSTERING REPRESENTATIONS AND METHODS

ADRIAN HEILBUT

Boston University, Graduate School of Arts and Sciences

and

College of Engineering, 2016

Major Professor: Eric Kolaczyk, Professor, Mathematics and Statistics

ABSTRACT

Rather than designing focused experiments to test individual hypotheses, scientists now commonly acquire measurements using massively parallel techniques, for *post hoc* interrogation. The resulting data is both high-dimensional and structured, in that observed variables are grouped and ordered into related subspaces, reflecting both natural physical organization and factorial experimental designs. Such structure encodes critical constraints and clues to interpretation, but typical unsupervised learning methods assume exchangeability and fail to account adequately for the structure of data in a flexible and interpretable way. In this thesis, I develop computational methods for exploratory analysis of structured high-dimensional data, and apply them to study gene expression regulation in Parkinson's (PD) and Huntington's diseases (HD).

BOMBASTIC (Block-Organized, Model-Based, Tree-Indexed Clustering) is a methodology to cluster and visualize data organized in pre-specified subspaces, by combining independent clusterings of blocks into hierarchies. BOMBASTIC provides a formal specification of the block-clustering problem and a modular implementation that facilitates integration, visualization, and comparison of diverse datasets and rapid exploration of alternative analyses.

These tools, along with standard methods, were applied to study gene expression in mouse models of neurodegenerative diseases, in collaboration with Dr. Myriam Heiman and Dr. Robert Fenster. In PD, I analyzed cell-type-specific expression following levodopa treatment to study mechanisms underlying levodopa-induced dyskinesia (LID). I identified likely regulators of the transcriptional changes leading to LID and implicated signaling

pathways amenable to pharmacological modulation (Heiman, Heilbut *et al*, 2014). In HD, I analyzed multiple mouse models (Kuhn, 2007), cell-type specific profiles of medium spiny neurons (Fenster, 2011), and an RNA-Seq dataset profiling multiple tissue types over time and across an mHTT allelic series (CHDI, 2015). I found evidence suggesting that altered activity of the PRC2 complex significantly contributes to the transcriptional dysregulation observed in striatal neurons in HD.

Contents

List of Tables

List of Figures

Chapter 1

Introduction: Analysis of Structured (Biological) Data

1.1 Data-intensive biology

Technological improvements in measurement, computation, and storage over the last two decades have changed the way that many scientific investigations are conducted [52]. Rather than designing focused, individual experiments to test individual hypotheses, it has become common to acquire high-dimensional measurements using massively parallel techniques and then interrogate these datasets after the fact. Doing science this way demands very different statistical tools and approaches to those developed for traditional focused experiments designed to test pre-specified hypotheses.

In cell and molecular biology, for example, one can profile the states of tens of thousands of molecular species over multiple cell types (and even individual cells), conditions, and time points. Extracting scientific value from such datasets requires statistical methods to assist in efficiently generating and ranking interpretations and hypotheses. To be useful, these methods ought to satisfy at least three demands. First, scientists need methods — and practical software tools — for visualization and exploratory data analysis. Large datasets need to be presented in ways that facilitate identification of interesting and relevant aspects using human pattern recognition and intuition, even when specific biological questions and statistical hypotheses are formulated imprecisely or not at all. Second, data generated in different experiments must be easily integrated. For example, one would hope to be able to use the results of expensive systematic profiling efforts, such as the ENCODE project which characterized genomic functional elements and regulators [129], or the Allen Brain Atlas survey of gene expression in the brain [126], to help interpret

results of smaller but more directed studies. Tools should allow scientists to focus efforts on consideration of scientific hypotheses, rather than expending time and resources implementing repetitive *ad hoc* analyses. Finally, since quantities of data to be analyzed now exceed human capacities, and questions are often not easily formulated as simple classification or regression problems, unsupervised methods will be essential to extract useful knowledge at scale.

Biological data often has important structure, which comes both from natural biological organization and from experimental designs. Organisms are built from compartmentalized systems and specialized cells, and molecular activities are dynamic and depend on context and stimuli. Dynamics and context dependence greatly complicates the study of physiology, development, and disease mechanisms. However, all this structure reflects evolutionary, ontogenic, and biophysical constraints. Biological systems have evolved and develop over cycles of duplication and specialization, so regulatory systems are reused and repurposed over different cell types and contexts. It is therefore critical to be able both to recognize common mechanisms and to identify those that are specialized or context-specific. The need to acquire measurements over multiple contexts and time increases experimental cost and complexity and introduces redundancy, but also affords opportunities to inform analyses with priors about continuity and causality.

Now that modern methods enable systematic measurements over many dimensions such as cell types, treatments, and time, analysis should make effective use of the resulting structure, and of the relationships to underlying experimental and biological structure that provide constraints and clues to interpretation. Specifically, we seek methods to:

- identify biologically relevant similarities and differences among cell types and contexts
- integrate prior knowledge – especially systematic molecular profiling experiments – with data from more focused experiments, over multiple contexts, perturbations, and time-points
- concisely and non-redundantly define, represent and identify bio-molecular state and state-spaces

- efficiently generate comprehensible, mechanistically grounded, and experimentally verifiable biological hypotheses

Unifying these requirements is a need to identify subsets of molecules with similar behaviours that have common causes or consequents. From a statistical and machine learning perspective, this is just clustering, but to be biologically relevant, clusters should relate to biochemical mechanisms and explanations. Even the large, high-resolution datasets now being generated remain insufficient for inference of causal structure directly from observations, especially in complex multicellular organisms. But by making better use of structured data acquired across experimental designs and multiple contexts, one can incorporate knowledge of biochemical mechanism and the embedded evolutionary history and constraints that inform how mechanisms are reused and refined, and perhaps at least attempt to automatically generate and rank possible mechanistic hypotheses to explain observations.

1.2 Clustering and structured data

Most existing unsupervised clustering [59] and supervised feature selection algorithms [45] presume that objects and/or features are exchangeable. We are interested in the case in which objects or variables are grouped in subspaces, ordered, or arranged in hierarchies. Such data arises naturally from experimental designs with multiple factors, where objects are measured over multiple experimental modalities or endowed with multiple kinds of prior information. For example, in a hypothetical biological experiment, depicted in Figure 1.1, expression of each gene might be quantified by RNA-seq over a panel of cell types, each treated with a dozen different drugs, with measurements taken at six time-points. Each gene may be annotated with genomic features, such as occurrence of cis-regulatory motifs in promoters and inferences about context- and location-dependent binding of regulatory proteins from chromatin immunoprecipitation experiments.

For such data, the analysis goals described above are not well satisfied by existing clustering methods. We offer some methods to cluster and model structured, high-dimensional data that seek to better exploit known structure and prior data, and relate clusters to potential causal explanations. We focus on the situation in which structure of

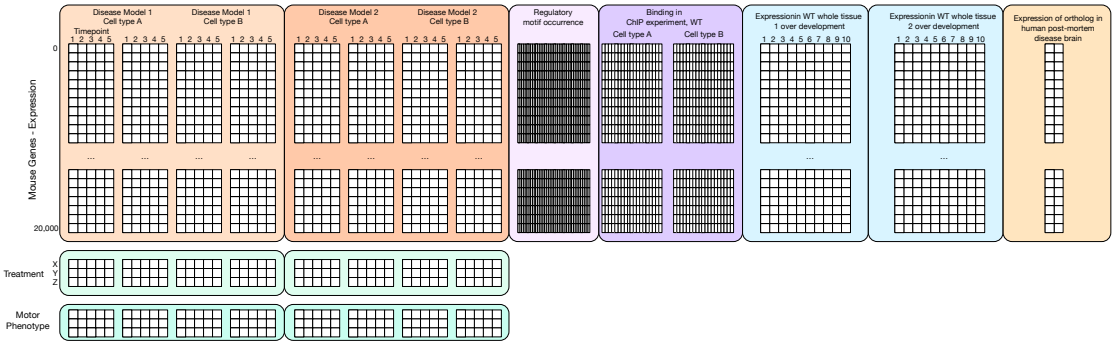


Figure 1.1: Cartoon of complex, structured data integrated from multiple sources

the input data is assumed to be fixed by experimental or analysis designs, and propose BOMBASTIC (Block-Organized, Model-Based, Tree-Indexed Clustering), a framework for interactive clustering that is guided and constrained by the structures of the inputs. (The case in which structure is incompletely specified, and must be learned, shall be left for future consideration.)

1.3 Motivating problems in molecular neuroscience and neurodegenerative disease

Neuroscience presents some particularly pathological examples of structured data analysis problems involving processes that are both dynamic and context-dependent. There are thousands of molecularly distinct neuronal subtypes, and neurons are adaptive to changes in inputs over both long and short time-scales. Neurodegenerative diseases, such as Huntington Disease (HD) and Parkinson Disease (PD), preferentially affect specific anatomical regions and cell types. Explaining this selectivity is central to understanding pathophysiology and informing development of more effective and tolerable therapies.

1.3.1 Levodopa-induced dyskinesia

Parkinson disease is caused by death of cells of the *substantia nigra*, which normally provides dopaminergic inputs to the striatum. The main therapy for Parkinson's is to increase levels of dopamine in the brain by supplementation with the metabolic precursor to dopamine, L-DOPA [64]. Over time, however, L-DOPA therapy leads to maladaptive changes in medium-spiny neurons, resulting in abnormal striatal function and debilitating

levodopa-induced dyskinesias (LID) [60, 18], which ultimately limits the clinical utility of L-DOPA. Defining the molecular mechanisms responsible for LID is therefore important so that we may identify and prioritize targets to modulate to better manage Parkinson's disease and L-DOPA side-effects.

Levodopa-induced dykinesia results from dysfunction of specific sub-populations of striatal neurons, and develops over time. To investigate the mechanisms of LID pathogenesis, our collaborator, Dr. Myriam Heiman, conducted an experiment in which cell-type specific gene expression was measured in a mouse model of LID. The resulting dataset offers an example of the kind of high-dimensional, structured data discussed above; measurements over 20,000 genes are indexed by cell-type, time, and different L-DOPA treatment regimens.

1.3.2 Huntington Disease

The genetic cause of HD was discovered in 1993 to be a trinucleotide expansion in the HTT gene, which codes for a expanded polyglutamine repeat in the *huntingtin* protein [130]. Despite decades of study of *huntingtin* biochemistry and genetics, relatively little progress has been made in defining the molecular mechanisms mediating the disease process or in developing disease-modifying therapies. HD preferentially affects medium-spiny neurons in the striatum. Lying deep in the brain, these cells are difficult to study in living patients, necessitating the development of mouse models.

Two enduring mysteries of HD are the repeat-length dependence of age of onset of the disease, and the selective vulnerability of certain cell types to the effects of the mutant Huntingtin gene, despite its ubiquitous expression in both neuronal and non-neuronal cells. Among the many cellular changes that occur in vulnerable cells, gene expression is severely dysregulated, and this is thought to play a central role in HD pathophysiology. This has motivated many studies of the impact of mutant Huntingtin on gene expression (eg. [69]). To study the cell-type selectivity and repeat length dependence of expression changes in HD requires profiling expression over time, multiple cell types, and alternative mutant Htt alleles and models, again generating highly-structured, high-dimensional datasets to analyze.

1.4 Organization

The first part of this dissertation proposes some general methodology and software for clustering datasets with pre-defined structure. In Chapter 2, we more carefully define the structured clustering problem and review relevant prior work from several disparate fields. Chapter 3 describes BOMBASTIC, a software tool to facilitate clustering, comparisons, and visualization of such data.

Parts 2 and 3 describe biological analyses which either motivated or applied the methods that were developed. Chapter 4 contains an analysis of transcriptional dysregulation in mouse models of levodopa-induced dyskinesia associated with Parkinson's disease. Part 3 consists of three related analyses of gene expression data in HD models. Chapter 5 reviews the role of transcriptional dysregulation in HD. Chapter 6 describes a re-analysis of several published HD expression datasets to identify potential transcriptional regulators. In Chapter 7, we extend these analyses to cell-type specific expression data. In Chapter 8, we apply BOMBASTIC to interrogate an even larger collection of mRNA-seq expression measured across multiple time-points, tissues, and Huntington alleles.

Part I

Methods and Software for Clustering Structured Data

Chapter 2

Background: Clustering Structured Data

2.1 Introduction

Cluster analysis is one of the principal tools used for exploratory data analysis and unsupervised learning [59]. The objective of clustering is to group similar items together. Many different clustering algorithms have been invented; some are heuristic (eg. hierarchical clustering), others based on fitting to an underlying generative model (eg. mixture models), and many more are inspired by the idea of finding a lower-dimensional approximation of a dataset under various constraints (eg. NMF).

Another way to cluster clustering algorithms is by the type of output data generated. The simplest type of clustering algorithm learns a one-to-one function mapping each multivariate observation to a single discrete label from a finite set. A canonical example of such an algorithm is k-means.

The second major type of clustering algorithm takes multivariate observations and produces an ordered tree in which each observation is represented by a leaf. Agglomerative algorithms such as hierarchical clustering are familiar examples, in which the tree is constructed to minimize the sum of a distance metric along the branches. The resulting clustering tree can be then be cut at an arbitrary depth to produce a simple partitioning of the first type.

Most clustering algorithms operate on a set of objects represented as multidimensional vectors that form the rows of a matrix, in which the columns are considered to be exchangeable. That is, re-ordering the columns is assumed to not have an effect on the results of clustering. In many real-world situations, where data does have consider-

able structure, this may be neglecting useful information. We will refer to data in which columns are grouped and/or ordered as having *block structure*. For example, when data is measured over time, the order of observations is clearly important, and it is possible that certain time points are more informative than others. When data is measured across two different contexts (for example, gene expression time-series measured after treatment with two different drugs) the sets of columns corresponding to each context should be considered separately, and in the correct order. A clustering algorithm that produced the same results from a version of such data in which columns were randomly permuted would potentially be discarding information and producing less interpretable results.

A related, but distinct and much older [4] problem than clustering is that of constructing or learning taxonomies. Taxonomies can can be useful for discovering, representing, and reasoning about natural structure. Taxonomies also have important practical uses in identification or diagnostic keys, which provide a simple algorithm to identify objects by recursively partitioning a space of possible labels using a sequence of tests, which can be dichotomous or polytomous. Classical examples of these kinds of taxonomies are the keys in field guides used identify plants by visual features of their leaves, or diagnostic keys to identify a disease name by from a sequence of questions and tests. Taxonomies are constructed using a set of pre-defined discrete *taxonomic characters*, features by which the objects being classified can be distinguished, and these characters and the order in which they are used are carefully chosen to be easily observable and maximally informative (i.e. to partition the space of possibilities in a minimal number of steps). Very often, the features used to partition at each level also have functional or phylogenetic interpretations; good features "carve nature at its joints" [97].

When high-dimensional data is also highly structured, the problem of finding an informative and interpretable partitioning can involve aspects of both unsupervised clustering and of taxonomy construction. *Within* the subspaces (blocks) defined by experimental designs or different data types, one must rely on unsupervised clustering. Depending on the nature of the data, different algorithms or clustering objectives may be appropriate for partitioning within each of the subspaces. But combining information across pre-define subspaces is more akin to taxonomy construction. Instead of using taxonomic

characters derived from easily observable features, one can rely on unsupervised methods to learn the distinctions at each level, and then combine those learned features into a taxonomy. Since high-dimensional, block-structured data is increasingly common, clustering it in this semi-supervised manner is becoming an important practical problem, not generally addressed by currently available tools.

Since trees and hierarchies arise in many clustering algorithms, it is also worth clarifying a distinction between the approach sketched out above and some existing methods. While hierarchical clustering and other agglomerative clustering algorithms involve learning a tree, such a clustering tree is not a taxonomy in the sense we have defined above, because the branching of the tree at every level is determined by the same distance metric. The branching at different depths in a hierarchical clustering tree represents differing scales of viewing a projection of the data onto a single dimension, rather than representing conceptually different types of categorizations (i.e.. distinct taxonomic characters).

2.1.1 A motivating example: CHDI HD Allelic Series Dataset

As an example of biological data with significant structure, we consider a recent dataset generated and made available by the CHDI Foundation (CHDI) (http://www.hdinhd.org). To study the mechanisms of Huntington's Disease, CHDI measured gene expression in an allelic series of knock-in mouse models of HD, heterozygous for mutant Huntingtin genes with CAG repeats of varying lengths (WT, Q20, Q80, Q92, Q111, Q140, Q175). Each type of mouse was studied over development, and phenotype monitored by a variety of motor and behavioural assays. Mice were sacrificed at 2 months, 6 months, and 10 months of age, and at each of these time-points, mRNA (23,351 genes) and miRNA (626 miRNAs) expression was measured from various brain structures and other tissues: striatum, cortex, liver, cerebellum, and hippocampus. Several more tissues – Gonadal adipose, Intestinal white adipose, corpus callosum, Hypothalamus, Brainstem, Skin, Gastrocnemius, Heart, and Brown Adipose – were also characterized in the WT and Q175 mice at the 6 month time point only. The organization of this data is depicted in Figure 2.1.

It is often not obvious how or where to begin looking at such data. A naïve strategy might be to lay all of the data out in a very wide matrix, and cluster the rows of this

matrix. However, a biologist might ask if some genes have similar profiles in the striatum, but not in the cerebellum? What if the dysregulation due to HD is only apparent at the 10 month time-point? Or if the dysregulation begins only when the repeat length exceeds some critical value? There are many such questions to ask, and it could be tedious to explicitly write programs to consider every possible question serially. Instead, one might prefer to parameterize the space of all possible scientific questions and their attendant computational analyses, to allow questions to be posed and answered (almost) as quickly and easily as they could be formulated.

2.1.2 Importance and sources of block structure

Many biological datasets have natural structure, because the common set of genes encoded in the genome are reused in different ways across different tissue and cell types and in different responses to stimuli. Experimental designs, by choosing which cell types to measure, and at which time points, encapsulate prior knowledge and hypotheses about the experimental contexts thought to be relevant to specific scientific questions. Retaining this structure throughout analysis is therefore important, so that results can be related back to the scientific questions that motivated experimental design decisions.

Data is often acquired from multiple contexts simultaneously in profiling experiments that measure every accessible or potentially relevant context. For any particular scientific question, however, only a subset of contexts or times may be relevant, and some may be more relevant than others. Sometimes these decisions are completely obvious and unambiguous, and are implicit in the way an experimenter chooses to analyze data. Often, however, the choices of subsets of the data to examine are less obvious. It would therefore be useful to be able to easily and rapidly choose alternative subsets of contexts of interest to formulate and prioritize alternative scientific questions.

Clustering that makes use of subspace structure also produces more interpretable results, because a potentially large number of possible partitionings can be succinctly encoded as paths through a tree. This can enable navigation through the resulting space of partitionings along scientifically meaningful dimensions. For example, one would like to be able to express queries such as "the set of genes encoding transcription factors that

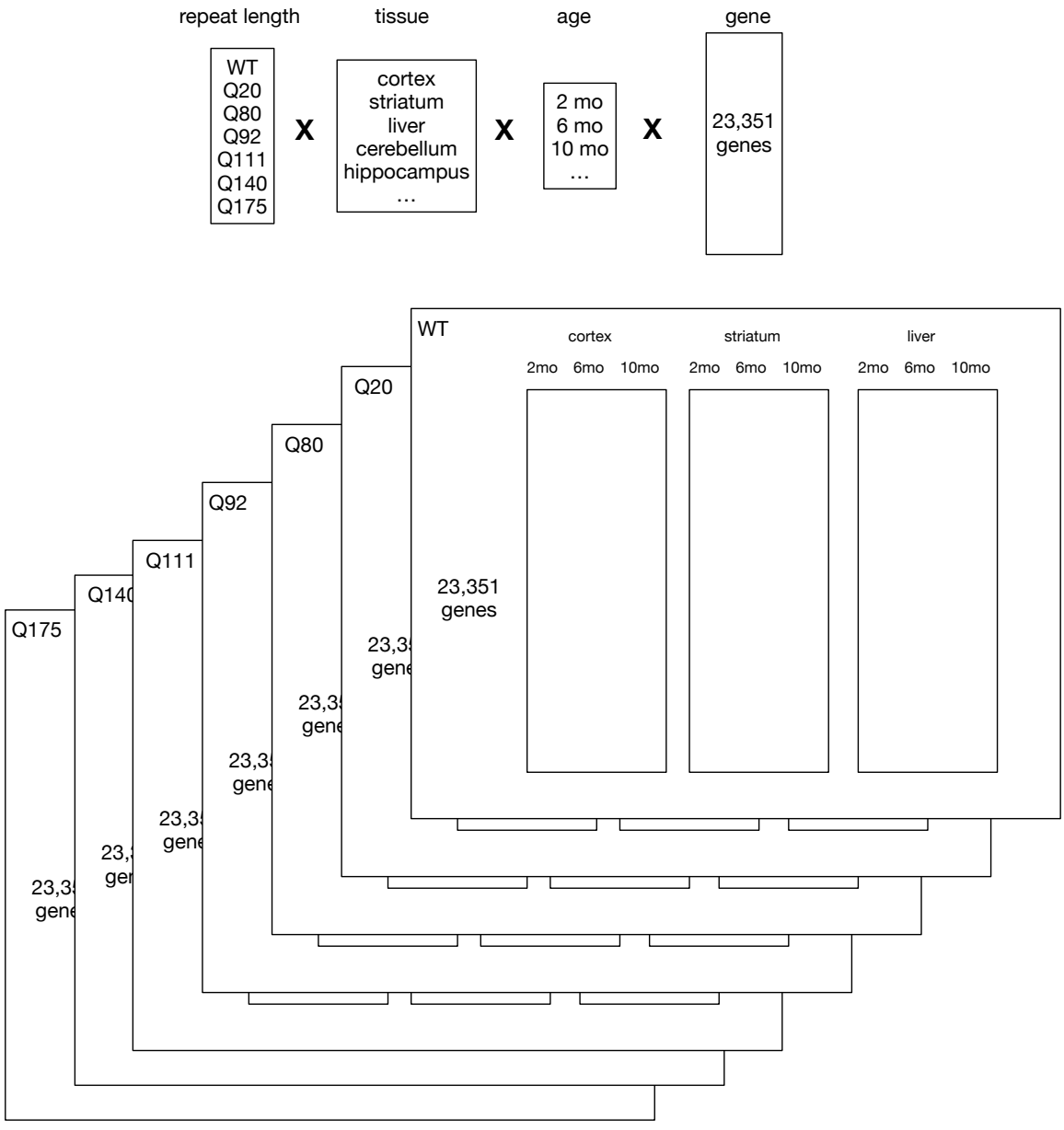


Figure 2.1: Schematic outline showing multidimensional structure of CHDI Allelic Series mRNA expression dataset

are down-regulated beginning at 6 months of age and continue to decline monotonically in the striatum of HD model mice, which have striatal-specific expression in the wild-type, and which have constant expression in the cortex of HD mice and in the striatum of wild-type mice", as opposed to "cluster #317".

Statistical considerations also motivate preservation of subspace structure in clustering. When distributions, dynamic ranges, noise characteristics, and sampling resolution vary, different transformations, clustering algorithms, and parameter settings may be optimal for each block, so a block-based approach might be better at recovering the true underlying structure than simply treating all the data homogeneously. Time- or location-indexed data also demands special consideration to impose prior beliefs about continuity and to be able to effectively distinguish differences in shapes, relative magnitudes of changes, and absolute levels, and any of these characteristics may differ between subspaces.

Useful applications of structured clustering will require integration of the clustering algorithms and models themselves with tools for visualization, navigation and filtering of results, and we will need effective representations to expressively and efficiently specify and manipulate desired analyses and computations. Next, we review relevant prior work from three areas: clustering algorithms for structured data (and for learning structure), with a focus on applications to biological data; methods for visualization of clustering results and for navigation and querying of complex datasets; and declarative approaches to formalizing search queries and specifying statistical graphics and analyses.

2.2 Clustering structured data

2.2.1 Numerical taxonomy

In the 1960s and 1970s, biological taxonomists discovered computers and developed the field of *numerical taxonomy* [118, 32], which applied quantitative similarity metrics and numerical clustering algorithms to the task of classifying organisms. These methods were applicable both to phenetics, in which organisms are classified by phenotypic features, as well as to the cladistic and phylogenetic methods that have come to dominate modern

taxonomy. Classifying organisms by phenotype requires first choosing a set of characters to use. For example, to classify micro-organisms, Sneath [117, 32] proposes using characteristics such as morphology (number of flagella, shape of spores), biochemical properties (anaerobic vs aerobic; oxidase activity), drug sensitivity (penicillin sensitive?), etc. When used to classify objects in a taxonomy, such properties are referred to as *taxonomic characters*. In the task of classifying or identifying organisms, these characters have typically been chosen by experts to be easily observable and informative.

The distinctions between the different *kinds* of characters used in phenetics are usually clear-cut: morphology is one thing, and drug sensitivity is another, and there are typically a limited number of easily observable characteristics to work with. When these methods are applied to more abstract and plentiful data, it may be less obvious whether different variables reflect different characters, which variables should be considered together as groups, and which variables and combinations should be used at all to construct a useful taxonomy or partitioning of objects.

2.2.2 Biclustering, 3D biclustering, and Plaid Models

Biclustering methods address the question of finding subsets of variables that are relevant to distinguish only subsets of objects, usually for the case where there is no prior suggestion of a natural organization of the variables. Biclustering [82] was popularized by [23] as a method that "groups items based on a similarity measure that depends on context", relaxing the assumption of standard clustering methods in which all conditions (columns) are given equal weight. The objective in biclustering is to simultaneously discover subsets of both genes and conditions with similar profiles. This is potentially a much more computationally difficult problem than the one to be addressed by the BOMBASTIC method to be introduced in this dissertation (Chapter 3), which assumes that the column subsets are pre-specified, and that a user explicitly chooses which blocks to use.

Either ordinary clustering or biclustering can be extended to data with higher dimensional structure. For example, if gene expression is measured across both time and multiple conditions, the resulting data set can be imagined as a three dimensional matrix, or equivalently, as a set of two-dimensional matrices that are aligned on one or both axes.

TRICLUSTER [143] is a graph-based clustering algorithm that extends the idea of biclustering to three-way data, such as that indexed by gene, condition, and time. TRICLUSTER searches for clusters that are *homogenous* across two of the dimensions, such as genes that have the same temporal pattern over all of the experimental conditions. While this is one potentially useful objective, note that it might also be biologically interesting to discover clusters that have *different* patterns in different conditions, or clusters that exist only in a single condition.

Strauch et al. [124] proposed an interesting 'two-step' algorithm for 3-dimensional genes-time-condition data. In an example application, the levels of 23,000 genes were measured under 9 abiotic stress conditions, each at 6 to 9 time points. For their two-step algorithm, k-means clustering is first used to cluster data for only one of the conditions. The profiles for the corresponding genes in each of the other conditions are then compared to their cluster assignments learned in the first condition. The modules are categorized as either single-response modules, which cluster only in the seed condition but not in the others, coherent-response modules, which cluster together and have the same temporal pattern in all conditions, or as independent response modules, which cluster in other conditions, but have distinct dynamic profiles in each condition. This entire process is then repeated using each of the conditions as the seed to learn initial clusters. The Strauch et al. algorithm is thus able to identify clusters across subsets of conditions whether or not the exact profiles are condition-dependent.

EDISA [127] extends the ISA biclustering algorithm, which performs matrix factorization with some additional thresholding and constraints. EDISA extends to 3-way data by considering a fixed time-course vector for each (object, condition) pair. EDISA iteratively samples observations from the data and assigns them to modules, which as in [124], are categorized as being single-response (clustering in one condition only), coherent response (similar profiles over all conditions), or independent responses (a common set of genes with potentially different profiles in different conditions).

2.2.3 Functional and time-course clustering

Clustering time-indexed and other functional data has motivated development of specialized algorithms, and many strategies are reviewed in [58]. The simplest approach is to use standard clustering algorithms with the common distance metrics such as correlation. This mostly ignores the time-dependent character of the data. A common improvement is to transform the raw observations to a more meaningful basis, for example by fitting splines, and then using the parameters of the spline fits as the inputs to standard algorithms, as in [81].

2.2.4 Time-course clustering for biological data

Several algorithms have been developed specifically for clustering biological time-course data, which tends to be short and noisy. STEM, the short time-series expression miner, developed by Ernst et al. [35] is a notable example of gene expression time course clustering. STEM begins with the idea of enumerating all possible patterns using a fixed, quantized step size between successive time points. To reduce the number of clusters, STEM proposes a greedy algorithm to maximize the diversity of the chosen set of potential cluster profiles for a specified number of clusters. The choice of these patterns is independent of the data, and is determined solely by the quantization scheme and number of clusters specified. Genes are then assigned to profiles based on the correlation coefficient between the measured profile and the cluster pattern. STEM also supports comparing clusterings between two conditions, using the hypergeometric test to assess overlap between the set of genes assigned to each cluster in each condition.

2.2.5 Model-based clustering of multi-factor data

Many standard clustering methods, such as k-means, can also be viewed as fitting probabilistic generative models to data [38]. For multi-factor data such as gene expression measured over conditions and time, hierarchical mixture models can be used to model the effects of the various factors. For example, Jörnsten and Keles [62] proposed using 2-level Gaussian mixture models for clustering, fitting the models using expectation-

maximization. The parameters of such models can be interpreted in different ways to explicitly encode alternative scientific questions, such as modeling differential expression between conditions at each time point separately, modeling the trajectories of differential expression between time points, or comparing expression levels at individual time points.

2.3 Visualizing Clustering Results

2.3.1 Multiple clustering visualization

Once clusters have been found, by whatever methods are used, the clusters must be presented to the user in an accessible way. For visualizing clustered data matrices in biology, heat maps and 'cluster-grams' [34] are ubiquitous. Such views are static and rows can be arranged in only one order. While a carefully chosen ordering can be used to emphasize relationships between different subsets of columns, heat maps on their own do not provide an effective way to compare alternative blocks and orderings.

StratomeX [74] is an interesting tool developed for visualization of cancer subtypes, where a set of samples are partitioned using a variety of different types of data. StratomeX has similar motivations to BOMBASTIC and shares the concept of composing analyses by relating blocks of data. Each (fixed) partitioning by some type of data (e.g. RNA expression, miRNA expression, mutation status) is represented in a column, and partitions within each data type are drawn as blocks. Ribbons are drawn to represent intersections between partitions across blocks, adopting the Parallel Sets idea originally presented in [68]. Earlier applications of the parallel coordinates / parallel sets method to compare multiple partitionings include [144] and [46]. StratomeX also provides 'dependent' columns that display a representation of a dependent variable within a selected subset of the data.

StratomeX was described as a visualization technique aimed specifically at comparing pre-computed cancer subtype stratifications. A related method, Domino [42], has also recently been proposed to aid in the manipulation of subsets across multiple tabular datasets, reinforcing the emerging recognition of the importance of this class of data analysis problems. BOMBASTIC has been developed contemporaneously with both of these systems [48], and while it includes a visualization component, aims to provide a

more generic methodology for structured clustering, as well as to offer specializations for time-indexed data.

2.3.2 Faceted Search

An alternative way to frame the problem of analyzing and discovering scientifically interesting subsets of items within a large dataset is as an information retrieval or search task, in which the scientist's job is to formulate queries. Unfortunately, scientists' queries are often vague and uncertain. One important technique that has been developed in the field of information retrieval to help users explore complex databases in the face of uncertaintly and poorly specified queries is faceted navigation and search [134, 104], and BOMBASTIC may also be considered as an attempt to provide a dynamic, faceted navigation system to query structured quantitative data.

Faceted search extends the notion of a fixed, hierarchical taxonomy to permit dynamic, iterative composition of *facets* drawn from separate taxonomies that describe different aspects of objects. Each individual facet is itself a "hierarchy formed using a [distinct] characteristic of division" [134] (i.e. a facet is similar to a taxonomic character, though facets may be hierarchical themselves). Specific points within these facet hierarchies can be selected, and choices from multiple facets can be *combined* to define sets of objects matching all of the chosen predicates. The number of facets used, and the order in which selections are made is flexible, and a faceted classification system is "hospitable" to extension with new facets that do not fit into the existing hierarchies.

The original faceted search system was the "colon classification" library system by Ranganathan in 1933 [134]. More recently, faceted search has become nearly ubiquitous in both e-commerce and document information retrieval systems. Faceted databases can be queried with combinations of boolean predicates on the facets, and this process can be facilitated by providing interfaces that list the possible parameters for each facet. Modern *faceted navigation* extends the parametric search idea by dynamically updating the interface to show only the allowable remaining parameters as a user iteratively refines a search.

An extension in some modern systems is to allow classification and identity within a

facet to be either fixed or dynamic. In the colon classification system, the taxonomies used for each the facets (describing aspects such as location (Earth->USA->Massachusetts->Boston) or time (20th century -> Late 20th century -> 1980s) are pre-defined. Modern faceted search systems are often used for semi-structured datasets that may have some rigidly defined facets, but also permit 'dynamic facets' to be defined by full-text queries. These 'search' facets have typically been unstructured, although a recent extension has been to construct hierarchical facets dynamically [27, 3] using unsupervised methods such as topic models.

2.3.3 Dynamic Queries

Clustering provides a mechanism to assign discrete, hierarchically organized category labels to observations, against which to formulate queries. However, many of the properties on which one might like to filter are continuous variables, such as parameters used in clustering algorithms or any other statistics computed from the data. Dynamic query interfaces [1] provide graphical representations of parametric queries and statistical summarizations of data, allowing users to interactively select subsets of observations by direct interaction to define regions of interest on plots of variables and their distributions, and to see how the distributions of different variables relate to each other. An important design goal of BOMBASTIC will be to facilitate queries over both discrete categories and associated continuous properties simultaneously.

2.4 Formalizing Statistical Graphics and Analyses

The practical implementation of any computational method requires some formal representation that a computer can execute. For methods that are potentially complicated and may change frequently, it is worth considering how to best specify computations in ways that are easy both to understand and manipulate.

Most standard statistical algorithms, including those for clustering, have readily available implementations in open source packages, such as those offered by the R project [98] or Python's scikit-learn [95].

Composition of algorithms to perform more complex analyses is typically done by

writing programs in imperative or functional languages. For certain classes of problems that tend to recur often, writing ad-hoc programs for every new instance may become tiresome. Instead, a common approach is to develop domain-specific languages that can more succinctly and elegantly encode programs to solve particular classes of problems.

Two problem domains in which this strategy has been fruitfully applied are the querying of databases and construction of statistical graphics. In both cases, specialized declarative languages enable the specification of desired outputs, and those specifications can be automatically transformed into the detailed sequences of operations needed to produce those results. This enables more concise expression of programs which are often easier for users to understand and manipulate because program semantics map closely to the user's conceptualization of the problem at hand.

2.4.1 Database Query Languages

Database query languages are one of the most familiar applications of declarative programming. Relational databases [25] are defined using formal data definition languages, and modified and queried using SQL, structured query language. On-line analytical processing (OLAP) offers a model for querying multidimensional data [103], using a declarative query language called MDX (Multi-Dimensional eXpressions). Both SQL and MDX provide basic analytical operations, such as computing summary statistics (eg. means, sums, max, min). However, such languages do not provide support for more complex analyses, such as clustering.

2.4.2 Graphics Algebras

In statistical graphics, Wilkinson's Grammar of Graphics [141] has been extremely influential. The Grammar of Graphics (GoG) formalized the specification and construction of many kinds of statistical visualizations. Wilkinson's formal syntax included six kinds of statements: Data, Transformations, Scales, a Coordinate system, Visual elements (and rules to specify their aesthetics), and Guides (such as axis labels). These statements provide a concise and easily manipulated specification of a visualization, and can be automatically rendered to produce graphical output. The GoG was implemented in R by the

ggplot2 package, which [140] has become one of the most popular ways to construct visualizations in R.

Stolte's Polaris system [122, 123] developed a declarative algebra, along with a corresponding visual representation, to facilitate interactive manipulation and specification of visualizations of relational data that may have hierarchically structured dimensions. Polaris is primarily distinguished from GoG by the model of the data on which it operates [122]. GoG proposed its own data model to represents sets of variables, and provides a variety of operators to support statistical transformations and relations between variables. Polaris, in contrast, was designed to work strictly on data conforming to the relational model.

GoG and Polaris are similar, however, in that they are primarily designed to operate on datasets with fixed structure. While both systems support basic summarizations over a fixed set of dimensions, they are not designed to facilitate clustering and other such algorithms that project an entire dataset into a new space, nor to facilitate the visualization and comparison of clustering results.

Neither relational query languages nor graphics algebras capture the semantics of clustering. A central goal of BOMBASTIC is therefore to offer a declarative formalism to specify clustering analyses, which could be more succinct and more easily manipulated than the *ad hoc* imperative code typically used.

2.5 Summary

Structured high-dimensional data is becoming more prevalent. Clustering such data in a way that makes optimal use of the structure may benefit from an extension of the standard view of clustering with old ideas borrowed from taxonomy construction and some newer ideas from information retrieval. Special kinds of data, such as time series, often demand specialized clustering algorithms, and different subspaces of a dataset may require application of different algorithms and parameter choicess. Exploring the large space of potential alternative clustering analyses would benefit from having succinct and easily manipulated computational representations. Declarative programming approaches, which have found successful application to both querying and visualization of databases,

may offer a means to formalize and automate the clustering analysis of structured data.

In the next chapter, we propose BOMBASTIC, an attempt to formalize the specification of clustering analyses on structured data, and implement software to construct, query and visualize structured clustering analyses and their results.

Chapter 3

BOMBASTIC: Block-Organized, Model-based, Tree-indexed Clustering

BOMBASTIC is motivated by the idea that many scientific questions can be framed in terms of comparisons between clusterings or partitionings of various subspaces that describe different aspects of the same set of objects. In this chapter, we attempt to formalize and decompose the problem of clustering structured data, and propose methods to address each of these subproblems. We then describe the design of software that implements these methods. Finally, we compare and contrast BOMBASTIC to related work, and discuss how BOMBASTIC, within its limited scope of structured clustering, offers an example of a declarative, interactive, and visual approach to the construction and exploration of complex analyses. An application and evaluation of BOMBASTIC on a real biological data analysis problem will be described in Chapter 8.

BOMBASTIC offers a simple and general methodology for clustering data that is organized into multiple predefined blocks or subspaces. We assume that these blocks define distinct taxonomic characters, and then use transformations and clustering algorithms of choice to learn partitionings of the observations *within* each block. By treating blocks independently, optimal algorithms and parameters can be used to cluster objects within each block. Re-combining the blocks into taxonomies allows one to make more fine-grained distinctions between objects and to compare and contrast clusters between blocks. Although the underlying statistical methods used are not novel, by formalizing and automating the construction of these trees of clusterings, BOMBASTIC enables more efficient and comprehensible enumeration and exploration of potentially large spaces of alternative clusterings and trees.

3.0.1 Problems Addressed

Accounting for block structure in clustering will require:

- Clear specification of the block structure of input data
- Clustering each block individually, using appropriate algorithms and parameters. In particular, it is important to provide interpretable and queryable clustering outputs for time- and location-indexed data.
- Selection and recombination of block clusterings into a taxonomy
- Efficient interactive exploration of the resulting tree of partitionings

3.1 BOMBSTIC Methods

3.1.1 Overview

We describe the main concepts and steps involved in BOMBASTIC, which are summarized in Figure 3.1.

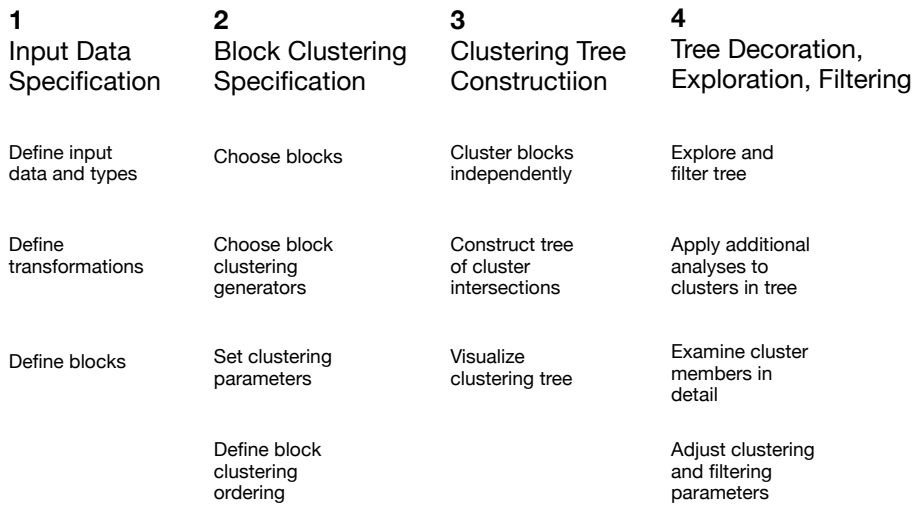


Figure 3.1: High-level overview of BOMBASTIC methods and usage

First, some data to analyze is chosen. The types of the items being analyzed must be specified explicitly and precisely, to ensure that only meaningful combinations of clus-

terings are permitted, and to allow annotation of results with ancillary information. For example, one might be clustering mouse genes, indexed by gene symbol.

Blocks are defined; these specify a sequence of the data columns to be used as inputs to some clustering algorithm. In the CHDI data (Chapter 8), one might define a block for the expression time-course in the wild-type mouse striatum, additional blocks for the expression time courses in each of the knock-in models. Blocks might also be defined based on relative expression in each cell type from a different dataset [30], to allow filtering for cell-type-specific genes.

BlockClusteringResults are produced by a BlockClusteringGenerator. Clustering data requires the specification of the Block itself, the clustering algorithm to use, and any parameters required by the clustering algorithm.

A BlockClusteringResult contains a partitioning of that block into clusters, as well as any ancillary statistics (on each object) from the clustering algorithm, such as parameters of any transformations applied (eg. centering and scaling) or statistics associated with the cluster assignment itself (such as a measure of fit, cluster assignment probability, or p-value). A BlockClusteringResult may be queried to report the membership of one or more of the clusters, and these sets may be filtered by predicates defined on any of the attached statistics.

A sequence of BlockClusteringGenerators produces a sequence of corresponding BlockClusteringResults.

Such a sequence of n BlockClusteringResults (or FilteredResults) implicitly defines a tree of depth n+1, in which each node contains a subset of the items. The root of this tree contains all items in the domain. The children of each node X are formed by taking the intersections between the items in X and each of the clusters in the next BlockClusteringResult in the sequence. This procedure is explained in more detail in Figure 3.2.

This tree makes explicit all of the potential relationships and intersections between the BlockClusterings in a sequence. The leaves represent the intersections between clusters for all of the combinations of clusters produced by the cross-product of all of the clusterings, while the internal nodes reflect successively refined subsets. The leaves will be the same regardless of the ordering of the BlockClusterings in a sequence, while the

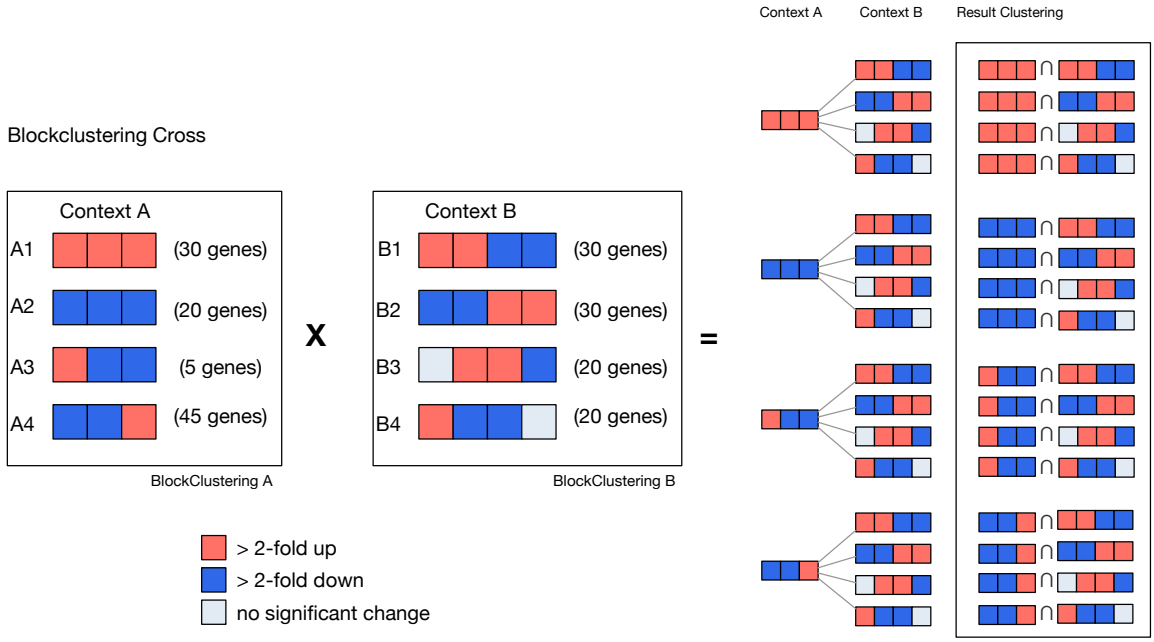


Figure 3.2: Cartoon example showing basic BOMBASTIC cross-product operation between two BlockClusteringResults. Applying this operation recursively to a sequence of BlockClusteringResults generates a tree.

nodes along the internal paths depend on the sequence in which the BlockClusterings were specified.

Picking a node in this tree represents a selection of the items falling into a particular combination of clusters, or behaviours over the specified blocks. The set of items in every node, or in a particular node, can then be further analyzed by inspection of the items, their raw data, and annotations, or by applying some other function to the set of items specified by that node (such as an over-representation test.) Such analyses can be applied to a single node at a time, or to all of the nodes in the tree, and the results can then be encoded into a visualization of the tree.

3.1.2 BOMBASTIC Concepts

We briefly describe the main concepts that need to be represented by BOMBASTIC objects.

Dataset

A dataset consists of a matrix of numerical values, with rows indexed by an object identifier and columns indexed by some observation identifier, along with row (object) and column (observation) metadata.

Transformations

A transformation takes a dataset and produces a new one which has the same row index as the original.

Filters

A filter takes a dataset and produces a new one which has the same row index type as the input, but which contains only a subset of the rows.

Blocks

A Block specifies a sequence of columns chosen from a dataset to be used as input to clustering or visualization, and retains a reference back to the source dataset.

Block Clustering Algorithm

A BlockClusteringAlgorithm takes as input a Block and any required parameters, and produces a partitioning (BlockClustering) along with a dataset containing any statistics associated with the clustering. We assume a hard clustering in which each item is assigned to exactly one cluster.

Block Clustering Generator

A BlockClusteringGenerator binds a Block with a particular BlockClusteringAlgorithm and settings for any parameters required by the clustering algorithm, and produces a BlockClusteringResult. These results may be computed on-the-fly, or pre-computed.

Block Clustering Sequence

Any particular BOMBASTIC analysis is defined by a sequence of BlockClusteringGenerators. Typically this sequence is constructed interactively by dragging around available block clusterings.

Block Clustering Result

A BlockClusteringResult contains the set of clusters and any associated statistics produced by the application of a BlockClusteringAlgorithm to a single block, with a particular set of parameters. Each cluster has an integer ID, a label generated by the clustering algorithm, the parameters that define the identity of the cluster (eg. the centroid or vector of contrasts that it represents), along with the set of items assigned to it.

Block Clustering Result Sequence

The resulting sequence of BlockClusteringResults provides the data that drives the rest of the BOMBASTIC interface.

Block Clustering Result Sequence Query

Any combination of clusters can be extracted from a Block Clustering result sequence. For each BlockClusteringResult in the sequence, a query specifies one or more of the clusters, and optionally, a set of inequalities that further filter the selection by the values of clustering statistics.

Block Clustering Result Tree

A sequence of BlockClusteringResults can also be used to generate a tree that represents all of the intersections between clusters (this is explained in more detail in the next section and in Figure 3.2). This tree can be constructed explicitly, so that analyses can be applied to the sets of clustered objects represented by each node, and so that it can be drawn as a figure.

3.1.3 Block Clustering Methods

BOMBASTIC is intended to be modular and agnostic about the specific clustering algorithms used to produce the independent block clusterings. The major restriction is that clusterings are assumed to be 'hard'; each object should be assigned to exactly one cluster. Initially, we implemented several very simple but widely-applicable algorithms.

Binary label assignments

The simplest possible 'clustering' is to partition objects based the value of some binary label, (eg. indicating membership in a some set). This is useful for restricting analyses to subsets of objects of interest; for example, when clustering genes, one might wish to investigate only the set of transcription factors. Calling such restrictions a clustering allows us to implement this frequent task within the general BOMBASTIC framework.

Real Filters

A slightly more complicated but equally common scenario is that of filtering objects based on the values of some associated real-valued statistics. In gene expression analysis applications, for example, one often wants to filter genes by fold-change or variance. It is therefore useful to be able to define a clustering of objects by specifying a set of ranges for some statistic. Such ranges may be specified interactively by selecting regions on a histogram. While there are well-established tools to filter quantitative data in this manner (eg. Spotfire), providing partitioning based on such filters allows this task to fit naturally into the BOMBASTIC scheme and to be used in combination with other clustering algorithms.

Testing individual contrasts

Building on the above two clustering types is the common case of partitioning objects based on the results of a statistical test for a single contrast between two conditions. This produces a partitioning of the objects at multiple levels (eg. up-regulated, down-regulated, unchanged) based on a combination of filters on both statistical significance

and data values.

Trivial Indexed Quantized Contrast Clustering (TIQCC)

Time-course data can be represented as a sequence of contrasts between successive time-points. Given an observation vector $\boldsymbol{x} = [x_0, x_1, ..x_t]$, one can construct the vector of contrasts $\boldsymbol{c} = [x_1 - x_0, x_2 - x_1 \ldots x_n - x_{n-1}]$

This leads to a *very* simple way to cluster such data, which we call Trivial Indexed Quantized Contrast Clustering (TIQCC). We define a set of q levels for quantization (eg. the intervals between log2 fold changes of $[-\infty, -2, -1, 0, 1, 2, +\infty]$), and then enumerate all possible quantized contrast vectors. We then quantize each contrast vector, and assign it to the matching cluster. (Since many of the possible clusters may be unoccupied, one can start with the data and keep track of only those clusters that have support).

This scheme has several useful properties:

- Observations are clustered by shape, rather than absolute level. This is particularly important when analyzing bio-molecular data, since there are wide variations in dynamic range, and biological relevance is not necessarily related to absolute levels. Moreover, many common experimental techniques measure only relative changes.
- Every potential pattern will be represented; one does not have to choose the number of clusters to use, and even rare patterns will be represented by clusters.
- The number of quantization levels can be adjusted to generate more or less granular clusterings
- The quantization method can easily be extended to incorporate other statistics about the contrasts. For example, if we perform significance tests for each contrast, they can be used as a filter in the quantization and assignment to clusters.
- The algorithm is extremely simple and fast, and scales linearly with the number of observations.

A limitation of TIQCC is that it is only suitable for relatively short time-courses, since if q is the number of quantization levels for each delta, and t the number of time-points,

the number of possible clusters will be q^{t-1}.

Ernst et al. also proposed a time series clustering algorithm based on quantized patterns [36], although the patterns used were not exhaustive, and the choice of patterns to be used as clusters was independent of the observed data.

Another related approach is to cluster time-courses by their derivatives, after transformation to splines. For example, Dejean [29] described an algorithm in which time-course data was smoothed and represented by cubic splines. The profiles were then clustered (using k-means) on the first derivatives of those functions, to cluster the profiles by their shapes rather than absolute levels.

The same approach can also be used for the case in which two conditions are compared over time, such as disease vs. normal. In such an experiment, often it is the contrast between disease and normal that is of the greatest interest, and so the vector of these contrasts can be used without comparing successive time points, although if relevant, one could also cluster by the time-dependent changes in the disease vs. normal changes.

Scaled, centered K-means

Another simple algorithm suitable for clustering short time-course data is k-means. When the shapes of the time-courses are of interest, the data may first be transformed by centering and then scaling to a uniform range. Importantly, the parameters used in these transformations are recorded and propagated to the resulting BlockClusteringResult, so that queries can be specified both on cluster shape (eg. a cluster that is monotonically increasing) and which satisfy additional criteria (eg. having expression above a baseline level and spanning a dynamic range of at least a 4-fold change).

3.1.4 BOMBASTIC Tree

3.1.5 Tree visualization, decoration and interactive filtering

The tree of cluster intersections induced by a sequence of Block Clusterings can be explored in two ways. A cross-filter view (Figure 3.3) provides a compact representation of clusters and permits selection of combinations of clusters. When multiple clusters are

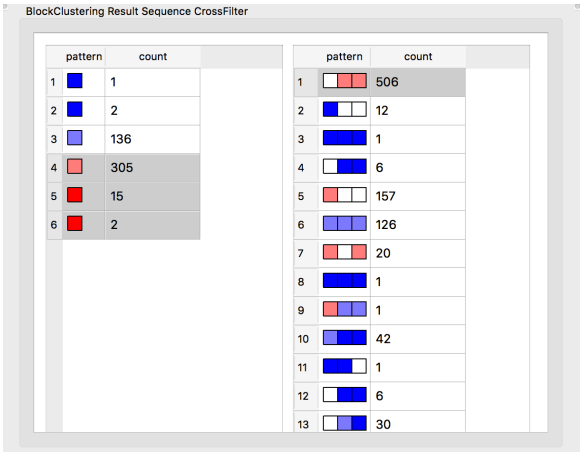


Figure 3.3: BOMBASTIC cross-filter / cluster selection interface. Two BlockClusteringResults are shown. The table on the left shows a partitioning of genes for a test performed at a single contrast (expression change due to HD in the brainstem at 6 months; see Ch. 8). The table on the right shows a TIQCC clustering for the combinations of HD vs. normal contrasts over time in striatum. For example, the current selection in the example (highlighted in grey) queries for the set of genes that were significantly up-regulated (> 1.5-fold) in the HD condition in Brainstem at 6mo **and** which were also up-regulated in HD starting at 6mo, and continuing at 10mo, in striatum.

selected within a block, the resulting selection is the union of those clusters. Selections across blocks represent the intersections between the selections within each block.

The tree may also be drawn explicitly (Figure 3.4, panel 5), which allows all of the resulting subsets to be seen, and the sizes of clusters can be reflected in the visualization. If additional analyses (eg. over-representation tests) are applied to these clusters, the drawing of the tree can be decorated to indicate those clusters with statistically significant or potentially interesting results, or clusters exhibiting over-representation of a specified set.

3.1.6 Pickset annotation and interpretation

After a user has selected a particular set of objects (one or more nodes in the tree), two common analysis tasks are examination of the set members in detail, and functional or mechanistic interpretation. A tabular view (Figure 3.4, panel 6) displays the currently picked set, along with annotations and primary data. To support generation of functional or mechanistic hypotheses, analyses (initially, over-representation tests) are automatically

applied to the current set (Figure 3.4, panels 7 and 8). Even when used in a manual mode, this offers an easy way to browse the results of such analyses applied to a large number of possible subsets.

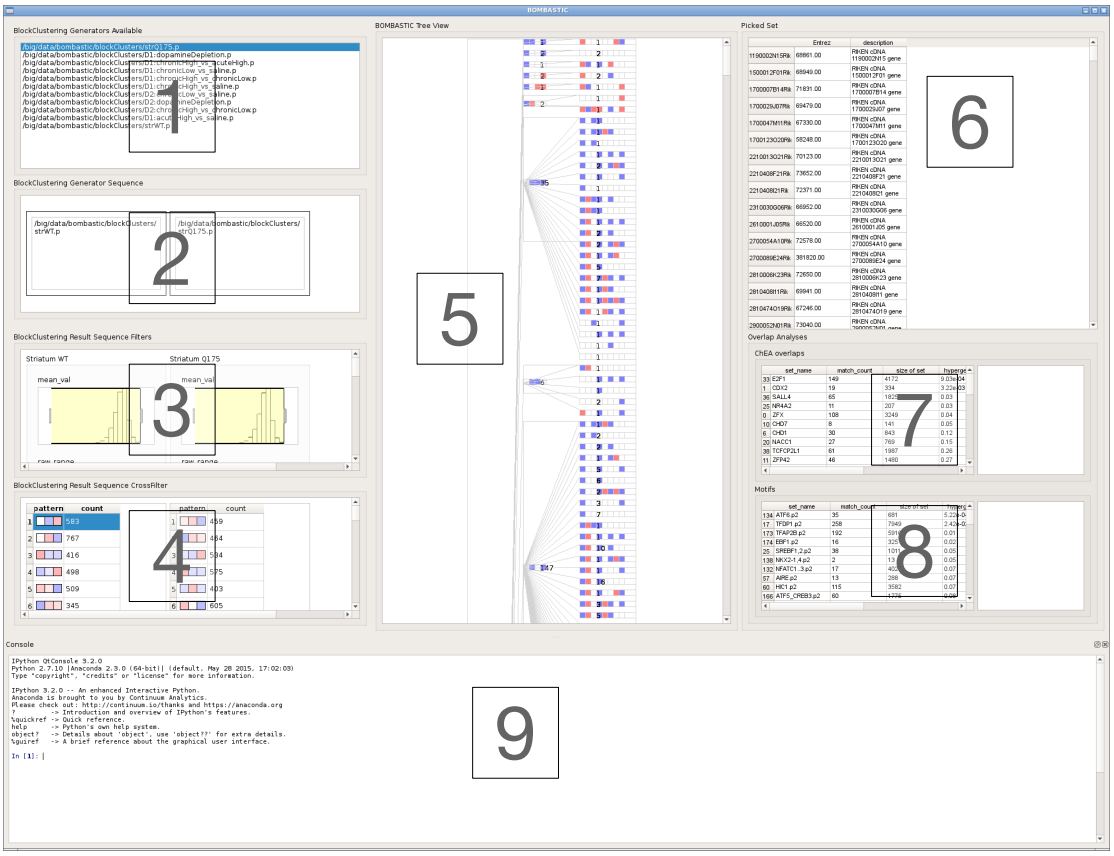
3.2 Implementation

The methods described above were implemented in Python. The user interface was developed using the PyQt (`https://riverbankcomputing.com/software/pyqt/intro`) bindings to the cross-platform Qt 5 framework (`http://www.qt.io`).

The core module defines the python objects representing the concepts outlined in section 3.1.2, and the algorithms for clustering, filtering and tree construction. The user interfaces are implemented in four modules, each of which implements Qt objects to provide models that wrap the underlying BOMBASTIC structures, views that render the interface, and controllers. The AnalysisTemplate module provides support to to manage specification and setup of analyses. The BCResult module display BlockClusteringResults in a grid view and implements selection and filtering. The BCTree module rendering the clustering trees. The Pickset module supports display of the currently selected cluster, shows annotation associated with the individual items, and shows the results of further analyses performed on the selected cluster. A screenshot of the BOMBASTIC interface is shown in Figure 3.4.

3.3 Discussion

The main contribution of BOMBASTIC is to explicate a generalized and modular methodology for block-organized clustering that is relevant to many data analysis problems in biology and to provide a software implementation that facilitates efficient visualization, filtering, and exploration of a large space of potential analyses and their results.



Overview of BOMBASTIC interface. To construct an analysis, the user drags blocks from a menu of available datasets (1) to assemble a sequence of BlockClusteringGenerators (2). Such a sequence suffices to specify the generation of the rest of the analyses, clustering each block independently. Statistics associated with each resulting BlockClustering are shown in histograms (3), which may be used to interactively select and filter subsets of the data. A cross-filter view of BlockClusteringResults is shown in (4), and the full representation of the clustering tree in (5). A user may select any combination of clusters and their intersections using either the cross-filter or the tree view, and the objects (eg. genes) comprising that cluster can then be interrogated in detail (6), and additional analyses (eg. computing over-representation of associated regulatory motifs) applied to annotate the constituents of the currently selected node (7, 8). The entire system is scriptable and can be controlled through an integrated python console (9).

Figure 3.4: Screenshot and overview of BOMBASTIC interface.

3.3.1 Benefits over traditional methods

Comparing sizes of subsets with Venn diagrams

Comparison of discretized changes in gene expression between several groups offers one of the simplest examples of a useful application of BOMBASTIC. Genes might be measured in different contexts and classified as being up-regulated, down-regulated, or unchanged. Venn diagrams are very often used to visualize comparisons between 2, 3, or 4 groups. Beyond 5 groups, however, Venn diagrams become so visually complex that they are unhelpful. Venn diagrams also can only show the sizes of intersections, and the identity (eg. time-course patterns) of each set is indicated only by colors or labels. In contrast, the cross-filter and tree visualizations provided by BOMBASTIC can efficiently and comprehensibly display intersections across an arbitrary number of sets, and the identity of each set and intersection can be directly encoded in the visualization.

Concatenation and clustering

Instead of clustering blocks independently, one could concatenate data and employ standard algorithms. Doing so immediately presents the choice of which blocks to use, which is part of the problem addressed by BOMBASTIC. Once a desirable collection of columns had been concatenated, one would likely be able to find the same partitioning that would be produced by a combination of independent clusterings, assuming that individual blocks were of comparable sizes and had similar characteristics. If the blocks were of different sizes or contained data with very different distributions, it would be necessary to develop specialized clustering algorithms or objective functions to account for this.

3.3.2 Comparisons to related approaches and systems

BOMBASTIC was first presented at VIZBI 2013 [48] in April 2013, and was also described at the NYAS Data Science Learning and Applications to Biomedical and Health Sciences Workshop in January 2016 [47]. There are a number of earlier and contemporaneously developed systems addressing the same class of problems, having both similarities and differences to BOMBASTIC.

Declarative visualization algebras As was reviewed in the previous chapter, BOMBASTIC is inspired by declarative visualization techniques such as Polaris/Tableau [122] and the Grammar of Graphics [141, 140]. These tools, however, formalize the problem of mapping a fixed tabular dataset (potentially with hierarchically structured dimensions) into graphical representations. Their algebras do not include primitives for clustering or for combining combinations of clusterings into taxonomies. Such tools also do not aim to provide an interactive interface that relates the summary visualizations of a clustering (e.g. the cross-filter or tree views of BOMBASTIC) to analyses and visualizations of the constituents of particular clusters (i.e. the individual genes and results of over-representation analysis).

STEM STEM [36], the short time-series expression miner, was an early and influential tool for analysis of biological time-course data. For time-course clustering, the simple TICQ approach we have proposed makes fewer assumptions than the STEM method and avoids attempting to prune the space of possible patterns, while preserving the possibility of identifying clusters that might be sparsely populated. STEM also offered a tool for comparing membership between the clusters of clusterings from two contexts. BOMBASTIC generalizes this to an arbitrary number of independent clusterings, and can generate the combined clustering result formed by all of the intersections.

StratomeX and Domino StratomeX [74] is a visualization tool aimed at the problem of comparing cancer subtype stratifications. BOMBASTIC is distinguished from StratomeX by its goal of providing a formalization of the clustering problem, in which the combination of two block clusterings produces a new partitioning which can viewed used as a 'first-class' clustering itself, whereas StratomeX is primarily described as a visualization method to compare and relate fixed, alternative stratifications.

Gratzl and colleagues [42], (from the same group that developed StratomeX), also recently proposed Domino, a system for "extracting, comparing, and manipulating subsets across multiple tabular datasets". Like BOMBASTIC, Domino recognizes that performing comparisons across heterogenous datasets is an important problem not well served by existing tools. Domino provides a number of relationship operators to connect blocks that indexed by the same object types, and even supports selection of clusters across mul-

tiple partitionings. BOMBASTIC again differs from Domino in providing a cross-product operator that explicitly constructs a new partitioning by combining two independent partitionings, as well as in providing an the explicit tree view. Furthermore, BOMBASTIC supports conducting analyses (eg. over-representation tests) systematically over each subset in the resulting tree, to facilitate searches for interesting paths and cluster combinations, whereas Domino relies more heavily on selection of subsets under the direction of the user. Finally, neither StratomeX nor Domino are specifically intended for clustering of time-course data, which was a major motivation for BOMBASTIC and the simple but effective TIQCC method that it implements.

3.4 Future Work

3.4.1 Implementation improvements and additional features

Interface for Block and Clustering specification

In the current implementation, BlockClusterings are constructed using the python API, and then pre-computed BlockClusteringResults are saved for use by the analysis interface. It would be preferable to provide an interface to configure new BlockClusterings on new data, as well as to facilitate computation of BlockClusterings with different parameter settings through the graphical interface.

Additional Pickset analyses

The existing implementation, which is specialized for analysis of genes, performs over-representation analysis of cis-regulatory motifs and chromatin-binding regulators for subsets defined by any selected clusters of combinations of clusters. These are among the simplest possible analyses that can be done. Application of more sophisticated and modern models of transcriptional regulation would extend the utility of BOMBASTIC for analysis of gene expression data.

3.4.2 Empirical and theoretical analysis of advantages of block clustering

We have claimed that when the data in different blocks have different characteristics (eg. dynamic range, variability, number of actual clusters, and distribution of objects over clusters), it makes intuitive sense to cluster blocks independently and then combine clusters. It will be valuable and necessary to more rigorously study the conditions when this is actually true, and to compare the results and performance of clustering using BOMBASTIC to alternative methods.

3.4.3 Searching for informative clusterings and orderings

A key advantage of a formal representation for alternative block clusterings is the possibility of *automatically* enumerating and searching over alternative analyses to identify those that are potentially informative and of scientific interest. For example, given a (reasonably small) set of blocks or block clusterings, one could enumerate all possible subsequences and generate the corresponding trees. The associated clusters could used as input to over-representation tests (or better models), and each tree assigned a score and ranked based on whether, or how many of its clusters had statistically significant overlaps.

3.4.4 Empirical user studies

Since one of the goals of BOMBASTIC is to make it easier for end-user scientists to more efficiently explore large, structured datasets, it will be important to evaluate its utility in the hands of a larger sample of users. We plan to accomplish this by curating sets of relevant datasets for which BOMBASTIC might be useful to make comparisons, beginning with the neurodegenerative disease datasets discussed in the other chapters. Since user interaction with BOMBASTIC is restricted to the operations permitted by the formalism and interface, (i.e. selection, ordering, and filtering of Blocks and BlockClusterings), it is straightforward to record these interactions for later analysis. To maximize the population of potential users, we may also develop a web-based version of the interface.

Part II

Parkinson's Disease

Chapter 4

Analysis of Transcriptional Dysregulation in Models of Levodopa-induced Dyskinesia

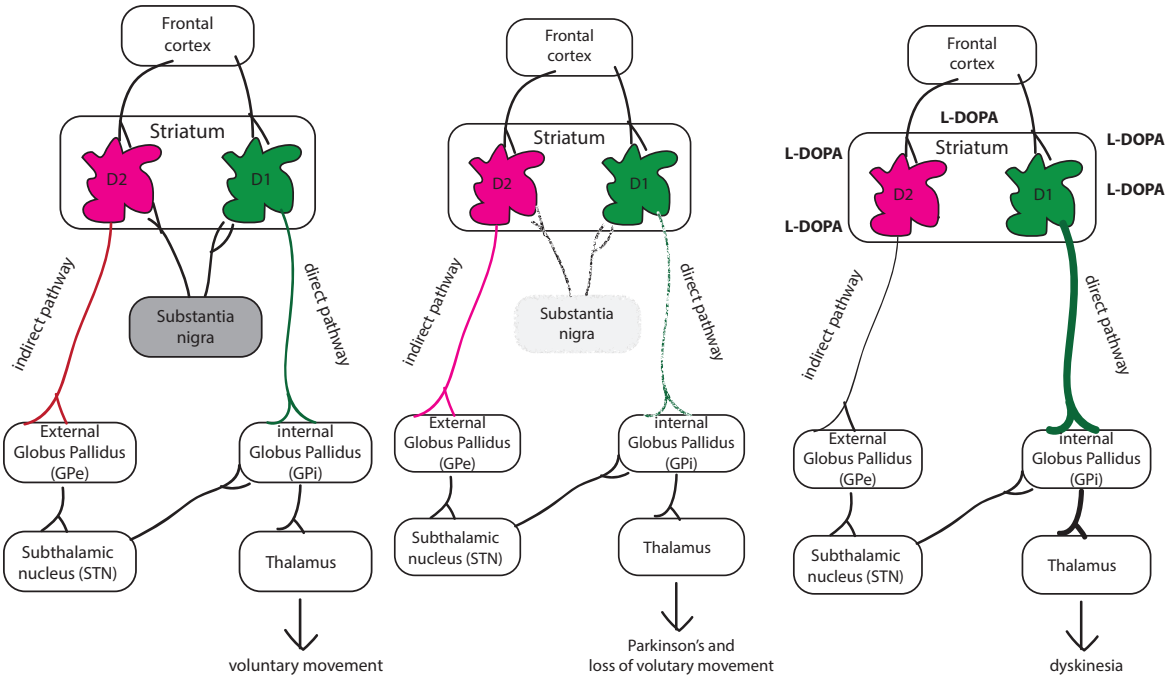
The analysis and parts of the text and organization of this chapter were joint work with Myriam Heiman and published in:
Heiman M, Heilbut A, Francardo V, Kulicke R, Fenster RJ, Kolaczyk ED, Mesirov JP, Surmeier DJ, Cenci MA, Greengard P. *"Molecular adaptations of striatal spiny projection neurons during levodopa-induced dyskinesia"*, PNAS, 2014 Mar 25 111(12):4578-83. .

4.1 Introduction and Background

4.1.1 Parkinson's Disease

Parkinson's Disease (PD) is a movement disorder that affects over 1 million patients in the United States, and 10 million globally. Parkinson's is characterized by akinesia, bradykinesia, rigidity, and tremor, and deficiencies in motor coordination and in movements that are normally automatic [64]. These motor symptoms are the result of degeneration of dopamine-producing neurons in the *substantia nigra pars compacta* that project to the striatum, a subcortical structure consisting of the caudate nucleus and putamen which is particularly important for transmitting signals from the cortex to the basal ganglia and for planning and coordination of motor activity.

In the 1960s, Birkmayer and Hornykiewicz showed that supplementation with levodopa (L-DOPA; L-3,4-Dihydroxyphenylalanine), the physiological precursor to dopamine, dramatically improved motor functions in Parkinson's. L-DOPA, in combination with dopamine decarboxylase inhibition, remains the main pharmacotherapy for PD.



(a) Normal function (b) Parkinson's Disease (c) Levodopa-induced dyskinesia

Figure 4.1: Classical model of changes to the direct and indirect pathways in Parkinson's and Levodopa-induced dyskinesia. Normally, the activities of the direct pathway (primarily D1-expressing neurons) and indirect pathway (primarily D2-expressing neurons) are balanced and effectively transduce instructions from the cortex to facilitate voluntary movement. Both pathways are modulated by dopamine inputs from the substantia nigra. b) In Parkinson's disease, degeneration of the substantia nigra reduces the dopamine input to the striatum. This results in an increase in activity of the indirect pathway, which increases its inhibition of movement, and a decrease in activity of the direct pathway that normally facilitates voluntary movement. c) Excessive, non-physiological stimulation by dopamine from exogenous L-DOPA results in changes to the behavior of D1 and D2 neurons leading to overactivity of the direct pathway and underactivity of the indirect pathway, which causes involuntary movements (dyskinesia). Adapted from Figure 2 of Jenner, 2008 [60]

4.1.2 Levodopa-induced Dyskinesia (LID)

While levodopa is initially very effective, its clinical utility is eventually limited by dyskinesias that frequently develop with chronic treatment. It is thought that these effects are due to the intermittent and non-physiological kinetics of levodopa delivery to the medium-spiny neurons (MSNs) of the striatum, which, over time, cause maladaptive changes in these neurons. Moreover, the depletion of dopamine caused by PD prior to treatment primes and hypersensitize MSNs to the changes that lead to dyskinesia. While there have been efforts to optimize dosing and delivery of L-DOPA, systemic dosing inevitably produces non-physiological exposure of dopamine to the MSNs. It is therefore important to understand the biology of the strial adaptations due to L-DOPA so that pharmacological interventions to prevent or mitigate dyskinesias might be developed.

4.1.3 6-OHDA Hemiparkinsonian model

Rodent models recapitulate import aspects of Parkinson's disease and LID pathology. Stereotactic injection of the neurotoxin 6-OHDA (6-hydroxydopamine) can be used to create relatively specific lesions in the rodent brain [113]. Unilateral injection of 6-OHDA into the Medial Forebrain Bundle (MFB), the nerve bundle projecting from the substantia nigra pars compacta to the striatum, damages the substantia nigra and generates a hemiparkinsonian motor phenotype within a few weeks in mice. This motor phenotype can be ameliorated by administration of L-DOPA. Moreover, prolonged administration of L-DOPA produces a dyskinesia-like phenotype with abnormal involuntary movements, and therefore serves as a model for LID [80].

4.1.4 Intracellular signaling pathways and transcriptional dysregulation in LID

Studies of both non-human primate and rodent models of LID have found profound molecular changes in the neurons of the striatum as levodopa-induced dyskinesia emerges, affecting gene expression, protein expression, post-translational modifications, and synaptic organization [18]. Figure 4.2 summarizes some of the major pathways thought to play roles in dopamine- dependent signaling in the striatum.

Early studies using *in situ* hybridization and staining focused on changes in the expression of neurotransmitters and enzymes involved in neurotransmitter metabolism in the striatum, and found that expression changes of specific mRNAs were were correlated with the emergence of L-DOPA-induced dykinesia and that these expression changes were region- and cell-type specific. [19]

There are two main types of dopamine receptors expressed in the striatum, known as D_1 and D_2. These are G-protein coupled receptors that transduce signals through a number of mechanisms, include cAMP dependent signaling via the adenylyl cyclase-PKA-DARPPP-32 pathway, phospholipase C and IP3 signaling, and cross-talk with MAPK cascades [137], and ultimately lead to transcriptional regulation. Looking at changes in receptor levels in non-human primate models of both Parkinson's and LID, Aubert et al. observed changes in expression and sensitivity of D_1 receptor signaling, both at the level of the D_1 receptor itself and in the activity of downstream effectors including Cdk5 and DARPP-32 [8].

Dopamine depletion in PD leads to hypersensitivity of the neurons that normally respond to dopamine. It is believed that this sensitization is mainly due to changes in the signaling pathways downstream of the dopamine receptors, rather than simply an increase in the number of dopamine receptors expressed [106]. The D1Rs signal through activation of adenylyl cyclase, leading to a number of cAMP-depdendent downstream effects. In particular, excessive dopamine stimulation leads to PKA activity that causes activation of CREB and induction of immediate early genes, and PKA-dependent hyperphosphorylation of DARPP-32, which appears to play an central role in the development of dyskinesia through actions on the ERK/MAPK pathway [106]. Ultimately, these signals lead to profound changes in transcription.

Microarrays have been applied to characterize expression changes associated with the emergence of dykinesias in rat models [67, 43]. In these rat models, not all rats treated with L-DOPA develop dyskinesia over the course of the experiment, so analyses emphasized comparisons between levodopa-treated rats that develop dyskinesia vs. those that do not. In an early study interrogating 8000 genes (of which only 3000 were well annotated at the time), Konradi et al. [67] observed many genes differentially expressed

between dyskinetic and non-dyskinetic rats in striatal tissue. These genes were involved in diverse processes including ion homeostasis, neurotransmission, synaptic plasticity, kinases and phosphatases, stress response and apoptosis, and ribosomal proteins. More recently, Grunblatt [43] compared the effects of pulsatile dopaminergic stimulation, which tends to induce dyskinesia in the 6-OHDA rat model, to continuous dopaminergic treatment, which does not. Among the genes with the greatest expression differences in striatum between these two conditions were several growth factors including Neurotrophin 3, and genes involved in neurotransmission, including multiple glutamate and serotonin receptors. Since these microarray studies have relied on measurements of homogenized tissue, it has been impossible to know whether expression changes are specific to subtypes of neurons present in the striatum, despite the evidence from *in situ* studies and the standard model of striatal physiology which would suggest that the D_1 direct pathway MSNs are most relevant to LID.

4.1.5 Pharmacological Therapy for Parkinson Disease

The drugs used for PD were recently reviewed by Connolly and Lang [26]. Although levodopa remains the most effective treatment for the motor symptoms of PD, a number other agents are also used, especially at early stages of disease, to avoid the high risk of levodopa-associated dyskinesias and other side effects. For mild symptoms, monoamine oxidase type B (MAO-B) inhibitors, such as selegiline or rasagiline are often an initial therapy. By inhibiting MAO-B, which metabolizes dopamine, these increase levels of endogenous dopamine.

Once motor symptoms become severe, levodopa is the main treatment. When dopamine levels are increased, motor function can be restored ('on time'), but when these drugs are metabolized and levels drop, Parkinsonian symptoms return ('off time'). To manage the kinetics of L-DOPA treatment, it is often combined with MAO-B inhibitors, catechol-O-methyl transferase (COM-T) inhibitors, and carbidopa, an inhibibitor of dopamine decarboxylase (DDC), which all help to slow down dopamine metabolism, raising dopamine levels and smoothing out the effects and reducing the dose and frequency of L-DOPA or dopamine agonists required.

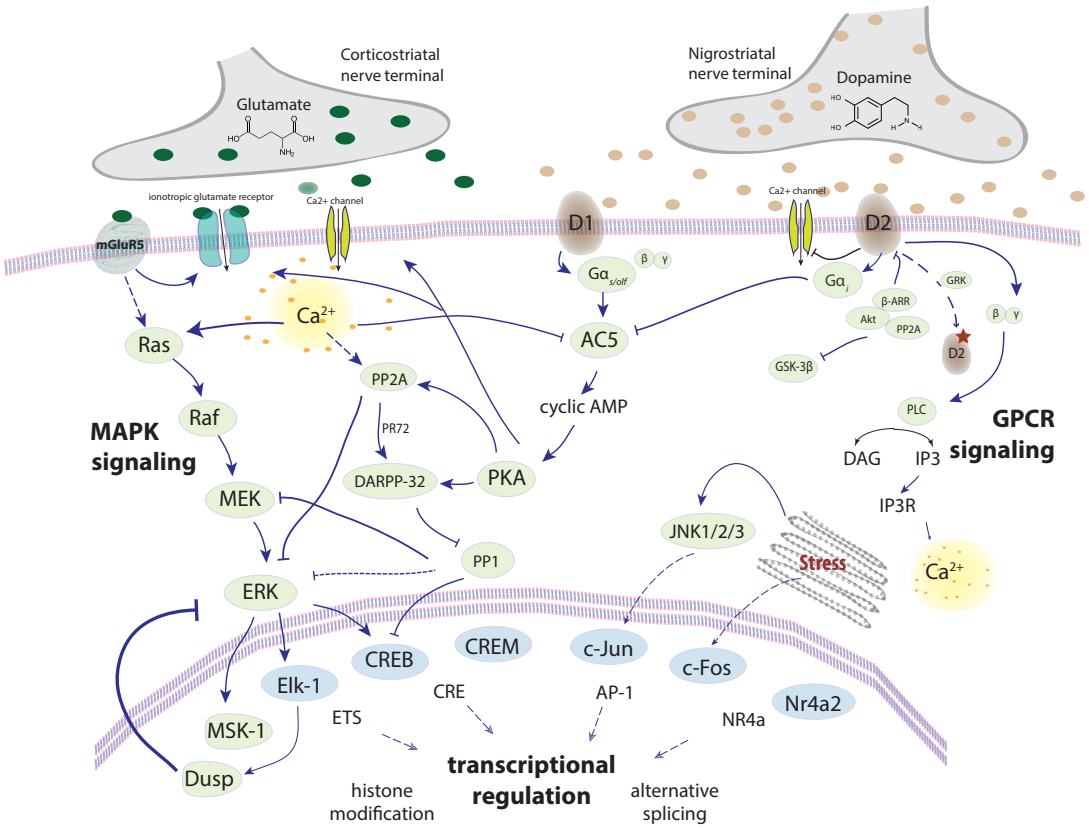


Figure 4.2: Summary of major pathways downstream of D1 and D2 dopamine receptors in striatal neurons. Adapted from Cenci, 2010 [18] with input from M. Heiman

While levodopa leads to increases in dopamine itself, which then acts on all of its physiological targets, a variety of dopamine agonists have been developed, each of which has distinct affinities for the various dopamine receptors. The are often used as adjuncts to levodopa and carbidopa to maximize 'on time' while reducing levodopa doses and side effects. The dopamine agonists used for PD, such as Cabergoline, Ropinirole and Pramiprexole, have higher affinity for the D_2 (and D_3 and D_4) receptors than for D_1.

4.1.6 Current Therapies and Therapeutic Targets in LID

Pharmacological factors contributing to LID and the approaches that have been explored to prevent or treat it are reviewed comprehensively by Schaeffer et al. [107]. While dopamine is the main player, many other neurotransmitter systems may also be potentially relevant to prevention and modulation of LID.

The primary strategy to prevent dyskinesia is simply to delay using L-DOPA, and then to use the smallest doses possible. Since the pulsatile changes in concentration produced by oral L-DOPA is thought to be one of the key factors contributing to LID, there has also been work to develop alternative formulations, such as intra-intestinal infusions, to provide more stable levels of dopamine.

Development of more selective dopamine receptor agonists has both provided tools to better manage L-DOPA therapies and revealed relationships between dyskinesia and the stimulaion of specific dopamine receptors (and the cell types that express them). The dopamine agonists typically used in PD mainly activate D_2 receptors, and tend to be less likely to cause LID. Whether the differences in dyskinesia development with these agonists is primarily due to differences in their kinetics or receptor selectivity is unclear; it has also been suggested that the function of D_3 receptors is important [14].

Many drugs have been evaluated for treatment of dyskinesias after they develop. Amantadine, an NMDA glutamate receptor antagonist, and clozapine, an atypical antipsychotic with both serotonin and dopamine receptor agonist activities are the the main drugs currently used to treat dyskinesia symptoms.

Among other strategies studied are NMDA receptor antagonists, mGluR antagonists, AMPA antagonists, anticonvulsants, 5HT agonists, α-adrenergic antagonists, opiod antag-

onists, endocannabinoid antagonists, adenosine A_{2A} antagonists, and nicotinic receptor antagonists, none of which have been established to reduce LID in the clinic [107].

Most of the preclinical studies of potential drugs for LID rely on either rat or primate models of Parkinsons and dyskinesia generated either by 6-OHDA or MPTP. Motor phenotypes are the primary endpoints evaluated, although recent studies have also included molecular measurements. For example, BN82451 / IPS-082451 is a compound originally investigated for its anti-oxidant activity, since oxidative stress is thought to be a common feature of many neurodegenerative diseases [21]. Further study revealed that it acts through multiple targets and mechanisms: it blocks neuronal Na^+ channels, reducing glutamate release and thus reducing excitotoxicity; it inhibits cyclooxygenase activity, reducing inflammation; it provides general protection against oxidative stress; and protects against toxicity caused by mitochondrial dysfunction. Interestingly, IPS-082451 was shown to ameliorate levodopa-induced dyskinesias in both rat [120] and primate models [6], and it reversed LID-associated up-regulation of cFos, FosB, Arc, and Nur77/Nr4A1, but did not have an effect on GAD67, Homer, PDyn, and PPE genes which are also up-regulated in LID.

4.1.7 TRAP

To enable measurement of gene expression from specific cell types, Heiman et al. developed Translating Ribosome Affinity Purification (TRAP) [49]. In TRAP, transgenic mice express an EGFP-tagged copy of the of the L10a large ribosomal subunit under the control of a cell-type specific promoter. This allows the tagged ribosomes from the targeted cell type to be affinity-purified using an antibody to EGFP, and any mRNA molecules that were being translated are brought along with the ribosomes. This population of mRNAs can then be purified and characterized using microarrays or RNA sequencing.

4.2 Heiman TRAP LID study

The pathophysiology of both Parkinson's and LID are are influenced by how the distinct motor pathways and cell types in the striatum are changed. Since earlier studies have meaured gene expression changes from whole tissue, it has been difficult to untangle

precisely how different cell types are affected, and when cells are homogenized, expression changes in one cell type could be obscured by opposing changes in other cell types. Using the TRAP technology in combination with a mouse model of PD and LID enables much more precise definition of changes in the relevant cell types.

4.3 Experimental Design and Data

To allow comparisions between expression changes in D1 to D2 MSNs, TRAP mice expressing eGFP-taggged ribosomes under control of either the Drd1a or the Drd2 promoter were studied. In these mice, the Parkinson's model of dopamine deficiency is first induced by unilateral injection of 6-OHDA into the MFB, and compared to a mock lesion in which mice were subjected to the same surgery, but with injection of ascorbate. This allowed assessment of the transcriptional changes due to dopamine depletion alone, modeling the Parkinsonian condition. Since L-DOPA does not produce a significant phenotype in non-Parkinsonian mice, only the 6-OHDA (dopamine depleted) mice were then subjected to levodopa treatment. As levodopa dose is a critical factor influencing development of dyskinesia, both low-dose and high-dose L-DOPA regimens were applied, and compared to a saline control. Since L-DOPA is likely to have acute effects on striatal neuron functions, in addition to the the chronic effects which are though to be more relevant to emergence of dyskinesia, transcriptional changes were also assessed following acute L-DOPA administration in the D1 arm of the study. Dykinesia phenotype was measured by counting abnormal involuntary movements to derive an AIM score, based on a previously validated scale [39]. Each mouse was then sacrificed so that a profile of gene expression in the targeted cell type could be assayed using the Affymetrix 430 2.0 microarray. Figure 4.3 summarizes the experimental groups and the number of subjects in each.

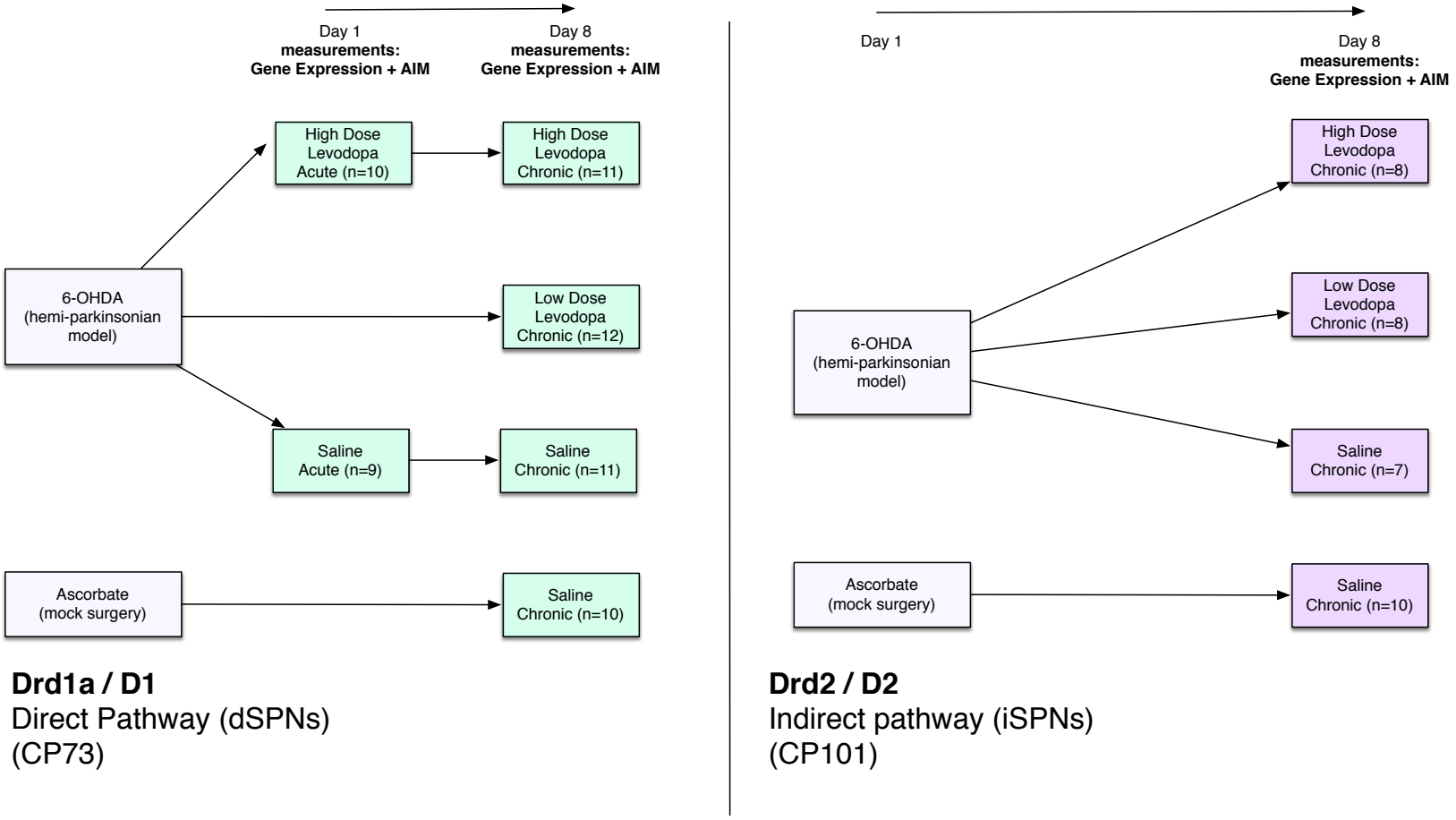


Figure 4.3: Outline of Heiman LID study experimental design

4.4 Methods

4.4.1 Differential expression analysis

TRAP-purified mRNAs from either Drd1a- or Drd2-expressing SPNs were reverse-transcribed, amplified, and used to interrogate Affymetrix 430 2.0 GeneChip microarrays. Affymetrix CEL files were processed and normalized using the RMA algorithm from the Bioconductor "affy" package [40]. For each (Dose, Cell Type) group, log2 fold-change for each probe-set was computed as the difference in mean expression compared with the matched saline-treated group. Significance of differences between groups was calculated by Welch's t test using scipy.stats or R [98]. To report counts for comparisons between groups, we defined significantly differentially expressed genes as those having any probe-set with greater than 1.5-fold change and a Benjamini–Hochberg adjusted P value from Welch's t test < 0.10. Source code and data files to replicate all statistical analyses are provided on the Web site http://pd.sciencespace.org and at http://github.com/aheilbut/PDmouse. Dataset S20 contains all statistical results for all probe-sets, and Table 4.26 provides links to complete files with all data tables discussed.

4.4.2 Linear modeling of AIM scores from L-DOPA dose and expression

Since there is variability in the timing and severity of dyskinesias both in the clinic and in these mouse models, one of the initial questions considered was whether there were genes associated specifically with the emergence of dyskinesia, distinct from other expression changes associated with L-DOPA treatment but which might not be directly related to dyskineia. To test the hypothesis that differences in gene expression may be correlated to variation in AIMs severity, we considered two sets of nested linear models relating expression of each probe-set, L-DOPA dose, and AIM score: $AIM \sim Expression + C(Dose)$, $AIM \sim C(Dose)$, and $AIM \sim Expression$, as well as $Expression \sim AIM + C(Dose)$, $Expression \sim C(Dose)$, and $Expression \sim AIM$. $C(Dose)$ refers to the factor variable representing high- or low-dose levodopa treatment. Models were fit using the "ols" procedure in the python statsmodels module [108]. Comparing these models allowed assessment of whether expression was correlated with AIM score, and whether that cor-

relation was more than would have been expected given the common dependence of dyskinesia and expression on levodopa dose. This process distinguishes three possible sets of genes: (i) dose-dependent genes with the expected correlation with dyskinesia severity (i.e., significant differential expression across dose, and significant association of AIM score and dose, but nonsignificant associa- tion of AIM score and expression, adjusting for dose); (ii) dose- dependent genes with excess correlation with dyskinesia (i.e., as in i, but with significant association of AIM score and expression, adjusting for dose); and (iii) genes with expression independent of dose yet correlated with dyskinesia (i.e., as in ii, but without significant differential expression between doses). Fig. **??** shows theoretical examples of each of these types of possible probe- sets. Dataset S16 reports statistics for all model fits and comparisons, to enable comparisons among models and sorting probe-sets by correlations with AIMs, statistical significance, or magnitudes of expression changes. Probe-sets are sorted by the significance of the multiple correlation for the model $Expression \sim C(Dose) + AIM$, after filtering for significant changes of 1.5-fold or greater between the high- and low-dose groups.

4.4.3 Pathways Overlap Analysis

For each treatment group, the set of statistically significant differentially expressed genes (Benjamini–Hochberg FDR, cut-off of 0.10), independent of magnitude of change, was compared against the Wikipathways gene sets to compute over- laps. Statistical significance of gene set overlaps was assessed by a hypergeometric test.

4.4.4 Multiple Hypothesis Testing Adjustment

P values from all statistical tests were adjusted using the Benjamini–Hochberg procedure with "multicomp.multipletests" in python statsmodels [108] to control false-discovery rate over all probe-sets. Bonferroni-adjusted and nominal P values are also reported.

Full details on experimental methods and biological reagents are provided in the text and supplement of Heiman, 2014. [51]

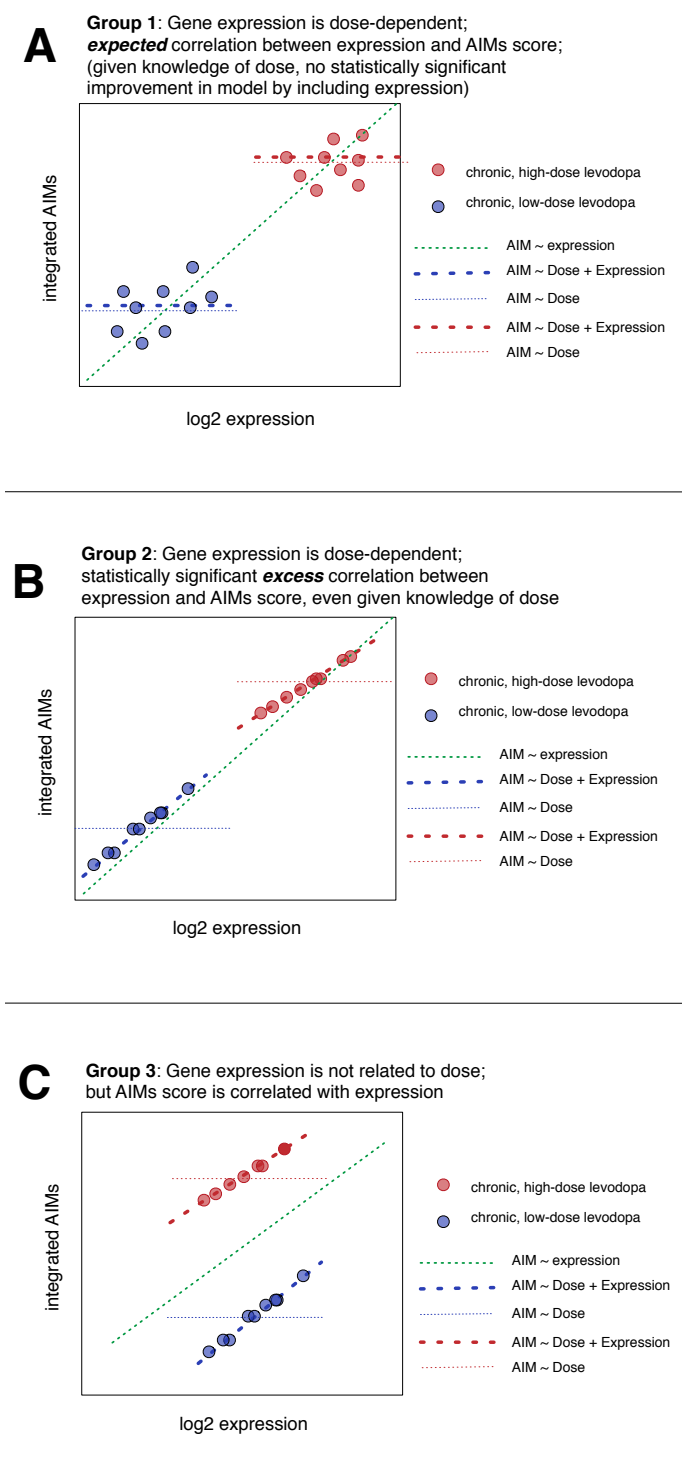


Figure 4.4: Hypothetical examples of genes with different relationships to AIM scores that are distinguished by comparisons between alternative linear models. Each panel shows a scatter plot of AIM score vs. log2 expression. Each point represents one gene in one mouse; its expression is encoded by horizontal position, and the integrated AIM score of that mouse is encoded by vertical position. Blue points represent measurements from mice treated with low-dose levodopa; red points are from mice treated with the high dose of the drug. (A) An example of a gene in group 1. The expression of this gene is dependent on dose. Fine dotted lines show the predicted AIM score for a model of AIM score as a function of dose (blue for low-dose and red for high dose). If the AIM score is modeled as a function of gene expression alone (AIM ~ Expression), there is a significant correlation between AIM score and expression (green dotted line). Red and blue large dashed lines depict AIM score prediction from a model using both dose and expression (AIM ~ Dose + Expression). For group 1, there is no significant difference between the dashed and dotted lines; conditional on knowledge of dose, AIM score is not correlated with expression. (B) A hypothetical gene in group 2. For this gene, there is a significant difference between the dashed and dotted lines; the model (AIM ~ Dose + Expression) fits significantly better than (AIM ~ Dose). (C) A hypothetical gene in group 3, for which AIM score is not well modeled by gene expression alone (AIM ~ expression), because expression measurements overlap between the dose groups. However, there is again a significant difference between (AIM ~ Dose + Expression) and (AIM ~ Dose) models.

4.5 Results

As expected from earlier studies, Drd1a MSNs were found to have many more significant expression changes than do Drd2 cells following L-DOPA treatment. Figure 4.5 shows a graphical summary of statistically significant changes over all of the experimental contrasts involving chronic L-DOPA treatment.

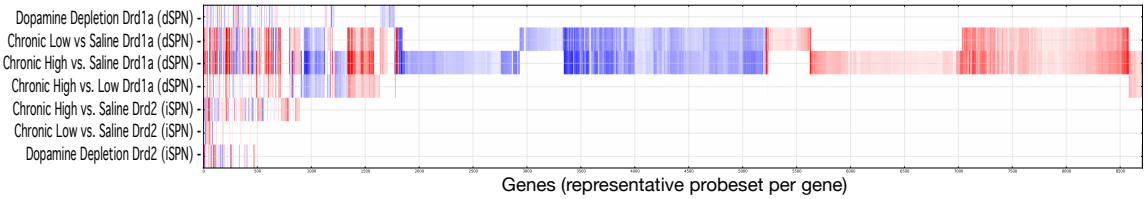


Figure 4.5: Genome-wide heatmap showing statistically significant expression changes over experimental contrasts. Drd1a cells exhibit many more changes than Drd2a cells.

4.5.1 Effects of Striatal Dopamine Depletion on SPNs

In D1 dSPNs, there were 226 genes, represented by 291 probesets, that were differentially expressed following striatal dopamine depletion, while 156 genes (196 probesets) were changed in the D2 iSPNs. The probesets with the largest changes (greater than 2-fold up- or down-regulated) are listed in Tables 4.1 and 4.2 respectively. To understand the major biological processes affected by these changes, we performed an overlap analysis against gene sets defined by pathways in the the Wikipathways database.

In D1 dSPNs, biological pathways with significant changes upon dopamine depletion (Table 4.5) included IL-3 signaling, the EGFR1 signaling pathway, and regulators of the MAPK signaling pathway. Among genes associated with MAPK signaling are numerous *Dusp* family (Dual specificity phosphatase) genes, which are downregulated. *Dusp*s are normally negative regulators of MAPK signaling, so their downregulation may lead to disinhibition and supersensitivity of the ERK pathways that respond to dopamine signaling in dSPNs.

In D2 iSPNs, pathways associated with genes that were changed on dopamine depletion included TGF-β and G-protein signaling pathways 4.6.

Only 22 genes have significant changes in both dSPNs and iSPNs (Table 4.7). Most,

Probeset	Gene Symbol	Gene Description	P-value	log2 FC
1438967_x_at	Amhr2	anti-Mullerian hormone type 2 receptor	$2.09 \cdot 10^{-3}$	3.65
1457021_x_at	Amhr2	anti-Mullerian hormone type 2 receptor	$1.63 \cdot 10^{-4}$	2.35
1437226_x_at	Marcksl1	MARCKS-like 1	$8.6 \cdot 10^{-3}$	1.69
1452473_at	Prr15	proline rich 15	$7.26 \cdot 10^{-2}$	1.58
1438852_x_at	Mcm6	minichromosome maintenance deficient 6 (MIS5 homolog, S. pombe) (S. cerevisiae)	$2.59 \cdot 10^{-2}$	1.49
1434458_at	Fst	follistatin	$8.86 \cdot 10^{-2}$	1.41
1441306_at	6820408C15Rik	RIKEN cDNA 6820408C15 gene	$1.06 \cdot 10^{-2}$	1.29
1421365_at	Fst	follistatin	$5.62 \cdot 10^{-2}$	1.25
1415922_s_at	Marcksl1	MARCKS-like 1	$2.54 \cdot 10^{-2}$	1.24
1436919_at	Trp53i11	transformation related protein 53 inducible protein 11	$8.95 \cdot 10^{-3}$	1.22
1442180_at	Dleu7	deleted in lymphocytic leukemia, 7	$1.24 \cdot 10^{-2}$	1.22
1416406_at	Pea15a	phosphoprotein enriched in astrocytes 15A	$1.52 \cdot 10^{-4}$	1.21
			$5.45 \cdot 10^{-3}$	1.19
1429372_at	Sox11	SRY-box containing gene 11	$4.51 \cdot 10^{-3}$	1.19
1434436_at	Morc4	microrchidia 4	$6.27 \cdot 10^{-2}$	1.18
1455324_at	Plcxd2	phosphatidylinositol-specific phospholipase C, X domain containing 2	$2.43 \cdot 10^{-2}$	1.16
1436790_a_at	Sox11	SRY-box containing gene 11	$1.14 \cdot 10^{-2}$	1.15
1422130_at	Nptx1	neuronal pentraxin 1	$4.31 \cdot 10^{-2}$	1.13
1422662_at	Lgals8	lectin, galactose binding, soluble 8	$2.71 \cdot 10^{-3}$	1.13
1450724_at	Fam126a	family with sequence similarity 126, member A	$9.57 \cdot 10^{-4}$	1.13
1418726_a_at	Tnnt2	troponin T2, cardiac	$2.09 \cdot 10^{-3}$	1.11
1424967_x_at	Tnnt2	troponin T2, cardiac	$6.1 \cdot 10^{-4}$	1.11
1422661_at	Lgals8	lectin, galactose binding, soluble 8	$1.16 \cdot 10^{-3}$	1.1
			$5.33 \cdot 10^{-2}$	1.1
1435627_x_at	Marcksl1	MARCKS-like 1	$1.11 \cdot 10^{-2}$	1.09
1455628_at	Epb4.1l4b	erythrocyte protein band 4.1-like 4b	$2.35 \cdot 10^{-2}$	1.09
1418369_at	Prim1	DNA primase, p49 subunit	$1.45 \cdot 10^{-2}$	1.08
1453002_at	Sox11	SRY-box containing gene 11	$2.6 \cdot 10^{-2}$	1.08
1416410_at	Pafah1b3	platelet-activating factor acetylhydrolase, isoform 1b, subunit 3	$6.26 \cdot 10^{-3}$	1.07
1453125_at	Sox11	SRY-box containing gene 11	$5.63 \cdot 10^{-2}$	1.05
1416407_at	Pea15a	phosphoprotein enriched in astrocytes 15A	$2.43 \cdot 10^{-4}$	1.04

Table 4.1: Probe sets up-regulated > 2-fold upon dopamine depletion in D1 dSPNs

though not all of these changes were in opposing directions between the two cell types. The relatively small number of opposing changes suggests that the cell type specific responses are a product more of intrinsic differences in the cell types, rather than simply being due to the classical opposition between the $G_{\alpha/olf}$ and G_i pathways that interact with Drd1a and Drd2 receptors.

4.5.2 Effects of Levodopa Treatment on Dopamine-depleted SPNs

When dopamine-depleted hemiparkinsonian mice are treated with a low dose of levodopa, forelimb use on the affected side can be restored. After chronic levodopa treatment at a higher dose, however, these mice develop abnormal involuntary movements (AIMs) affecting axial, orofacial and limb muscles. Notably, in the mouse model, almost all mice do exhibit AIMs after high-dose levodopa treatment, which is different from some reports of rat models in which only some of levodopa-treated rats develop dyskinesias; it is unclear whether this reflects intrinsic biological differences between the models or simply a differences in the effective levodopa doses studied.

cc**Probeset**	**Gene Symbol**	**Gene Description**	**P-value**	fc
1443722_at			0.080227716	-1.004929565
1420860_at	Itga9	integrin alpha 9	0.014470969	-1.01259148
1417602_at	Per2	period circadian clock 2	0.017863723	-1.020555574
1420998_at	Etv5	ets variant gene 5	0.018811118	-1.025240512
1451705_a_at	Oprm1	opioid receptor, mu 1	0.028843744	-1.046919705
1458413_at	Fbxw8	F-box and WD-40 domain protein 8	0.044367617	-1.049376773
1429072_at	Col6a4	collagen, type VI, alpha 4	0.066905288	-1.059686436
1426721_s_at	Tiparp	TCDD-inducible poly(ADP-ribose) polymerase	0.012875622	-1.061443354
1455956_x_at	Ccnd2	cyclin D2	0.006889739	-1.063499724
1438672_at	Parvb	parvin, beta	0.085468055	-1.066935787
1420654_a_at	Gbe1	glucan (1,4-alpha-), branching enzyme 1	0.008596545	-1.074902655
1454884_at	Zbtb46	zinc finger and BTB domain containing 46	0.018117194	-1.078537366
1419606_a_at	Tnnt1	troponin T1, skeletal, slow	0.069373426	-1.079877924
1416700_at	Rnd3	Rho family GTPase 3	0.069847635	-1.083252055
1425608_at	Dusp3	dual specificity phosphatase 3 (vaccinia virus phosphatase VH1-related)	0.000543978	-1.092687546
1419144_at	Cd163	CD163 antigen	0.020635694	-1.094528415
1459941_at	Clvs1	clavesin 1	0.060311638	-1.103306147
1449484_at	Stc2	stanniocalcin 2	0.093986755	-1.110465455
1453334_at	B230216N24Rik	RIKEN cDNA B230216N24 gene	0.0240589	-1.118632232
1420462_at	Il1rapl2	interleukin 1 receptor accessory protein-like 2	0.060311638	-1.118771412
1429952_at	Mospd4	motile sperm domain containing 4	0.011228591	-1.133904911
1416123_at	Ccnd2	cyclin D2	0.000174909	-1.135277657
1450082_s_at	Etv5	ets variant gene 5	0.046103495	-1.135690637
1430332_a_at	Gusb	glucuronidase, beta	0.056932472	-1.135741667
1456280_at	Clspn	claspin	0.034930075	-1.135761804
1448754_at	Rbp1	retinol binding protein 1, cellular	0.007469616	-1.13643762
1449133_at	Sprr1a	small proline-rich protein 1A	0.032054256	-1.139575307
1437950_at	Fam149a	family with sequence similarity 149, member A	0.048940099	-1.146935888
1436405_at	Dock4	dedicator of cytokinesis 4	0.004212193	-1.177432288
1416122_at	Ccnd2	cyclin D2	0.00951358	-1.182960456
1421979_at	Phex	phosphate regulating gene with homologies to endopeptidases on the X chromosome (hypophosphatemia, vitamin D resistant rickets)	0.006889739	-1.18405527
1450212_at	Fmnl1	formin-like 1	0.005802704	-1.19325141
1435852_at	Spred3	sprouty-related, EVH1 domain containing 3	0.025986042	-1.227385994
1449374_at	Pipox	pipecolic acid oxidase	0.034102497	-1.228193555
1454256_s_at			0.016786616	-1.228486073
1423606_at	Postn	periostin, osteoblast specific factor	0.021082199	-1.236529732
1429637_at	Fam198b	family with sequence similarity 198, member B	0.019229388	-1.269591877
1449584_at	Dgkg	diacylglycerol kinase, gamma	0.021082199	-1.273989947
1449188_at	Midn	midnolin	0.002258877	-1.279165662
1426210_x_at	Parp3	poly (ADP-ribose) polymerase family, member 3	0.010683401	-1.279187142
1450029_s_at	Itga9	integrin alpha 9	0.072565857	-1.285791837
1441914_x_at	Fgf3	fibroblast growth factor 3	0.029209964	-1.302902536
1443888_at	AU023762	expressed sequence AU023762	0.006889739	-1.303651692
1416805_at	Fam198b	family with sequence similarity 198, member B	0.001013089	-1.304524411
1450445_at	Phex	phosphate regulating gene with homologies to endopeptidases on the X chromosome (hypophosphatemia, vitamin D resistant rickets)	0.033923248	-1.330576397
1415834_at	Dusp6	dual specificity phosphatase 6	0.001013089	-1.333785975
1451969_s_at	Parp3	poly (ADP-ribose) polymerase family, member 3	0.066752042	-1.334633606
1428142_at	Etv5	ets variant gene 5	0.001013089	-1.346831722
1423505_at	Tagln	transgelin	0.028843744	-1.361606086
1423506_a_at	Nnat	neuronatin	0.004746895	-1.37884082
1449519_at	Gadd45a	growth arrest and DNA-damage-inducible 45 alpha	0.027034826	-1.399956207
1439985_at	Abcc12	ATP-binding cassette, sub-family C (CFTR/MRP), member 12	0.001538811	-1.40523437
1455760_at	Slc9a5	solute carrier family 9 (sodium/hydrogen exchanger), member 5	0.004041789	-1.467262759
1430127_a_at	Ccnd2	cyclin D2	0.007117806	-1.471328167
1438796_at	Nr4a3	nuclear receptor subfamily 4, group A, member 3	0.047872093	-1.613834029
1442754_at	C030013G03Rik	RIKEN cDNA C030013G03 gene	0.059472728	-1.676542519
1427683_at	Egr2	early growth response 2	0.056172044	-1.739839538
1426278_at	Ifi27l2a	interferon, alpha-inducible protein 27 like 2A	0.000223113	-1.874592417
1432073_at			0.027627299	-1.948721277
1427682_a_at	Egr2	early growth response 2	0.015739617	-1.958022672
1436094_at	Vgf	VGF nerve growth factor inducible	0.000338438	-2.137487929
1455275_at	E530001K10Rik	RIKEN cDNA E530001K10 gene	0.00259659	-2.167667125

Table 4.2: Probe sets down-regulated > 2-fold upon dopamine depletion in D1 dSPNs

Probeset	Gene Symbol	Gene Description	P-value	log2 FC
1460330_at	Anxa3	annexin A3	9.067E-05	4.236560507
1422825_at	Cartpt	CART prepropeptide	0.001082682	3.890656887
1442754_at	C030013G03Rik	RIKEN cDNA C030013G03 gene	0.004014201	3.273084675
1416505_at	Nr4a1	nuclear receptor subfamily 4, group A, member 1	0.000590965	2.827500676
1422860_at	Nts	neurotensin	1.0186E-05	2.59660318
1420720_at	Nptx2	neuronal pentraxin 2	0.000640343	2.59021096
1423100_at	Fos	FBJ osteosarcoma oncogene	0.073043831	2.426296369
1434528_at	Aard	alanine and arginine rich domain containing protein	0.029706557	2.238873825
1459145_at	A930033H14Rik	RIKEN cDNA A930033H14 gene	0.090322637	2.194986711
1450347_at	Syt10	synaptotagmin X	0.083831393	1.90848127
1460043_at			0.033272303	1.908017505
1450708_at	Scg2	secretogranin II	0.000359282	1.89818906
1437247_at	Fosl2	fos-like antigen 2	0.01223872	1.819254844
1419592_at	Unc5c	unc-5 homolog C (C. elegans)	0.055733024	1.693009328
1436094_at	Vgf	VGF nerve growth factor inducible	2.16893E-05	1.653749189
1423851_a_at	Shisa2	shisa homolog 2 (Xenopus laevis)	0.004128802	1.652623308
1450117_at	Tcf7l1	transcription factor 7 like 1 (T cell specific, HMG box)	0.028776312	1.649445606
1434243_s_at	Tomm70a	translocase of outer mitochondrial membrane 70 homolog A (yeast)	0.000590965	1.61797314
1419647_a_at	Ier3	immediate early response 3	0.004268132	1.605653794
1451342_at	Spon1	spondin 1, (f-spondin) extracellular matrix protein	0.026420145	1.560673332
1422053_at	Inhba	inhibin beta-A	0.027196734	1.554124256
1454256_s_at			0.078707444	1.541844266
1425110_at	Sorcs3	sortilin-related VPS10 domain containing receptor 3	0.001082682	1.493188138
1437841_x_at	Csdc2	cold shock domain containing C2, RNA binding	0.033372134	1.4831511
1434877_at	Nptx1	neuronal pentraxin 1	0.001802109	1.476337609
1418817_at	Chmp1b	charged multivesicular body protein 1B	0.033982044	1.467821129
1417018_at	Efemp2	epidermal growth factor-containing fibulin-like extracellular matrix protein 2	0.064099936	1.467137043
1435917_at	Ociad2	OCIA domain containing 2	0.061775407	1.46679379
1426225_at	Rbp4	retinol binding protein 4, plasma	0.034849593	1.442362381
1423852_at	Shisa2	shisa homolog 2 (Xenopus laevis)	0.013124915	1.427286112
1449037_at	Crem	cAMP responsive element modulator	0.024174349	1.412913365
1429643_a_at	Pde1c	phosphodiesterase 1C	0.012755763	1.403677843
1452729_at	Dpm3	dolichyl-phosphate mannosyltransferase polypeptide 3	0.01223872	1.394366092
1422256_at	Sstr2	somatostatin receptor 2	0.067651577	1.378763774
1416701_at	Rnd3	Rho family GTPase 3	0.041834303	1.371417518
1453387_at	4833432E10Rik	RIKEN cDNA 4833432E10 gene	0.089660297	1.361192572
1449286_at	Ntng1	netrin G1	0.029706557	1.357176164
1426036_a_at	Pde1c	phosphodiesterase 1C	0.024839636	1.329982363
1424831_at	Cpne2	copine II	0.064950324	1.307232065
1443523_at	Fam135b	family with sequence similarity 135, member B	0.008980017	1.282381322
1431422_a_at	Dusp14	dual specificity phosphatase 14	0.012782502	1.263071042
1447669_s_at	Gng4	guanine nucleotide binding protein (G protein), gamma 4	0.092661651	1.237418904
1456186_at	Prdm11	PR domain containing 11	0.092958072	1.236601239
1419425_at	Cnr1	cannabinoid receptor 1 (brain)	0.033372134	1.220632394
1443558_s_at	Nt5dc3	5′-nucleotidase domain containing 3	0.005072121	1.207482426
1459299_at	Myo3b	myosin IIIB	0.081406023	1.200579447
1440374_at	Pde1c	phosphodiesterase 1C	0.054528039	1.175234449
1422931_at	Fosl2	fos-like antigen 2	0.089660297	1.174816953
1435472_at	Kremen1	kringle containing transmembrane protein 1	0.017888879	1.164250654
1453187_at	Ociad2	OCIA domain containing 2	0.000318085	1.158705508
1435628_x_at			0.017888879	1.150086915
1441894_s_at	Grasp	GRP1 (general receptor for phosphoinositides 1)-associated scaffold protein	0.060058661	1.128127817
1441728_at	Scn1a	sodium channel, voltage-gated, type I, alpha	0.062577766	1.124587057
1435621_at	Far2	fatty acyl CoA reductase 2	0.030661148	1.11686058
1417192_at	Tomm70a	translocase of outer mitochondrial membrane 70 homolog A (yeast)	4.47069E-05	1.099563773
1425608_at	Dusp3	dual specificity phosphatase 3 (vaccinia virus phosphatase VH1-related)	0.017888879	1.086737514
1440147_at	Lgi2	leucine-rich repeat LGI family, member 2	0.000318085	1.06240188
1429714_at	Sumf2	sulfatase modifying factor 2	0.068869598	1.054498875
1426106_a_at	Syt6	synaptotagmin VI	0.005289356	1.054310344
1423285_at	Coch	coagulation factor C homolog (Limulus polyphemus)	0.060310752	1.054209852
1438796_at	Nr4a3	nuclear receptor subfamily 4, group A, member 3	0.091163869	1.039769818
1431253_s_at	Tbc1d9	TBC1 domain family, member 9	0.035751362	1.037413877
1457040_at	Lgi2	leucine-rich repeat LGI family, member 2	0.052041173	1.029388582
1427057_at	Nt5dc3	5′-nucleotidase domain containing 3	0.004268132	1.022546826
1423985_at			0.00160732	1.011936336
1435598_at	BB319198	expressed sequence BB319198	0.083593179	1.010679529
1417944_at	Gng4	guanine nucleotide binding protein (G protein), gamma 4	0.068602361	1.010314872
1423786_at	8430410A17Rik	RIKEN cDNA 8430410A17 gene	0.029706557	1.005074113

Table 4.3: Probe sets up-regulated > 2-fold upon dopamine depletion in D2 iSPNs

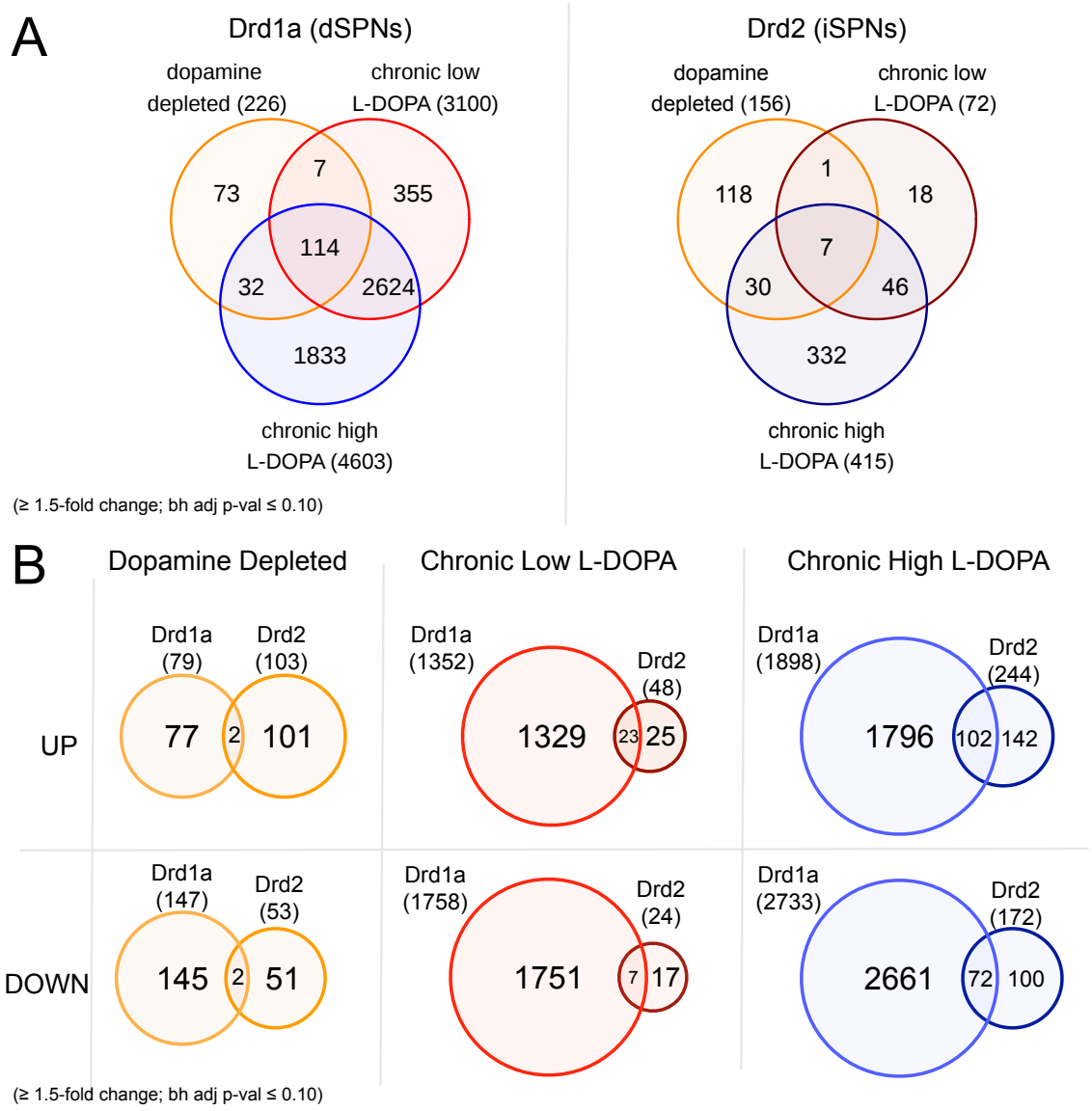


Figure 4.6: Genome-wide analysis of gene expression changes induced by dopamine depletion and levodopa treatment. (A) Venn diagrams showing the total numbers of genes changing across treatments in Drd1a (dSPN) cells (Left) and Drd2 (iSPN) cells (Right) for statistically significant changes (Benjamini– Hochberg adjusted P value < 0.10) of 1.5-fold or greater. (B) Venn diagrams comparing the numbers of genes up-regulated and down-regulated by each treatment between Drd1a (dSPN) and Drd2a (iSPN) cells for statistically significant changes (Benjamini–Hochberg adjusted P value < 0.10) of 1.5-fold or greater.

Probeset	Gene Symbol	Gene Description	P-value	log2 FC
1452135_at	Gpx6	glutathione peroxidase 6	$3.24 \cdot 10^{-2}$	-2.6
1438200_at	Sulf1	sulfatase 1	$4.18 \cdot 10^{-2}$	-2.22
1455753_at	Fam163b	family with sequence similarity 163, member B	$1.31 \cdot 10^{-2}$	-1.96
1456642_x_at	S100a10	S100 calcium binding protein A10 (calpactin)	$3.96 \cdot 10^{-2}$	-1.93
1457437_at	Fam163b	family with sequence similarity 163, member B	$3.16 \cdot 10^{-2}$	-1.92
1426065_a_at	Trib3	tribbles homolog 3 (Drosophila)	$7.87 \cdot 10^{-2}$	-1.62
1418726_a_at	Tnnt2	troponin T2, cardiac	$6.41 \cdot 10^{-2}$	-1.58
1436493_at	Ctxn2	cortexin 2	$4.53 \cdot 10^{-2}$	-1.48
1423836_at	Zfp503	zinc finger protein 503	$4.55 \cdot 10^{-2}$	-1.44
1416762_at	S100a10	S100 calcium binding protein A10 (calpactin)	$7.83 \cdot 10^{-3}$	-1.35
1416368_at	Gsta4	glutathione S-transferase, alpha 4	$3.07 \cdot 10^{-2}$	-1.35
1441302_at	LOC100502835	uncharacterized LOC100502835	$5.57 \cdot 10^{-2}$	-1.34
1419200_at	Fxyd7	FXYD domain-containing ion transport regulator 7	$1.93 \cdot 10^{-2}$	-1.31
1421992_a_at	Igfbp4	insulin-like growth factor binding protein 4	$3.68 \cdot 10^{-2}$	-1.31
1422821_s_at	Stard5	StAR-related lipid transfer (START) domain containing 5	$8.97 \cdot 10^{-2}$	-1.3
1438399_at	Pex5l	peroxisomal biogenesis factor 5-like	$8.73 \cdot 10^{-2}$	-1.29
1434868_at	4933431E20Rik	RIKEN cDNA 4933431E20 gene	$5.12 \cdot 10^{-2}$	-1.26
1448554_s_at			$3.16 \cdot 10^{-2}$	-1.24
1457052_at	Kcng1	potassium voltage-gated channel, subfamily G, member 1	$8.96 \cdot 10^{-2}$	-1.16
1449129_a_at	Kcnip3	Kv channel interacting protein 3, calsenilin	$6.4 \cdot 10^{-4}$	-1.13
1436193_at	Man1c1	mannosidase, alpha, class 1C, member 1	$6.26 \cdot 10^{-2}$	-1.13
1428221_at	Klhdc8b	kelch domain containing 8B	$1.52 \cdot 10^{-2}$	-1.1
1424534_at	Mmd2	monocyte to macrophage differentiation-associated 2	$2.35 \cdot 10^{-2}$	-1.09
1435307_at	Ankrd34b	ankyrin repeat domain 34B	$4.74 \cdot 10^{-2}$	-1.03
1427070_at	Snx21	sorting nexin family member 21	$8.63 \cdot 10^{-2}$	-1.02
1455961_at			$3.07 \cdot 10^{-2}$	-1.02
1434115_at	Cdh13	cadherin 13	$8.97 \cdot 10^{-2}$	-1.01

Table 4.4: Probe sets down-regulated > 2-fold upon dopamine depletion in D2 iSPNs

Effects of levodopa treatment in D1 dSPNs

In D1 dSPNs, over 3100 genes (4545 probe sets) had expression changes following dopamine depletion and low-dose levodopa treatment; 1,352 genes were up-regulated and 1,758 were down-regulated. Pathways represented in this set of genes are listed in Table 4.10. Many of the same pathways affected by dopamine depletion alone are further affected by levodopa. Notably, many of the Dusp genes that regulate ERK/MAPK signaling and were downregulated with dopamine depletion, have increased expression following chronic low-dose L-DOPA administration. This is presumably a homeostatic feedback response to increased ERK / MAPK signaling stimulated by L-DOPA (Figure 4.7).

With high-dose levodopa treatment, 4,603 genes (7,118 probe sets) have significant differential expression. 1,898 of theese genes were up-regulated, and 2,733 were down-regulated. 3,635 of the genes changed with high-dose levodopa also had statistically significant changes with the lower dose. The top 50 genes changing are listed in Table 4.11. Most of the pathways represented among genes changing with a high dose (Table 4.13) were the same as those observed using a low dose. One pathway that only appeared to be overrepresented at the higher dose was Regulation of Actin Cytoskeleton, which

Set Name	Matches	Genes in Overlap	Size of Set	p-value	Bonf adj p-val	b-h FDR adj p-val
Hypertrophy Model	5	NR4A3, DUSP14, ANKRD1, HBEGF, JUND	20	$4 \cdot 10^{-4}$	$4.12 \cdot 10^{-2}$	$1.84 \cdot 10^{-2}$
IL-3 Signaling Pathway	11	PTK2, BCL2L11, SOCS2, PPP2CA, PRKCB, MAP2K1, JAK1, KCNIP3, PAK1, HRAS1, BCL2	102	$5.57 \cdot 10^{-4}$	$5.74 \cdot 10^{-2}$	$1.84 \cdot 10^{-2}$
Electron Transport Chain	12	NDUFV3, NDUFA7, NDUFA2, NDUFA1, UQCR11, NDUFS2, NDUFS3, COX6B1, COX7A2, ATP5G1, ATP5E, ATP5D	119	$5.87 \cdot 10^{-4}$	$6.04 \cdot 10^{-2}$	$1.84 \cdot 10^{-2}$
EGFR1 Signaling Pathway	15	EPS15, RPS6KA3, APPL1, NDUFA13, JUND, PRKCB, SNCA, SPRY2, MAP2K1, ELK1, HRAS1, JAK1, DUSP1, PAK1, PTPN5	176	$7.8 \cdot 10^{-4}$	$8.03 \cdot 10^{-2}$	$1.84 \cdot 10^{-2}$
MAPK signaling pathway	14	RPS6KA3, STMN1, PPM1B, GADD45A, PRKCB, JUND, MAP2K1, ELK1, ECSIT, SRF, DUSP6, DUSP1, PAK1, PTPN5	160	$8.92 \cdot 10^{-4}$	$9.19 \cdot 10^{-2}$	$1.84 \cdot 10^{-2}$
Oxidative phosphorylation	8	NDUFV3, NDUFA7, NDUFA2, NDUFS2, NDUFS3, ATP5G1, ATP5E, ATP5D	69	$1.95 \cdot 10^{-3}$	0.2	$3.34 \cdot 10^{-2}$
ErbB signaling pathway	6	PTK2, HBEGF, MAP2K1, ELK1, NRG3, HRAS1	46	$3.9 \cdot 10^{-3}$	0.4	$5.73 \cdot 10^{-2}$
Signaling of Hepatocyte Growth Factor Receptor	5	ELK1, MAP2K1, PTK2, PAK1, HRAS1	34	$4.91 \cdot 10^{-3}$	0.51	$6.07 \cdot 10^{-2}$
estrogen signalling	8	TAF13, MAP2K1, ELK1, POLR2A, CREBBP, BCL2, HRAS1, POLR2J	81	$5.3 \cdot 10^{-3}$	0.55	$6.07 \cdot 10^{-2}$
Kit Receptor Signaling Pathway	7	SOCS5, PRKCB, SPRED1, SPRED2, MAP2K1, MITF, HRAS1	68	$7.11 \cdot 10^{-3}$	0.73	$6.71 \cdot 10^{-2}$
TGF Beta Signaling Pathway	6	CTNNB1, HRAS1, FST, SKIL, JAK1, CREBBP	52	$7.17 \cdot 10^{-3}$	0.74	$6.71 \cdot 10^{-2}$
B Cell Receptor Signaling Pathway	12	NFATC2, PRKCB, BCL2L11, ARPC3, CCND2, CTNNB1, MAP2K1, ELK1, HNRNPK, DUSP6, PTK2, BCL2	163	$8.12 \cdot 10^{-3}$	0.84	$6.74 \cdot 10^{-2}$
Myometrial Relaxation and Contraction Pathways	12	RGS14, RGS20, ATP2A2, CAMK2D, CAMK2G, RAMP1, PKIA, PRKCB, RGS4, RGS7, IGFBP6, RGS2	164	$8.51 \cdot 10^{-3}$	0.88	$6.74 \cdot 10^{-2}$
Diurnally regulated genes with circadian orthologs	6	IDI1, SUMO3, PER2, UGP2, ERC2, NCKAP1	55	$9.4 \cdot 10^{-3}$	0.97	$6.91 \cdot 10^{-2}$
Calcium Regulation in the Cardiac Cell	11	KCNJ3, RGS14, CAMK2G, ATP2A2, CAMK2D, PRKCB, PKIA, RGS4, RGS20, RGS2, RGS7	154	$1.36 \cdot 10^{-2}$	1	$8.81 \cdot 10^{-2}$
T Cell Receptor Signaling Pathway	10	DLG1, RASGRP2, NFATC2, PTK2, CTNNB1, MAP2K1, DUSP3, TUBA4A, PAK1, CREBBP	134	$1.37 \cdot 10^{-2}$	1	$8.81 \cdot 10^{-2}$
Circadian Exercise	6	IDI1, SUMO3, PER2, UGP2, ERC2, NCKAP1	61	$1.53 \cdot 10^{-2}$	1	$9.24 \cdot 10^{-2}$
G1 to S cell cycle control	6	GADD45A, CCND2, RPA3, MCM6, CCNG2, PRIM1	62	$1.64 \cdot 10^{-2}$	1	$9.4 \cdot 10^{-2}$
IL-7 Signaling Pathway	5	MAP2K1, CCND2, BCL2L11, HRAS1, JAK1	46	$1.75 \cdot 10^{-2}$	1	$9.48 \cdot 10^{-2}$

Table 4.5: Wikipathways pathways over-represented among genes changed upon dopamine depletion in D1 dSPNs

Set Name	Matches	Genes in Overlap	Size of Set	p-value	Bonf adj p-val	b-h FDR adj p-val
Myometrial Relaxation and Contraction Pathways	11	RGS4, PRKCB, RGS2, ATP2A2, FOS, GNAS, GUCY1A3, SFN, GNG4, IGFBP6, IGFBP4	164	$5.27 \cdot 10^{-6}$	$4.17 \cdot 10^{-4}$	$4.17 \cdot 10^{-4}$
TGF Beta Signaling Pathway	5	SMAD4, MAPK3, FOS, SKIL, INHBA	52	$4.07 \cdot 10^{-4}$	$3.22 \cdot 10^{-2}$	$1.55 \cdot 10^{-2}$
Calcium Regulation in the Cardiac Cell	8	RGS4, ATP2A2, GNAS, PRKCB, SFN, GNG4, RGS2, SLC8A3	154	$5.88 \cdot 10^{-4}$	$4.65 \cdot 10^{-2}$	$1.55 \cdot 10^{-2}$
Splicing factor NOVA regulated synpatic proteins	4	DAB1, CADM3, NTNG1, TERF2IP	42	$1.63 \cdot 10^{-3}$	0.13	$3.22 \cdot 10^{-2}$
One carbon metabolism and related pathways	4	AHCYL1, GAD1, GAD2, GPX6	45	$2.11 \cdot 10^{-3}$	0.17	$3.34 \cdot 10^{-2}$
Selenium metabolism/Selenoproteins	3	GPX6, CREM, FOS	26	$3.71 \cdot 10^{-3}$	0.29	$4.88 \cdot 10^{-2}$
G Protein Signaling Pathways	5	PDE1C, GNG4, PDE7B, PRKCB, GNAS	99	$7.09 \cdot 10^{-3}$	0.56	$8 \cdot 10^{-2}$
Alanine and aspartate metabolism	2	GAD1, GAD2	12	$8.89 \cdot 10^{-3}$	0.7	$8.17 \cdot 10^{-2}$
Kit Receptor Signaling Pathway	4	PRKCB, KITL, SPRED1, SPRED2	68	$9.31 \cdot 10^{-3}$	0.74	$8.17 \cdot 10^{-2}$
Homologous recombination	2	ATM, NBN	13	$1.04 \cdot 10^{-2}$	0.82	$8.23 \cdot 10^{-2}$
Biogenic Amine Synthesis	2	GAD1, GAD2	14	$1.21 \cdot 10^{-2}$	0.95	$8.66 \cdot 10^{-2}$

Table 4.6: Wikipathways pathways over-represented among genes changed upon dopamine depletion in D2 iSPNs

Probeset	Gene Symbol	Gene Description	D1 p-val	D1 log2 FC	D2 p-val	D2 log2 FC
1442754_at	C030013G03Rik	RIKEN cDNA C030013G03 gene	$5.95 \cdot 10^{-2}$	-1.68	$4.01 \cdot 10^{-3}$	3.27
1416123_at	Ccnd2	cyclin D2	$1.75 \cdot 10^{-4}$	-1.14	0.33	-0.78
1448229_s_at	Ccnd2	cyclin D2	$1.8 \cdot 10^{-3}$	-0.94	0.12	-0.68
1455956_x_at	Ccnd2	cyclin D2	$6.89 \cdot 10^{-3}$	-1.06	$3.07 \cdot 10^{-2}$	-0.83
1430127_a_at	Ccnd2	cyclin D2	$7.12 \cdot 10^{-3}$	-1.47	0.72	-0.32
1416122_at	Ccnd2	cyclin D2	$9.51 \cdot 10^{-3}$	-1.18	0.15	-0.84
1434745_at	Ccnd2	cyclin D2	$2.6 \cdot 10^{-2}$	-0.99	$6.41 \cdot 10^{-2}$	-0.77
1449037_at	Crem	cAMP responsive element modulator	$7.97 \cdot 10^{-2}$	-0.82	$2.42 \cdot 10^{-2}$	1.41
1437841_x_at	Csdc2	cold shock domain containing C2, RNA binding	$1.95 \cdot 10^{-2}$	0.58	$3.34 \cdot 10^{-2}$	1.48
1423845_at	Csdc2	cold shock domain containing C2, RNA binding	$3.59 \cdot 10^{-2}$	0.76	0.35	0.78
1451147_x_at	Csdc2	cold shock domain containing C2, RNA binding	$3.99 \cdot 10^{-2}$	0.7	0.34	0.8
1431422_a_at	Dusp14	dual specificity phosphatase 14	$5.33 \cdot 10^{-2}$	-0.81	$1.28 \cdot 10^{-2}$	1.26
1434472_at	Dusp3	dual specificity phosphatase 3 (vaccinia virus phosphatase VH1-related)	$1.75 \cdot 10^{-4}$	-0.65	$3.72 \cdot 10^{-3}$	0.57
1425608_at	Dusp3	dual specificity phosphatase 3 (vaccinia virus phosphatase VH1-related)	$5.44 \cdot 10^{-4}$	-1.09	$1.79 \cdot 10^{-2}$	1.09
1456769_at	Dusp3	dual specificity phosphatase 3 (vaccinia virus phosphatase VH1-related)	0.11	-0.64	$3.86 \cdot 10^{-2}$	0.85
1448807_at	Hrh3	histamine receptor H3	$1.57 \cdot 10^{-2}$	0.84	$8.38 \cdot 10^{-2}$	-0.9
1417933_at	Igfbp6	insulin-like growth factor binding protein 6	$1.33 \cdot 10^{-2}$	-0.76	$3.16 \cdot 10^{-2}$	1
1457052_at	Kcng1	potassium voltage-gated channel, subfamily G, member 1	$3.2 \cdot 10^{-2}$	-0.69	$8.96 \cdot 10^{-2}$	-1.16
1423506_a_at	Nnat	neuronatin	$4.75 \cdot 10^{-3}$	-1.38	$8.98 \cdot 10^{-3}$	0.9
1422130_at	Nptx1	neuronal pentraxin 1	$4.31 \cdot 10^{-2}$	1.13	0.28	1.26
1434877_at	Nptx1	neuronal pentraxin 1	$7.63 \cdot 10^{-2}$	0.73	$1.8 \cdot 10^{-3}$	1.48
1438796_at	Nr4a3	nuclear receptor subfamily 4, group A, member 3	$4.79 \cdot 10^{-2}$	-1.61	$9.12 \cdot 10^{-2}$	1.04
1421080_at	Nr4a3	nuclear receptor subfamily 4, group A, member 3	$9.25 \cdot 10^{-2}$	-0.76	0.39	0.73
1416406_at	Pea15a	phosphoprotein enriched in astrocytes 15A	$1.52 \cdot 10^{-4}$	1.21	0.55	-0.61
1416407_at	Pea15a	phosphoprotein enriched in astrocytes 15A	$2.43 \cdot 10^{-4}$	1.04	$4.12 \cdot 10^{-2}$	-0.7
1426225_at	Rbp4	retinol binding protein 4, plasma	$5.38 \cdot 10^{-2}$	-0.7	$3.48 \cdot 10^{-2}$	1.44
1416700_at	Rnd3	Rho family GTPase 3	$6.98 \cdot 10^{-2}$	-1.08	0.42	1.13
1416701_at	Rnd3	Rho family GTPase 3	0.21	-0.82	$4.18 \cdot 10^{-2}$	1.37
1450708_at	Scg2	secretogranin II	$2.4 \cdot 10^{-2}$	-0.75	$3.59 \cdot 10^{-4}$	1.9
1423851_a_at	Shisa2	shisa homolog 2 (Xenopus laevis)	$8.12 \cdot 10^{-2}$	-0.98	$4.13 \cdot 10^{-3}$	1.65
1423852_at	Shisa2	shisa homolog 2 (Xenopus laevis)	0.31	-0.73	$1.31 \cdot 10^{-2}$	1.43
1460439_at	Sik3	SIK family kinase 3	$7.08 \cdot 10^{-2}$	-0.71	$2.11 \cdot 10^{-2}$	0.78
1426106_a_at	Syt6	synaptotagmin VI	$6.89 \cdot 10^{-3}$	-0.84	$5.29 \cdot 10^{-3}$	1.05
1426721_s_at	Tiparp	TCDD-inducible poly(ADP-ribose) polymerase	$1.29 \cdot 10^{-2}$	-1.06	0.91	0.25
1452160_at	Tiparp	TCDD-inducible poly(ADP-ribose) polymerase	0.15	-0.64	$9.03 \cdot 10^{-2}$	0.99
1424967_x_at	Tnnt2	troponin T2, cardiac	$6.1 \cdot 10^{-4}$	1.11	0.11	-1.47
1418726_a_at	Tnnt2	troponin T2, cardiac	$2.09 \cdot 10^{-3}$	1.11	$6.41 \cdot 10^{-2}$	-1.58
1436094_at	Vgf	VGF nerve growth factor inducible	$3.38 \cdot 10^{-4}$	-2.14	$2.17 \cdot 10^{-5}$	1.65

Table 4.7: Genes with probe-sets changed > 2-fold (in any direction) upon dopamine depletion in *both* D1 dSPNs and D2 iSPNs

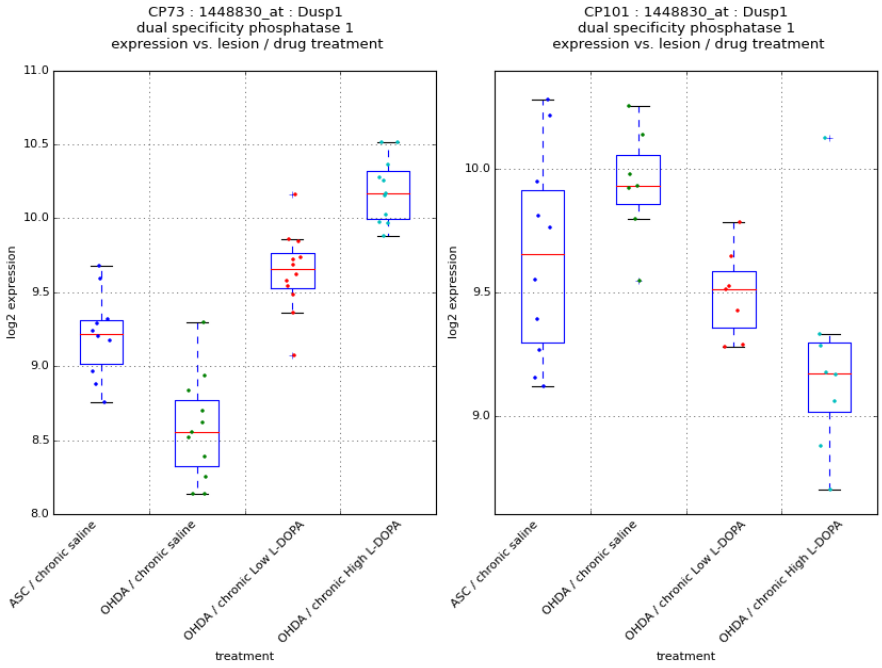


Figure 4.7: Expression changes in Dusp1, showing homeostatic responses to depletion of dopamine and L-DOPA treatment. Left panel, Drd1a dSPNs; Right panel , Drd2 iSPNs

may reflect more extensive structural changes and synaptic remodelling occurring under the high dose regimen.

Transcriptional regulators of the D1 dSPN response

To identify regulators of the dSPN response to chronic levodopa, we assembled data on occurrence of conserved motifs in promoters [94], and used the hypergeometric test to assess significance of overrepresentation of motifs among genes with altered expression.

Among genes that were up-regulated with chronic high-dose levodopa, the CREB, AP-1 (eg. Fos and Jun), and ERK-dependent (eg. Elk) motifs were among the most over-represented (Table 4.14), in addition to several motifs that are known to be common in neuronal-expressed genes, such as Sp1.

Probeset	Gene Symbol	Gene Description	P-value	log2 FC
1421079_at	Nr4a3	nuclear receptor subfamily 4, group A, member 3	$3.08 \cdot 10^{-10}$	5.8
1455034_at	Nr4a2	nuclear receptor subfamily 4, group A, member 2	$2.24 \cdot 10^{-12}$	5.73
1438796_at	Nr4a3	nuclear receptor subfamily 4, group A, member 3	$5.76 \cdot 10^{-8}$	5.55
1450750_a_at	Nr4a2	nuclear receptor subfamily 4, group A, member 2	$2.78 \cdot 10^{-10}$	5.37
1429475_at	Ubash3b	ubiquitin associated and SH3 domain containing, B	$1.85 \cdot 10^{-6}$	5.02
1423851_a_at	Shisa2	shisa homolog 2 (Xenopus laevis)	$6.58 \cdot 10^{-12}$	4.87
1421080_at	Nr4a3	nuclear receptor subfamily 4, group A, member 3	$2.24 \cdot 10^{-12}$	4.72
1427682_a_at	Egr2	early growth response 2	$8.19 \cdot 10^{-7}$	4.65
1441228_at	Apold1	apolipoprotein L domain containing 1	$7.34 \cdot 10^{-11}$	4.46
1447863_s_at	Nr4a2	nuclear receptor subfamily 4, group A, member 2	$6.58 \cdot 10^{-12}$	4.43
1423852_at	Shisa2	shisa homolog 2 (Xenopus laevis)	$6.69 \cdot 10^{-9}$	4.4
1427683_at	Egr2	early growth response 2	$7.93 \cdot 10^{-6}$	4.36
1451163_at	Tinf2	Terf1 (TRF1)-interacting nuclear factor 2	$2.47 \cdot 10^{-11}$	4.35
1422931_at	Fosl2	fos-like antigen 2	$7.3 \cdot 10^{-12}$	4.21
1436805_at	Ubash3b	ubiquitin associated and SH3 domain containing, B	$6.88 \cdot 10^{-10}$	4.2
1422053_at	Inhba	inhibin beta-A	$5.47 \cdot 10^{-9}$	4.13
1416700_at	Rnd3	Rho family GTPase 3	$2.07 \cdot 10^{-9}$	4.11
1435703_at	Ubash3b	ubiquitin associated and SH3 domain containing, B	$3.79 \cdot 10^{-8}$	4.07
1437247_at	Fosl2	fos-like antigen 2	$1.86 \cdot 10^{-11}$	4.02
1417263_at	Ptgs2	prostaglandin-endoperoxide synthase 2	$9.25 \cdot 10^{-9}$	3.92
1416701_at	Rnd3	Rho family GTPase 3	$1.63 \cdot 10^{-8}$	3.88
1436387_at			$4.83 \cdot 10^{-9}$	3.79
1425671_at	Homer1	homer homolog 1 (Drosophila)	$4.38 \cdot 10^{-9}$	3.77
1435872_at			$1.2 \cdot 10^{-7}$	3.73
1422256_at	Sstr2	somatostatin receptor 2	$3.44 \cdot 10^{-7}$	3.71
1420720_at	Nptx2	neuronal pentraxin 2	$2.4 \cdot 10^{-7}$	3.62
1433599_at	Baz1a	bromodomain adjacent to zinc finger domain 1A	$6.69 \cdot 10^{-9}$	3.62
1449188_at	Midn	midnolin	$3.57 \cdot 10^{-11}$	3.61
1434815_a_at	Mapkapk3	mitogen-activated protein kinase-activated protein kinase 3	$4.87 \cdot 10^{-6}$	3.6
1421396_at	Pcsk1	proprotein convertase subtilisin/kexin type 1	$1.29 \cdot 10^{-8}$	3.59
1439764_s_at	Igf2bp2	insulin-like growth factor 2 mRNA binding protein 2	$1.01 \cdot 10^{-5}$	3.58
1449960_at	Nptx2	neuronal pentraxin 2	$1.74 \cdot 10^{-9}$	3.55
1447930_at	Baz1a	bromodomain adjacent to zinc finger domain 1A	$2.5 \cdot 10^{-9}$	3.54
1418322_at	Crem	cAMP responsive element modulator	$2.32 \cdot 10^{-10}$	3.5
1460275_at	Gpr3	G-protein coupled receptor 3	$2.24 \cdot 10^{-12}$	3.46
1431422_a_at	Dusp14	dual specificity phosphatase 14	$1.96 \cdot 10^{-9}$	3.43
1418687_at	Arc	activity regulated cytoskeletal-associated protein	$2.24 \cdot 10^{-12}$	3.42
1459941_at	Clvs1	clavesin 1	$1.59 \cdot 10^{-9}$	3.41
1434350_at	Csrnp1	cysteine-serine-rich nuclear protein 1	$9.18 \cdot 10^{-9}$	3.4
1417695_a_at	Soat1	sterol O-acyltransferase 1	$7.31 \cdot 10^{-8}$	3.38
1449037_at	Crem	cAMP responsive element modulator	$1.4 \cdot 10^{-8}$	3.31
1435458_at	Pim1	proviral integration site 1	$1.88 \cdot 10^{-8}$	3.3
1417696_at	Soat1	sterol O-acyltransferase 1	$7.45 \cdot 10^{-8}$	3.29
1428834_at	Dusp4	dual specificity phosphatase 4	$1.08 \cdot 10^{-8}$	3.27
1422134_at	Fosb	FBJ osteosarcoma oncogene B	$1 \cdot 10^{-10}$	3.26
1436305_at	Rnf217	ring finger protein 217	$1.66 \cdot 10^{-11}$	3.26
1449405_at	Tns1	tensin 1	$1.87 \cdot 10^{-9}$	3.24
1419647_a_at	Ier3	immediate early response 3	$2.24 \cdot 10^{-12}$	3.23
1453590_at	Arl5b	ADP-ribosylation factor-like 5B	$3.33 \cdot 10^{-9}$	3.22
1450971_at	Gadd45b	growth arrest and DNA-damage-inducible 45 beta	$1.23 \cdot 10^{-6}$	3.13

Table 4.8: Top 50 genes up-regulated upon dopamine depletion in D1 dSPNs after chronic low-dose levodopa treatment

Probeset	Gene Symbol	Gene Description	P-value	log2 FC
1422313_a_at	Igfbp5	insulin-like growth factor binding protein 5	$5.38 \cdot 10^{-3}$	-2.2
1424470_a_at	Rapgef3	Rap guanine nucleotide exchange factor (GEF) 3	$5.89 \cdot 10^{-3}$	-2.18
1415800_at	Gja1	gap junction protein, alpha 1	$3.66 \cdot 10^{-5}$	-2.17
1437937_at	Ccbp2	chemokine binding protein 2	$4.16 \cdot 10^{-4}$	-2.15
1455720_at	Adamts2	a disintegrin-like and metallopeptidase (reprolysin type) with thrombospondin type 1 motif, 2	$2.22 \cdot 10^{-6}$	-2.09
1429514_at	Ppap2b	phosphatidic acid phosphatase type 2B	$1.76 \cdot 10^{-5}$	-2.08
1449365_at	S1pr5	sphingosine-1-phosphate receptor 5	$3.11 \cdot 10^{-5}$	-2.04
1452114_s_at	Igfbp5	insulin-like growth factor binding protein 5	$5.37 \cdot 10^{-5}$	-2.02
1438852_x_at	Mcm6	minichromosome maintenance deficient 6 (MIS5 homolog, S. pombe) (S. cerevisiae)	$2.82 \cdot 10^{-4}$	-2
1455556_at	Notch2	notch 2	$4.73 \cdot 10^{-3}$	-1.98
1447223_at			$1.4 \cdot 10^{-4}$	-1.91
1443129_at			$8.01 \cdot 10^{-6}$	-1.9
1439293_at	Fam214a	family with sequence similarity 214, member A	$1.54 \cdot 10^{-4}$	-1.89
1433639_at	Fam117a	family with sequence similarity 117, member A	$2.27 \cdot 10^{-3}$	-1.89
1452473_at	Prr15	proline rich 15	$2.8 \cdot 10^{-3}$	-1.87
1423284_at	Mansc1	MANSC domain containing 1	$5.29 \cdot 10^{-8}$	-1.79
1436600_at	Tox3	TOX high mobility group box family member 3	$1.79 \cdot 10^{-5}$	-1.79
1429089_s_at	2900026A02Rik	RIKEN cDNA 2900026A02 gene	$2.26 \cdot 10^{-6}$	-1.78
1451245_at	Lrrc3b	leucine rich repeat containing 3B	$1.49 \cdot 10^{-5}$	-1.76
1434893_at	Atp1a2	ATPase, Na+/K+ transporting, alpha 2 polypeptide	$1.55 \cdot 10^{-3}$	-1.75
1456967_at	Trim66	tripartite motif-containing 66	$1.55 \cdot 10^{-4}$	-1.74
1428332_at	Pik3ip1	phosphoinositide-3-kinase interacting protein 1	$1.31 \cdot 10^{-8}$	-1.73
1456603_at	Fam101b	family with sequence similarity 101, member B	$8.35 \cdot 10^{-6}$	-1.73
1429764_at	Fam101b	family with sequence similarity 101, member B	$2.44 \cdot 10^{-5}$	-1.73
1421840_at	Abca1	ATP-binding cassette, sub-family A (ABC1), member 1	$9.94 \cdot 10^{-5}$	-1.72
1456047_at			$4.52 \cdot 10^{-6}$	-1.72
1455972_x_at	Hadh	hydroxyacyl-Coenzyme A dehydrogenase	$1.09 \cdot 10^{-2}$	-1.7
1450799_at	Adcyap1r1	adenylate cyclase activating polypeptide 1 receptor 1	$5.59 \cdot 10^{-4}$	-1.69
1419063_at	Ugt8a	UDP galactosyltransferase 8A	$6.84 \cdot 10^{-4}$	-1.69
1423367_at	Wnt7a	wingless-related MMTV integration site 7A	$3.61 \cdot 10^{-7}$	-1.67
1442831_at			$4.72 \cdot 10^{-3}$	-1.65
1435407_at			$2.35 \cdot 10^{-6}$	-1.64
1459838_s_at	Btbd11	BTB (POZ) domain containing 11	$1.81 \cdot 10^{-2}$	-1.63
1428758_at	Tmem86a	transmembrane protein 86A	$5.78 \cdot 10^{-4}$	-1.63
1456005_a_at	Bcl2l11	BCL2-like 11 (apoptosis facilitator)	$1.47 \cdot 10^{-2}$	-1.63
1435125_at			$2.52 \cdot 10^{-5}$	-1.62
1424468_s_at	Phldb1	pleckstrin homology-like domain, family B, member 1	$2.74 \cdot 10^{-3}$	-1.62
1422529_s_at	Casq2	calsequestrin 2	$8.2 \cdot 10^{-5}$	-1.62
1439627_at	Zic1	zinc finger protein of the cerebellum 1	$7.48 \cdot 10^{-3}$	-1.62
1448127_at	Rrm1	ribonucleotide reductase M1	$1.75 \cdot 10^{-3}$	-1.61
1460136_at	AW047481	expressed sequence AW047481	$6.24 \cdot 10^{-4}$	-1.61
1417520_at	Nfe2l3	nuclear factor, erythroid derived 2, like 3	$3.12 \cdot 10^{-5}$	-1.61
1439715_at	Osgepl1	O-sialoglycoprotein endopeptidase-like 1	$1.53 \cdot 10^{-2}$	-1.61
1419064_a_at	Ugt8a	UDP galactosyltransferase 8A	$9.17 \cdot 10^{-4}$	-1.6
1438650_x_at	Gja1	gap junction protein, alpha 1	$2.08 \cdot 10^{-3}$	-1.59
1433489_s_at	Fgfr2	fibroblast growth factor receptor 2	$1.17 \cdot 10^{-4}$	-1.59
1443773_at	Ylpm1	YLP motif containing 1	$1.84 \cdot 10^{-5}$	-1.58
1442119_at	AI449212	expressed sequence AI449212	$1.53 \cdot 10^{-4}$	-1.58
1450712_at	Kcnj9	potassium inwardly-rectifying channel, subfamily J, member 9	$1.13 \cdot 10^{-4}$	-1.58
1455854_a_at	Ssh1	slingshot homolog 1 (Drosophila)	$3.48 \cdot 10^{-3}$	-1.57
1460560_at	Bahcc1	BAH domain and coiled-coil containing 1	$2.99 \cdot 10^{-7}$	-1.57

Table 4.9: Top 50 genes down-regulated upon dopamine depletion in D1 dSPNs chronic high-dose levodopa treatment

Set Name	Matches	Size of Set	p-value	Bonf adj p-val	b-h FDR adj p-val
mRNA processing	231	494	$2.67 \cdot 10^{-9}$	$4.05 \cdot 10^{-7}$	$4.05 \cdot 10^{-7}$
EGFR1 Signaling Pathway	95	176	$4.38 \cdot 10^{-8}$	$6.65 \cdot 10^{-6}$	$3.33 \cdot 10^{-6}$
Splicing factor NOVA regulated synpatic proteins	29	42	$3.95 \cdot 10^{-6}$	$6.01 \cdot 10^{-4}$	$2 \cdot 10^{-4}$
TCA Cycle	23	31	$5.92 \cdot 10^{-6}$	$8.99 \cdot 10^{-4}$	$2.25 \cdot 10^{-4}$
TGF-beta Receptor Signaling Pathway	78	155	$2.16 \cdot 10^{-5}$	$3.28 \cdot 10^{-3}$	$6.56 \cdot 10^{-4}$
IL-2 Signaling Pathway	44	77	$2.79 \cdot 10^{-5}$	$4.24 \cdot 10^{-3}$	$7.07 \cdot 10^{-4}$
Insulin Signaling	78	157	$3.83 \cdot 10^{-5}$	$5.83 \cdot 10^{-3}$	$8.32 \cdot 10^{-4}$
MAPK signaling pathway	78	160	$8.67 \cdot 10^{-5}$	$1.32 \cdot 10^{-2}$	$1.65 \cdot 10^{-3}$
T Cell Receptor Signaling Pathway	66	134	$2.01 \cdot 10^{-4}$	$3.05 \cdot 10^{-2}$	$3.39 \cdot 10^{-3}$
PluriNetWork	133	303	$2.28 \cdot 10^{-4}$	$3.47 \cdot 10^{-2}$	$3.47 \cdot 10^{-3}$
Signaling of Hepatocyte Growth Factor Receptor	22	34	$2.57 \cdot 10^{-4}$	$3.91 \cdot 10^{-2}$	$3.55 \cdot 10^{-3}$
B Cell Receptor Signaling Pathway	77	163	$3.37 \cdot 10^{-4}$	$5.13 \cdot 10^{-2}$	$4.17 \cdot 10^{-3}$
IL-6 signaling Pathway	51	100	$3.57 \cdot 10^{-4}$	$5.42 \cdot 10^{-2}$	$4.17 \cdot 10^{-3}$
IL-3 Signaling Pathway	51	102	$6.5 \cdot 10^{-4}$	$9.88 \cdot 10^{-2}$	$7.06 \cdot 10^{-3}$
TNF-alpha NF-kB Signaling Pathway	89	198	$9.44 \cdot 10^{-4}$	0.14	$9.57 \cdot 10^{-3}$
Electron Transport Chain	57	119	$1.24 \cdot 10^{-3}$	0.19	$1.18 \cdot 10^{-2}$
G13 Signaling Pathway	24	42	$1.82 \cdot 10^{-3}$	0.28	$1.63 \cdot 10^{-2}$
Diurnally regulated genes with circadian orthologs	29	55	$3.37 \cdot 10^{-3}$	0.51	$2.85 \cdot 10^{-2}$
Calcium Regulation in the Cardiac Cell	69	154	$3.64 \cdot 10^{-3}$	0.55	$2.91 \cdot 10^{-2}$
G Protein Signaling Pathways	47	99	$3.95 \cdot 10^{-3}$	0.6	$3 \cdot 10^{-2}$

Table 4.10: Wikipathways pathways over-represented among genes changed in D1 dSPNs upon dopamine depletion and chronic low-dose levodopa treatment

Probeset	Gene Symbol	Gene Description	P-value	log2 FC
1450750_a_at	Nr4a2	nuclear receptor subfamily 4, group A, member 2	$1.02 \cdot 10^{-17}$	7.29
1455034_at	Nr4a2	nuclear receptor subfamily 4, group A, member 2	$1.03 \cdot 10^{-11}$	7.22
1421079_at	Nr4a3	nuclear receptor subfamily 4, group A, member 3	$2.21 \cdot 10^{-9}$	6.31
1447863_s_at	Nr4a2	nuclear receptor subfamily 4, group A, member 2	$4.64 \cdot 10^{-13}$	5.98
1438796_at	Nr4a3	nuclear receptor subfamily 4, group A, member 3	$2.71 \cdot 10^{-8}$	5.89
1423851_a_at	Shisa2	shisa homolog 2 (Xenopus laevis)	$3.29 \cdot 10^{-11}$	5.65
1417263_at	Ptgs2	prostaglandin-endoperoxide synthase 2	$1.46 \cdot 10^{-15}$	5.59
1451163_at	Tinf2	Terf1 (TRF1)-interacting nuclear factor 2	$6.4 \cdot 10^{-11}$	5.48
1429475_at	Ubash3b	ubiquitin associated and SH3 domain containing, B	$5.82 \cdot 10^{-7}$	5.41
1421134_at	Areg	amphiregulin	$4.25 \cdot 10^{-13}$	5.21
1421080_at	Nr4a3	nuclear receptor subfamily 4, group A, member 3	$2.57 \cdot 10^{-11}$	5.1
1423852_at	Shisa2	shisa homolog 2 (Xenopus laevis)	$5.63 \cdot 10^{-9}$	5.05
1441228_at	Apold1	apolipoprotein L domain containing 1	$5.19 \cdot 10^{-11}$	4.91
1437166_at	Tinf2	Terf1 (TRF1)-interacting nuclear factor 2	$9.08 \cdot 10^{-10}$	4.88
1422256_at	Sstr2	somatostatin receptor 2	$3.44 \cdot 10^{-8}$	4.74
1419082_at	Serpinb2	serine (or cysteine) peptidase inhibitor, clade B, member 2	$7.49 \cdot 10^{-10}$	4.74
1422931_at	Fosl2	fos-like antigen 2	$4.99 \cdot 10^{-15}$	4.73
1427682_a_at	Egr2	early growth response 2	$6.18 \cdot 10^{-7}$	4.66
1427455_x_at			$3 \cdot 10^{-8}$	4.61
1417262_at	Ptgs2	prostaglandin-endoperoxide synthase 2	$2.58 \cdot 10^{-9}$	4.59
1416700_at	Rnd3	Rho family GTPase 3	$1.99 \cdot 10^{-9}$	4.55
1449960_at	Nptx2	neuronal pentraxin 2	$5.18 \cdot 10^{-10}$	4.51
1434815_a_at	Mapkapk3	mitogen-activated protein kinase-activated protein kinase 3	$1.3 \cdot 10^{-6}$	4.51
1422053_at	Inhba	inhibin beta-A	$5.55 \cdot 10^{-9}$	4.49
1436805_at	Ubash3b	ubiquitin associated and SH3 domain containing, B	$1.39 \cdot 10^{-9}$	4.48
1427660_x_at			$3.06 \cdot 10^{-8}$	4.39
1427683_at	Egr2	early growth response 2	$4.54 \cdot 10^{-6}$	4.38
1435872_at			$2.71 \cdot 10^{-8}$	4.38
1460275_at	Gpr3	G-protein coupled receptor 3	$9.95 \cdot 10^{-14}$	4.37
1435703_at	Ubash3b	ubiquitin associated and SH3 domain containing, B	$1.85 \cdot 10^{-8}$	4.35
1419647_a_at	Ier3	immediate early response 3	$4.99 \cdot 10^{-15}$	4.35
1450188_s_at	Lipg	lipase, endothelial	$1.11 \cdot 10^{-8}$	4.33
1420653_at	Tgfb1	transforming growth factor, beta 1	$4.22 \cdot 10^{-12}$	4.32
1417696_at	Soat1	sterol O-acyltransferase 1	$5.44 \cdot 10^{-9}$	4.32
1421396_at	Pcsk1	proprotein convertase subtilisin/kexin type 1	$3.32 \cdot 10^{-8}$	4.31
1437247_at	Fosl2	fos-like antigen 2	$1.17 \cdot 10^{-10}$	4.31
1450749_a_at	Nr4a2	nuclear receptor subfamily 4, group A, member 2	$3.58 \cdot 10^{-13}$	4.31
1428834_at	Dusp4	dual specificity phosphatase 4	$3.37 \cdot 10^{-10}$	4.3
1417695_a_at	Soat1	sterol O-acyltransferase 1	$2.91 \cdot 10^{-8}$	4.28
1420720_at	Nptx2	neuronal pentraxin 2	$2.73 \cdot 10^{-8}$	4.26
1424246_a_at	Tes	testis derived transcript	$5.18 \cdot 10^{-10}$	4.25
1447930_at	Baz1a	bromodomain adjacent to zinc finger domain 1A	$9.85 \cdot 10^{-10}$	4.22
1452417_x_at			$5.91 \cdot 10^{-7}$	4.21
1431057_a_at	Prss23	protease, serine, 23	$4.64 \cdot 10^{-13}$	4.2
1416701_at	Rnd3	Rho family GTPase 3	$4.89 \cdot 10^{-8}$	4.19
1452557_a_at	Igk	immunoglobulin kappa chain complex	$1.69 \cdot 10^{-6}$	4.18
1418687_at	Arc	activity regulated cytoskeletal-associated protein	$3.11 \cdot 10^{-12}$	4.16
1433599_at	Baz1a	bromodomain adjacent to zinc finger domain 1A	$5.23 \cdot 10^{-9}$	4.14
1418322_at	Crem	cAMP responsive element modulator	$1.44 \cdot 10^{-9}$	4.11
1436305_at	Rnf217	ring finger protein 217	$4.52 \cdot 10^{-14}$	4.06
1436387_at			$2.58 \cdot 10^{-9}$	4.02
1416554_at	Pdlim1	PDZ and LIM domain 1 (elfin)	$1.29 \cdot 10^{-10}$	4

Table 4.11: Top 50 genes up-regulated upon dopamine depletion in D1 dSPNs after chronic high-dose levodopa treatment

Probeset	Gene Symbol	Gene Description	P-value	log2 FC
1456351_at	Brd8	bromodomain containing 8	$1.7 \cdot 10^{-7}$	-2.9
1455556_at	Notch2	notch 2	$6.55 \cdot 10^{-5}$	-2.78
1456967_at	Trim66	tripartite motif-containing 66	$5.19 \cdot 10^{-7}$	-2.74
1424470_a_at	Rapgef3	Rap guanine nucleotide exchange factor (GEF) 3	$9.45 \cdot 10^{-4}$	-2.73
1452114_s_at	Igfbp5	insulin-like growth factor binding protein 5	$3.94 \cdot 10^{-6}$	-2.67
1429089_s_at	2900026A02Rik	RIKEN cDNA 2900026A02 gene	$2.33 \cdot 10^{-6}$	-2.64
1455720_at	Adamts2	a disintegrin-like and metallopeptidase (reprolysin type) with thrombospondin type 1 motif, 2	$2.23 \cdot 10^{-7}$	-2.61
1444139_at	Ddit4l	DNA-damage-inducible transcript 4-like	$2.1 \cdot 10^{-6}$	-2.55
1439332_at	Ddit4l	DNA-damage-inducible transcript 4-like	$2.01 \cdot 10^{-6}$	-2.43
1435407_at			$4.69 \cdot 10^{-7}$	-2.41
1423756_s_at	Igfbp4	insulin-like growth factor binding protein 4	$2.26 \cdot 10^{-4}$	-2.4
1455972_x_at	Hadh	hydroxyacyl-Coenzyme A dehydrogenase	$2.41 \cdot 10^{-4}$	-2.4
1459838_s_at	Btbd11	BTB (POZ) domain containing 11	$2.17 \cdot 10^{-4}$	-2.4
1422313_a_at	Igfbp5	insulin-like growth factor binding protein 5	$4 \cdot 10^{-3}$	-2.39
1433639_at	Fam117a	family with sequence similarity 117, member A	$1.42 \cdot 10^{-4}$	-2.35
1425096_a_at	Ptcd1	pentatricopeptide repeat domain 1	$3.71 \cdot 10^{-9}$	-2.33
1437937_at	Ccbp2	chemokine binding protein 2	$2.42 \cdot 10^{-4}$	-2.32
1425092_at	Cdh10	cadherin 10	$7.11 \cdot 10^{-10}$	-2.28
1437405_a_at	Igfbp4	insulin-like growth factor binding protein 4	$5.16 \cdot 10^{-5}$	-2.28
1440202_at			$6.61 \cdot 10^{-7}$	-2.26
1435125_at			$5.41 \cdot 10^{-6}$	-2.25
1419200_at	Fxyd7	FXYD domain-containing ion transport regulator 7	$8.43 \cdot 10^{-8}$	-2.24
1434672_at	Gpr22	G protein-coupled receptor 22	$3.06 \cdot 10^{-6}$	-2.22
1439715_at	Osgepl1	O-sialoglycoprotein endopeptidase-like 1	$7.12 \cdot 10^{-5}$	-2.19
1451245_at	Lrrc3b	leucine rich repeat containing 3B	$6.68 \cdot 10^{-8}$	-2.18
1454973_at	Atf7ip	activating transcription factor 7 interacting protein	$3.24 \cdot 10^{-5}$	-2.17
1455134_at	Tmem245	transmembrane protein 245	$5.09 \cdot 10^{-5}$	-2.14
1450511_at	Musk	muscle, skeletal, receptor tyrosine kinase	$1.55 \cdot 10^{-7}$	-2.11
1459703_at			$1.21 \cdot 10^{-6}$	-2.1
1434131_at	Rufy1	RUN and FYVE domain containing 1	$6.35 \cdot 10^{-6}$	-2.09
1436786_at	Sec14l3	SEC14-like 3 (S. cerevisiae)	$6.78 \cdot 10^{-4}$	-2.09
1450712_at	Kcnj9	potassium inwardly-rectifying channel, subfamily J, member 9	$5.69 \cdot 10^{-6}$	-2.08
1436610_at	Ankrd12	ankyrin repeat domain 12	$1.08 \cdot 10^{-6}$	-2.07
1453061_at	Elac1	elaC homolog 1 (E. coli)	$3.06 \cdot 10^{-6}$	-2.07
1447223_at			$7.29 \cdot 10^{-4}$	-2.07
1420870_at	Mllt10	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 10	$1.3 \cdot 10^{-6}$	-2.06
1425173_s_at	Golph3l	golgi phosphoprotein 3-like	$1.79 \cdot 10^{-5}$	-2.06
1437406_x_at	Igfbp4	insulin-like growth factor binding protein 4	$3.81 \cdot 10^{-5}$	-2.06
1436005_at	Sugp2	SURP and G patch domain containing 2	$1.55 \cdot 10^{-5}$	-2.05
1428332_at	Pik3ip1	phosphoinositide-3-kinase interacting protein 1	$3.97 \cdot 10^{-6}$	-2.04
1456005_a_at	Bcl2l11	BCL2-like 11 (apoptosis facilitator)	$3.61 \cdot 10^{-4}$	-2.03
1456674_at			$2.15 \cdot 10^{-4}$	-2.03
1451751_at	Ddit4l	DNA-damage-inducible transcript 4-like	$3.05 \cdot 10^{-5}$	-2.02
1437623_x_at	Xrcc3	X-ray repair complementing defective repair in Chinese hamster cells 3	$8.99 \cdot 10^{-6}$	-2.01
1436501_at	Mtus1	mitochondrial tumor suppressor 1	$5.36 \cdot 10^{-5}$	-2
1426434_at	Tmem43	transmembrane protein 43	$1.61 \cdot 10^{-5}$	-2
1439602_at	Fign	fidgetin	$1.54 \cdot 10^{-3}$	-1.99
1439293_at	Fam214a	family with sequence similarity 214, member A	$5.23 \cdot 10^{-5}$	-1.98
1424597_at	Wash	WAS protein family homolog	$5.75 \cdot 10^{-5}$	-1.98
1442139_at			$5.79 \cdot 10^{-5}$	-1.97
1453429_at	9530057J20Rik	RIKEN cDNA 9530057J20 gene	$1.9 \cdot 10^{-7}$	-1.97

Table 4.12: Top 50 genes down-regulated upon dopamine depletion in D1 dSPNs after chronic high-dose levodopa treatment

Set Name	Matches	Size of Set	p-value	Bonf adj p-val	b-h FDR adj p-val
EGFR1 Signaling Pathway	115	176	$2.72 \cdot 10^{-8}$	$4.19 \cdot 10^{-6}$	$4.19 \cdot 10^{-6}$
MAPK signaling pathway	103	160	$4.24 \cdot 10^{-7}$	$6.52 \cdot 10^{-5}$	$2.94 \cdot 10^{-5}$
Insulin Signaling	101	157	$5.73 \cdot 10^{-7}$	$8.83 \cdot 10^{-5}$	$2.94 \cdot 10^{-5}$
mRNA processing	266	494	$2.55 \cdot 10^{-5}$	$3.93 \cdot 10^{-3}$	$9.83 \cdot 10^{-4}$
Splicing factor NOVA regulated synpatic proteins	32	42	$3.38 \cdot 10^{-5}$	$5.21 \cdot 10^{-3}$	$1.04 \cdot 10^{-3}$
TGF-beta Receptor Signaling Pathway	93	155	$9.27 \cdot 10^{-5}$	$1.43 \cdot 10^{-2}$	$2.34 \cdot 10^{-3}$
TNF-alpha NF-kB Signaling Pathway	115	198	$1.06 \cdot 10^{-4}$	$1.64 \cdot 10^{-2}$	$2.34 \cdot 10^{-3}$
Signaling of Hepatocyte Growth Factor Receptor	26	34	$1.69 \cdot 10^{-4}$	$2.61 \cdot 10^{-2}$	$3.26 \cdot 10^{-3}$
TCA Cycle	24	31	$2.18 \cdot 10^{-4}$	$3.36 \cdot 10^{-2}$	$3.74 \cdot 10^{-3}$
G Protein Signaling Pathways	62	99	$2.55 \cdot 10^{-4}$	$3.93 \cdot 10^{-2}$	$3.93 \cdot 10^{-3}$
IL-6 signaling Pathway	62	100	$3.8 \cdot 10^{-4}$	$5.85 \cdot 10^{-2}$	$5.32 \cdot 10^{-3}$
MicroRNAs in cardiomyocyte hypertrophy	53	85	$8.09 \cdot 10^{-4}$	0.12	$1.04 \cdot 10^{-2}$
Regulation of Actin Cytoskeleton	86	149	$9.69 \cdot 10^{-4}$	0.15	$1.15 \cdot 10^{-2}$
G13 Signaling Pathway	29	42	$1.27 \cdot 10^{-3}$	0.2	$1.38 \cdot 10^{-2}$
PluriNetWork	162	303	$1.34 \cdot 10^{-3}$	0.21	$1.38 \cdot 10^{-2}$
T Cell Receptor Signaling Pathway	77	134	$2.05 \cdot 10^{-3}$	0.32	$1.97 \cdot 10^{-2}$
MAPK Cascade	21	29	$2.39 \cdot 10^{-3}$	0.37	$2.16 \cdot 10^{-2}$
Hypothetical Network for Drug Addiction	22	31	$2.86 \cdot 10^{-3}$	0.44	$2.29 \cdot 10^{-2}$
IL-2 Signaling Pathway	47	77	$2.92 \cdot 10^{-3}$	0.45	$2.29 \cdot 10^{-2}$
G1 to S cell cycle control	39	62	$3.03 \cdot 10^{-3}$	0.47	$2.29 \cdot 10^{-2}$

Table 4.13: Wikipathways pathways over-represented among genes changed in D1 dSPNs upon dopamine depletion and chronic high-dose levodopa treatment

Motif Name	Overlap	Total Genes in Group	Total motif occurrences	P-value	bh FDR adj p-val
TFAP2A,C.p2	1,363	1,898	11,209	$2.02 \cdot 10^{-83}$	0
TFDP1.p2	1,274	1,898	10,211	$1.16 \cdot 10^{-80}$	0
SP1.p2	1,590	1,898	14,315	$5.26 \cdot 10^{-80}$	0
TFAP2B.p2	1,142	1,898	8,716	$9.03 \cdot 10^{-80}$	0
ATF5_CREB3.p2	467	1,898	2,370	$4.71 \cdot 10^{-73}$	0
KLF4.p3	1,456	1,898	13,075	$8.74 \cdot 10^{-62}$	0
MAZ.p2	1,485	1,898	13,533	$1.51 \cdot 10^{-59}$	0
PATZ1.p2	1,415	1,898	12,837	$1.48 \cdot 10^{-53}$	0
PAX5.p2	975	1,898	7,682	$1.74 \cdot 10^{-53}$	0
HIC1.p2	878	1,898	6,807	$1.01 \cdot 10^{-48}$	0
JUN.p2	326	1,898	1,812	$2.03 \cdot 10^{-40}$	0
MAFB.p2	702	1,898	5,292	$6.24 \cdot 10^{-40}$	0
ELF1,2,4.p2	894	1,898	7,322	$4.04 \cdot 10^{-39}$	0
ELK1,4_GABPA,B1.p3	723	1,898	5,564	$3.27 \cdot 10^{-38}$	0
ATF4.p2	228	1,898	1,130	$1.54 \cdot 10^{-35}$	0
EGR1..3.p2	679	1,898	5,213	$1.61 \cdot 10^{-35}$	0
CREB1.p2	245	1,898	1,312	$1.35 \cdot 10^{-32}$	0
MZF1.p2	1,203	1,898	11,270	$4.49 \cdot 10^{-29}$	0
ZFP161.p2	590	1,898	4,556	$4.79 \cdot 10^{-29}$	0
GTF2I.p2	1,026	1,898	9,237	$1.2 \cdot 10^{-28}$	0
MTF1.p2	593	1,898	4,608	$1.81 \cdot 10^{-28}$	0
AHR_ARNT_ARNT2.p2	410	1,898	2,993	$9.91 \cdot 10^{-24}$	0
PAX2.p2	310	1,898	2,127	$9.52 \cdot 10^{-22}$	0
ATF6.p2	184	1,898	1,049	$2.83 \cdot 10^{-21}$	0
EHF.p2	440	1,898	3,550	$4.5 \cdot 10^{-17}$	0
NHLH1,2.p2	557	1,898	4,758	$8.65 \cdot 10^{-17}$	$1.78 \cdot 10^{-14}$
SPI1.p2	831	1,898	7,746	$4.29 \cdot 10^{-16}$	$7.06 \cdot 10^{-14}$
NRF1.p2	436	1,898	3,576	$1.17 \cdot 10^{-15}$	$1.93 \cdot 10^{-13}$
EP300.p2	270	1,898	1,963	$1.39 \cdot 10^{-15}$	$2.27 \cdot 10^{-13}$
HES1.p2	274	1,898	2,079	$1.58 \cdot 10^{-13}$	$2.47 \cdot 10^{-11}$
ATF2.p2	138	1,898	858	$4.04 \cdot 10^{-13}$	$6.26 \cdot 10^{-11}$
NFYA,B,C.p2	386	1,898	3,247	$3.61 \cdot 10^{-12}$	$5.56 \cdot 10^{-10}$
GFI1.p2	275	1,898	2,152	$5.05 \cdot 10^{-12}$	$7.72 \cdot 10^{-10}$
FEV.p2	605	1,898	5,575	$1.28 \cdot 10^{-11}$	$1.94 \cdot 10^{-9}$
SPIB.p2	628	1,898	5,868	$5.46 \cdot 10^{-11}$	$8.25 \cdot 10^{-9}$
TFCP2.p2	331	1,898	2,758	$6.36 \cdot 10^{-11}$	$9.53 \cdot 10^{-9}$
SNAI1..3.p2	746	1,898	7,210	$2.27 \cdot 10^{-10}$	$3.38 \cdot 10^{-8}$
ZNF148.p2	457	1,898	4,093	$2.94 \cdot 10^{-10}$	$4.35 \cdot 10^{-8}$
YY1.p2	1,078	1,898	11,035	$7.95 \cdot 10^{-10}$	$1.17 \cdot 10^{-7}$
bHLH_family.p2	418	1,898	3,714	$8.2 \cdot 10^{-10}$	$1.2 \cdot 10^{-7}$
TCF4_dimer.p2	818	1,898	8,090	$1.82 \cdot 10^{-9}$	$2.63 \cdot 10^{-7}$
ZBTB6.p2	221	1,898	1,789	$1.59 \cdot 10^{-8}$	$2.29 \cdot 10^{-6}$
RREB1.p2	676	1,898	6,598	$1.99 \cdot 10^{-8}$	$2.85 \cdot 10^{-6}$
FOXN1.p2	69	1,898	403	$2.76 \cdot 10^{-8}$	$3.93 \cdot 10^{-6}$
HIF1A.p2	152	1,898	1,139	$3.03 \cdot 10^{-8}$	$4.27 \cdot 10^{-6}$
MYFfamily.p2	545	1,898	5,225	$1.05 \cdot 10^{-7}$	$1.48 \cdot 10^{-5}$
ZEB1.p2	391	1,898	3,590	$1.64 \cdot 10^{-7}$	$2.28 \cdot 10^{-5}$
ARNT_ARNT2_BHLHB2_MAX_MYC_USF1.p2	205	1,898	1,691	$2.3 \cdot 10^{-7}$	$3.17 \cdot 10^{-5}$
SPZ1.p2	313	1,898	2,787	$2.43 \cdot 10^{-7}$	$3.34 \cdot 10^{-5}$
ETS1,2.p2	343	1,898	3,122	$5.12 \cdot 10^{-7}$	$6.96 \cdot 10^{-5}$
MYOD1.p2	270	1,898	2,407	$2.12 \cdot 10^{-6}$	$2.87 \cdot 10^{-4}$

Table 4.14: Motifs over-represented among genes up-regulated in D1 dSPNs upon dopamine depletion and chronic high-dose levodopa treatment

Overlap	Motif Name	Total Genes in Group	P-value	Total motif occurrences	bh FDR adj p-val
2,196	SP1.p2	2,733	$8.89 \cdot 10^{-78}$	14,315	0
1,716	TFDP1.p2	2,733	$1.88 \cdot 10^{-75}$	10,211	0
1,806	TFAP2A,C.p2	2,733	$2.3 \cdot 10^{-65}$	11,209	0
1,236	HIC1.p2	2,733	$1.92 \cdot 10^{-63}$	6,807	0
1,484	TFAP2B.p2	2,733	$6.91 \cdot 10^{-62}$	8,716	0
1,947	PATZ1.p2	2,733	$3.02 \cdot 10^{-50}$	12,837	0
2,021	MAZ.p2	2,733	$1.54 \cdot 10^{-48}$	13,533	0
1,300	PAX5.p2	2,733	$3.53 \cdot 10^{-48}$	7,682	0
1,934	KLF4.p3	2,733	$1.71 \cdot 10^{-39}$	13,075	0
822	ZFP161.p2	2,733	$1.46 \cdot 10^{-35}$	4,556	0
792	MTF1.p2	2,733	$1.01 \cdot 10^{-26}$	4,608	0
869	EGR1..3.p2	2,733	$2.86 \cdot 10^{-25}$	5,213	0
873	MAFB.p2	2,733	$6.91 \cdot 10^{-24}$	5,292	0
1,637	MZF1.p2	2,733	$1.47 \cdot 10^{-22}$	11,270	0
1,375	GTF2I.p2	2,733	$4.41 \cdot 10^{-21}$	9,237	0
530	AHR_ARNT_ARNT2.p2	2,733	$1.09 \cdot 10^{-19}$	2,993	0
611	NRF1.p2	2,733	$3.43 \cdot 10^{-19}$	3,576	0
769	NHLH1,2.p2	2,733	$4.97 \cdot 10^{-18}$	4,758	0
357	ZNF143.p2	2,733	$7.35 \cdot 10^{-17}$	1,904	$1.87 \cdot 10^{-14}$
598	bHLH_family.p2	2,733	$2.74 \cdot 10^{-13}$	3,714	$4.58 \cdot 10^{-11}$
846	ELK1,4_GABPA,B1.p3	2,733	$5.31 \cdot 10^{-13}$	5,564	$8.82 \cdot 10^{-11}$
362	HES1.p2	2,733	$2.5 \cdot 10^{-12}$	2,079	$4.12 \cdot 10^{-10}$
643	ZNF148.p2	2,733	$3.38 \cdot 10^{-12}$	4,093	$5.54 \cdot 10^{-10}$
570	ZEB1.p2	2,733	$1.53 \cdot 10^{-11}$	3,590	$2.49 \cdot 10^{-9}$
1,050	SNAI1..3.p2	2,733	$1.73 \cdot 10^{-11}$	7,210	$2.81 \cdot 10^{-9}$
1,064	ELF1,2,4.p2	2,733	$1.97 \cdot 10^{-11}$	7,322	$3.18 \cdot 10^{-9}$
1,157	TCF4_dimer.p2	2,733	$9.27 \cdot 10^{-11}$	8,090	$1.48 \cdot 10^{-8}$
257	SREBF1,2.p2	2,733	$1.19 \cdot 10^{-10}$	1,420	$1.9 \cdot 10^{-8}$
389	SOX2.p2	2,733	$1.05 \cdot 10^{-9}$	2,370	$1.65 \cdot 10^{-7}$
133	E2F1..5.p2	2,733	$9.5 \cdot 10^{-9}$	662	$1.49 \cdot 10^{-6}$
501	NFYA,B,C.p2	2,733	$2.39 \cdot 10^{-8}$	3,247	$3.74 \cdot 10^{-6}$
284	ARNT_ARNT2_BHLHB2_MAX_MYC_USF1.p2	2,733	$3.01 \cdot 10^{-8}$	1,691	$4.67 \cdot 10^{-6}$
451	HBP1_HMGB_SSRP1_UBTF.p2	2,733	$3.73 \cdot 10^{-8}$	2,891	$5.75 \cdot 10^{-6}$
333	SOX17.p2	2,733	$5.51 \cdot 10^{-8}$	2,048	$8.43 \cdot 10^{-6}$
294	ZBTB6.p2	2,733	$1.37 \cdot 10^{-7}$	1,789	$2.08 \cdot 10^{-5}$
454	LMO2.p2	2,733	$2.47 \cdot 10^{-7}$	2,958	$3.73 \cdot 10^{-5}$
178	TFEB.p2	2,733	$3.9 \cdot 10^{-7}$	1,003	$5.85 \cdot 10^{-5}$
427	SPZ1.p2	2,733	$7.46 \cdot 10^{-7}$	2,787	$1.11 \cdot 10^{-4}$
926	RREB1.p2	2,733	$1.45 \cdot 10^{-6}$	6,598	$2.15 \cdot 10^{-4}$
420	TFCP2.p2	2,733	$1.8 \cdot 10^{-6}$	2,758	$2.65 \cdot 10^{-4}$
246	SOX8,9,10.p2	2,733	$2.15 \cdot 10^{-6}$	1,503	$3.14 \cdot 10^{-4}$
195	SOX5.p2	2,733	$3.17 \cdot 10^{-6}$	1,153	$4.6 \cdot 10^{-4}$
192	HIF1A.p2	2,733	$4.68 \cdot 10^{-6}$	1,139	$6.74 \cdot 10^{-4}$
1,066	SPI1.p2	2,733	$5.64 \cdot 10^{-6}$	7,746	$8.06 \cdot 10^{-4}$
742	MYFfamily.p2	2,733	$5.83 \cdot 10^{-6}$	5,225	$8.27 \cdot 10^{-4}$
521	EHF.p2	2,733	$7.23 \cdot 10^{-6}$	3,550	$1.02 \cdot 10^{-3}$
177	ATF6.p2	2,733	$1.04 \cdot 10^{-5}$	1,049	$1.46 \cdot 10^{-3}$
126	FOXD1,D2.p2	2,733	$4.58 \cdot 10^{-5}$	723	$6.35 \cdot 10^{-3}$
243	TFAP4.p2	2,733	$6.33 \cdot 10^{-5}$	1,554	$8.7 \cdot 10^{-3}$
323	GFI1.p2	2,733	$1.03 \cdot 10^{-4}$	2,152	$1.41 \cdot 10^{-2}$
186	HMX1.p2	2,733	$1.24 \cdot 10^{-4}$	1,160	$1.67 \cdot 10^{-2}$

Table 4.15: Motifs over-represented among genes downregulated in D1 dSPNs upon dopamine depletion and chronic high-dose levodopa treatment

Effects of levodopa treatment in D2 iSPNs

In contrast to the dramatic changes observed in D1 dSPNs, D2 iSPNs had relatively few genes with significant expression changes upon dopamine depletion followed by chronic levodopa treatment. Only 72 genes (84 probe sets) changed, 48 of which were upregulated (Table 4.16) and 24 of which were down-regulated (Table 4.17).

With chronic high-dose levodopa treatment, 415 genes (533 probe sets) had altered expression; 244 were up-regulated and 172 down-regulated. Of these, 198 (represented by 252 probe sets) also changed significantly in dSPNs, and of these, only 62 moved in opposing directions. Pathway analysis of genes altered in iSPNs by the high dose of levodopa showed effects on Kit receptor signaling, IL-3 signaling, and ErbB signaling (Table 4.20).

Probeset	Gene Symbol	Gene Description	P-value	log2 FC
1434815_a_at	Mapkapk3	mitogen-activated protein kinase-activated protein kinase 3	0.099851262	2.847595061
1455034_at	Nr4a2	nuclear receptor subfamily 4, group A, member 2	0.056702009	2.465317778
1455197_at	Rnd1	Rho family GTPase 1	0.092234273	2.392312708
1452318_a_at	Hspa1b	heat shock protein 1B	0.035576689	2.38809829
1424638_at	Cdkn1a	cyclin-dependent kinase inhibitor 1A (P21)	0.027108534	2.261879119
1421679_a_at	Cdkn1a	cyclin-dependent kinase inhibitor 1A (P21)	0.023167893	2.248976967
1425990_a_at	Nfatc2	nuclear factor of activated T cells, cytoplasmic, calcineurin dependent 2	0.058814247	2.057795463
1426037_a_at	Rgs16	regulator of G-protein signaling 16	0.085400331	1.989217558
1453590_at	Arl5b	ADP-ribosylation factor-like 5B	0.036932861	1.834816958
1416266_at	Pdyn	prodynorphin	0.085400331	1.774771097
1425671_at	Homer1	homer homolog 1 (Drosophila)	0.023167893	1.711474986
1453851_a_at	Gadd45g	growth arrest and DNA-damage-inducible 45 gamma	0.014460931	1.697999076
1452484_at	Car7	carbonic anhydrase 7	0.035576689	1.688634582
1436387_at			0.014460931	1.611295695
1435935_at	2410131K14Rik	RIKEN cDNA 2410131K14 gene	0.080136817	1.565084094
1437884_at	Arl5b	ADP-ribosylation factor-like 5B	0.048348403	1.499866027
1434973_at	Car7	carbonic anhydrase 7	0.035576689	1.43784128
1422053_at	Inhba	inhibin beta-A	0.086314323	1.381189726
1435458_at	Pim1	proviral integration site 1	0.05237781	1.374734542
1439947_at	Cyp11a1	cytochrome P450, family 11, subfamily a, polypeptide 1	0.080136817	1.285778927
1435071_at	Zfyve1	zinc finger, FYVE domain containing 1	0.070255481	1.257651434
1455166_at	Arl5b	ADP-ribosylation factor-like 5B	0.085400331	1.247261929
1428860_at	Them6	thioesterase superfamily member 6	0.078947348	1.237726912
1433657_at	Fam78a	family with sequence similarity 78, member A	0.075195168	1.233448861
1422134_at	Fosb	FBJ osteosarcoma oncogene B	0.056702009	1.208101367
1428710_at	Rit1	Ras-like without CAAX 1	0.097716313	1.167709426
1422697_s_at	Jarid2	jumonji, AT rich interactive domain 2	0.080959089	1.150739467
1418300_a_at	Mknk2	MAP kinase-interacting serine/threonine kinase 2	0.02991536	1.147490068
1427975_at	Rasl10a	RAS-like, family 10, member A	0.099851262	1.103679943
1454725_at	Tra2a	transformer 2 alpha homolog (Drosophila)	0.00671035	1.102548155
1450971_at	Gadd45b	growth arrest and DNA-damage-inducible 45 beta	0.085400331	1.100850129
1441814_s_at	Rpain	RPA interacting protein	0.083323329	1.068681234
1455175_at	Phf13	PHD finger protein 13	0.080136817	1.058879112
1451236_at	Rerg	RAS-like, estrogen-regulated, growth-inhibitor	0.058814247	0.99549417
1423747_a_at	Pdk1	pyruvate dehydrogenase kinase, isoenzyme 1	0.058814247	0.938113321
1422705_at	Pmepa1	prostate transmembrane protein, androgen induced 1	0.035576689	0.935846258
1427225_at	Epn2	epsin 2	0.058814247	0.922234272
1456943_a_at	Dbndd2	dysbindin (dystrobrevin binding protein 1) domain containing 2	0.061173154	0.897223989
1448663_s_at	Mvd	mevalonate (diphospho) decarboxylase	0.035576689	0.888308845
1435867_at	Jhdm1d	jumonji C domain-containing histone demethylase 1 homolog D (S. cerevisiae)	0.075195168	0.864124435
1416011_x_at	Ehd1	EH-domain containing 1	0.075195168	0.83483475
1460645_at	Chordc1	cysteine and histidine-rich domain (CHORD)-containing, zinc-binding protein 1	0.048348403	0.829426127
1452155_a_at	Ddx17	DEAD (Asp-Glu-Ala-Asp) box polypeptide 17	0.045433445	0.81823156
1439968_x_at	Dbndd2	dysbindin (dystrobrevin binding protein 1) domain containing 2	0.035576689	0.81503305
1415975_at	Carhsp1	calcium regulated heat stable protein 1	0.014460931	0.80939916
1434343_at	Zfp954	zinc finger protein 954	0.085400331	0.8086057
1423630_at	Cygb	cytoglobin	0.088261285	0.76513596
1439182_at	D17Wsu92e	DNA segment, Chr 17, Wayne State University 92, expressed	0.098528063	0.761631939
1451431_a_at	Dbndd2	dysbindin (dystrobrevin binding protein 1) domain containing 2	0.00671035	0.743137776
1417001_a_at	D4Wsu53e	DNA segment, Chr 4, Wayne State University 53, expressed	0.082879924	0.730972123
1424883_s_at	Srsf7	serine/arginine-rich splicing factor 7	0.033388958	0.728450219
1423795_at	Sfpq	splicing factor proline/glutamine rich (polypyrimidine tract binding protein associated)	0.058814247	0.704586102
1424033_at	Srsf7	serine/arginine-rich splicing factor 7	0.03394856	0.684168197
1429373_x_at	Crtc2	CREB regulated transcription coactivator 2	0.098528063	0.643749066
1429048_at	Bloc1s2a	biogenesis of lysosome-related organelles complex-1, subunit 2A	0.058814247	0.611227815
1428872_at	Msl1	male-specific lethal 1 homolog (Drosophila)	0.035576689	0.607980756

Table 4.16: Genes up-regulated in D2 iSPNs with chronic low-dose levodopa treatment

cProbeset	Gene Symbol	Gene Description	P-value	log2 FC
1416505_at	Nr4a1	nuclear receptor subfamily 4, group A, member 1	0.04177213	-1.324496405
1436094_at	Vgf	VGF nerve growth factor inducible	0.070255481	-1.27593533
1420444_at	Slc22a3	solute carrier family 22 (organic cation transporter), member 3	0.048348403	-1.270652074
1417110_at	Man1a	mannosidase 1, alpha	0.087356874	-1.240375708
1417111_at	Man1a	mannosidase 1, alpha	0.056702009	-1.232952356
1448746_at	Nbn	nibrin	0.062051282	-1.206035406
1428930_at	Tmem29	transmembrane protein 29	0.087356874	-1.1690701
1455199_at	AI429214	expressed sequence AI429214	0.099851262	-1.165657993
1450757_at	Cdh11	cadherin 11	0.080136817	-1.156063482
1450417_a_at	Rps20	ribosomal protein S20	0.099472702	-1.105160689
1434819_at	St6gal2	beta galactoside alpha 2,6 sialyltransferase 2	0.085400331	-1.098995816
1453102_at	Flrt3	fibronectin leucine rich transmembrane protein 3	0.080959089	-1.081171488
1455085_at	1700086L19Rik	RIKEN cDNA 1700086L19 gene	0.045433445	-1.079688257
1433987_at	Hpcal4	hippocalcin-like 4	0.014460931	-1.060800935
1430348_at	2900019E01Rik	RIKEN cDNA 2900019E01 gene	0.088023725	-0.899192146
1428545_at	Tmem248	transmembrane protein 248	0.087356874	-0.88908774
1450708_at	Scg2	secretogranin II	0.099025401	-0.865774896
1436610_at	Ankrd12	ankyrin repeat domain 12	0.035576689	-0.862325461
1450520_at	Cacng3	calcium channel, voltage-dependent, gamma subunit 3	0.058814247	-0.86226335
1426865_a_at	Ncam1	neural cell adhesion molecule 1	0.058814247	-0.857099277
1442905_at			0.056702009	-0.801688011
1436449_at	Pcdh11x	protocadherin 11 X-linked	0.029867334	-0.772491409
1423991_at	Nop14	NOP14 nucleolar protein	0.097716313	-0.770120552
1438407_at	Dsel	dermatan sulfate epimerase-like	0.085400331	-0.743531852
1420514_at	Tmem47	transmembrane protein 47	0.085400331	-0.720390174
1419247_at	Rgs2	regulator of G-protein signaling 2	0.01541297	-0.662706361
1419248_at	Rgs2	regulator of G-protein signaling 2	0.061173154	-0.618142253
1436135_at			0.092234273	-0.587673133

Table 4.17: Genes down-regulated in D2 iSPNs with chronic low-dose levodopa treatment

Probeset	Gene Symbol	Gene Description	P-value	log2 FC
1442754_at	C030013G03Rik	RIKEN cDNA C030013G03 gene	$2.61 \cdot 10^{-2}$	-2.3
1438427_at	Fam120b	family with sequence similarity 120, member B	$6.19 \cdot 10^{-2}$	-1.88
1435171_at	2810416G20Rik	RIKEN cDNA 2810416G20 gene	$7.86 \cdot 10^{-2}$	-1.82
1417782_at	Cers4	ceramide synthase 4	$5.31 \cdot 10^{-2}$	-1.8
1434817_s_at	Rprd2	regulation of nuclear pre-mRNA domain containing 2	$1.83 \cdot 10^{-2}$	-1.79
1416888_at	Fadd	Fas (TNFRSF6)-associated via death domain	$6.06 \cdot 10^{-2}$	-1.78
1457347_at	Ryr1	ryanodine receptor 1, skeletal muscle	$4.96 \cdot 10^{-2}$	-1.71
1453309_at	9330179D12Rik	RIKEN cDNA 9330179D12 gene	$3.23 \cdot 10^{-2}$	-1.69
AFFX-TransRecMur/X57349_5_at	Tfrc	transferrin receptor	$8.3 \cdot 10^{-2}$	-1.69
1426338_a_at	Ntng1	netrin G1	$6.26 \cdot 10^{-2}$	-1.65
1416805_at	Fam198b	family with sequence similarity 198, member B	$3.61 \cdot 10^{-2}$	-1.64
1447513_at	Kcnd3	potassium voltage-gated channel, Shal-related family, member 3	$9.35 \cdot 10^{-2}$	-1.63
1456096_at	6430573F11Rik	RIKEN cDNA 6430573F11 gene	$7.85 \cdot 10^{-2}$	-1.55
1451060_at	Gpr146	G protein-coupled receptor 146	$4.83 \cdot 10^{-2}$	-1.54
1460020_at	Ankrd11	ankyrin repeat domain 11	$7.17 \cdot 10^{-2}$	-1.54
1454581_at	5330425B07Rik	RIKEN cDNA 5330425B07 gene	$4.63 \cdot 10^{-2}$	-1.52
1438562_a_at	Ptpn2	protein tyrosine phosphatase, non-receptor type 2	$5.22 \cdot 10^{-2}$	-1.48
1422811_at	Slc27a1	solute carrier family 27 (fatty acid transporter), member 1	$8.76 \cdot 10^{-2}$	-1.47
1416505_at	Nr4a1	nuclear receptor subfamily 4, group A, member 1	$1.67 \cdot 10^{-2}$	-1.44
1459897_a_at	Sbsn	suprabasin	$2.97 \cdot 10^{-2}$	-1.37
1429162_at	1500015A07Rik	RIKEN cDNA 1500015A07 gene	$8.59 \cdot 10^{-2}$	-1.35
1429107_at	Ubr3	ubiquitin protein ligase E3 component n-recognin 3	$1.64 \cdot 10^{-3}$	-1.34
1420418_at	Syt2	synaptotagmin II	$3.82 \cdot 10^{-2}$	-1.33
1456397_at	Cdh4	cadherin 4	$4.83 \cdot 10^{-2}$	-1.33
1452661_at	Tfrc	transferrin receptor	$3.45 \cdot 10^{-3}$	-1.31
1431110_at	Plxdc2	plexin domain containing 2	$1.44 \cdot 10^{-2}$	-1.31
1439630_x_at	Sbsn	suprabasin	$2.91 \cdot 10^{-2}$	-1.29
1431102_at	Cep350	centrosomal protein 350	$9.13 \cdot 10^{-2}$	-1.29
1423499_at	Sncaip	synuclein, alpha interacting protein (synphilin)	$1.51 \cdot 10^{-2}$	-1.27
1428010_at	Timm9	translocase of inner mitochondrial membrane 9	$9.13 \cdot 10^{-2}$	-1.26
1454973_at	Atf7ip	activating transcription factor 7 interacting protein	$3.51 \cdot 10^{-2}$	-1.25
1452145_at	H6pd	hexose-6-phosphate dehydrogenase (glucose 1-dehydrogenase)	$8.44 \cdot 10^{-2}$	-1.24
1455854_a_at	Ssh1	slingshot homolog 1 (Drosophila)	$9.55 \cdot 10^{-2}$	-1.23
1417018_at	Efemp2	epidermal growth factor-containing fibulin-like extracellular matrix protein 2	$4.35 \cdot 10^{-2}$	-1.22
1420444_at	Slc22a3	solute carrier family 22 (organic cation transporter), member 3	$3.31 \cdot 10^{-2}$	-1.22
1443523_at	Fam135b	family with sequence similarity 135, member B	$2.24 \cdot 10^{-2}$	-1.21
1418817_at	Chmp1b	charged multivesicular body protein 1B	$2.83 \cdot 10^{-2}$	-1.2
1435396_at	Stxbp6	syntaxin binding protein 6 (amisyn)	$1.69 \cdot 10^{-2}$	-1.2
1434446_at	Insr	insulin receptor	$3.87 \cdot 10^{-2}$	-1.2
1435598_at	BB319198	expressed sequence BB319198	$1.32 \cdot 10^{-2}$	-1.19
1456144_at	Nav3	neuron navigator 3	$2.5 \cdot 10^{-2}$	-1.18
1416221_at	Fstl1	follistatin-like 1	$6.7 \cdot 10^{-2}$	-1.18
1431828_a_at	Synj2	synaptojanin 2	$5.3 \cdot 10^{-2}$	-1.17
1433959_at	Zmat4	zinc finger, matrin type 4	$9.72 \cdot 10^{-2}$	-1.17
1448746_at	Nbn	nibrin	$1.44 \cdot 10^{-2}$	-1.16
1442918_at	Nav3	neuron navigator 3	$5.78 \cdot 10^{-2}$	-1.16
1449286_at	Ntng1	netrin G1	$6.49 \cdot 10^{-2}$	-1.16
1418265_s_at	Irf2	interferon regulatory factor 2	$6.21 \cdot 10^{-2}$	-1.15
1428724_at	Pcf11	cleavage and polyadenylation factor subunit homolog (S. cerevisiae)	$1.92 \cdot 10^{-2}$	-1.15

Table 4.18: Top 50 genes down-regulated in D2 iSPNs with chronic high-dose levodopa treatment

Probeset	Gene Symbol	Gene Description	P-value	log2 FC
1455197_at	Rnd1	Rho family GTPase 1	$6.33 \cdot 10^{-3}$	3.46
1417160_s_at	Wfdc18	WAP four-disulfide core domain 18	$1.03 \cdot 10^{-2}$	3.42
1450750_a_at	Nr4a2	nuclear receptor subfamily 4, group A, member 2	$6.33 \cdot 10^{-3}$	3.4
1455034_at	Nr4a2	nuclear receptor subfamily 4, group A, member 2	$6.33 \cdot 10^{-3}$	3.11
1447863_s_at	Nr4a2	nuclear receptor subfamily 4, group A, member 2	$4.26 \cdot 10^{-3}$	2.52
1452318_a_at	Hspa1b	heat shock protein 1B	$1.07 \cdot 10^{-2}$	2.49
1421679_a_at	Cdkn1a	cyclin-dependent kinase inhibitor 1A (P21)	$1.64 \cdot 10^{-3}$	2.47
1434815_a_at	Mapkapk3	mitogen-activated protein kinase-activated protein kinase 3	$7.65 \cdot 10^{-2}$	2.39
1438967_x_at	Amhr2	anti-Mullerian hormone type 2 receptor	$8.92 \cdot 10^{-2}$	2.38
1434458_at	Fst	follistatin	$2.6 \cdot 10^{-2}$	2.38
1453590_at	Arl5b	ADP-ribosylation factor-like 5B	$2.98 \cdot 10^{-4}$	2.38
1449226_at	Hic1	hypermethylated in cancer 1	$1.87 \cdot 10^{-2}$	2.34
1450344_a_at	Ptger3	prostaglandin E receptor 3 (subtype EP3)	$2.63 \cdot 10^{-2}$	2.22
1424638_at	Cdkn1a	cyclin-dependent kinase inhibitor 1A (P21)	$9.16 \cdot 10^{-3}$	2.21
1426037_a_at	Rgs16	regulator of G-protein signaling 16	$1.41 \cdot 10^{-2}$	2.18
1449141_at	Fblim1	filamin binding LIM protein 1	$2.97 \cdot 10^{-2}$	2.16
1453851_a_at	Gadd45g	growth arrest and DNA-damage-inducible 45 gamma	$7.55 \cdot 10^{-4}$	2.14
1451452_a_at	Rgs16	regulator of G-protein signaling 16	$1.17 \cdot 10^{-2}$	2.13
1460269_at	Pnmt	phenylethanolamine-N-methyltransferase	$6.67 \cdot 10^{-2}$	2.1
1416266_at	Pdyn	prodynorphin	$1.83 \cdot 10^{-2}$	2.09
1460275_at	Gpr3	G-protein coupled receptor 3	$1.47 \cdot 10^{-2}$	1.94
1437884_at	Arl5b	ADP-ribosylation factor-like 5B	$1.33 \cdot 10^{-3}$	1.94
1418569_at	Fblim1	filamin binding LIM protein 1	$5.98 \cdot 10^{-2}$	1.89
1447930_at	Baz1a	bromodomain adjacent to zinc finger domain 1A	$2.4 \cdot 10^{-2}$	1.84
1426973_at	Gpr153	G protein-coupled receptor 153	$6.65 \cdot 10^{-3}$	1.84
1425671_at	Homer1	homer homolog 1 (Drosophila)	$1.42 \cdot 10^{-4}$	1.81
1436387_at			$1.04 \cdot 10^{-3}$	1.8
1431413_at	Ramp1	receptor (calcitonin) activity modifying protein 1	$7.67 \cdot 10^{-2}$	1.77
1458711_at			$4.72 \cdot 10^{-2}$	1.76
1417263_at	Ptgs2	prostaglandin-endoperoxide synthase 2	$4.08 \cdot 10^{-2}$	1.76
1452295_at	Pmepa1	prostate transmembrane protein, androgen induced 1	$1.32 \cdot 10^{-2}$	1.74
1438928_x_at	Ninj1	ninjurin 1	$8.77 \cdot 10^{-2}$	1.72
1451642_at	Kif1b	kinesin family member 1B	$6.75 \cdot 10^{-2}$	1.69
1455265_a_at	Rgs16	regulator of G-protein signaling 16	$2.26 \cdot 10^{-2}$	1.61
1421365_at	Fst	follistatin	$4.5 \cdot 10^{-2}$	1.58
1455166_at	Arl5b	ADP-ribosylation factor-like 5B	$1.89 \cdot 10^{-2}$	1.58
1439947_at	Cyp11a1	cytochrome P450, family 11, subfamily a, polypeptide 1	$1.07 \cdot 10^{-2}$	1.58
1424670_s_at	Zfyve21	zinc finger, FYVE domain containing 21	$4.83 \cdot 10^{-2}$	1.56
1448694_at	Jun	Jun oncogene	$7.33 \cdot 10^{-2}$	1.55
1435999_at	Spink8	serine peptidase inhibitor, Kazal type 8	$7.73 \cdot 10^{-2}$	1.55
1435872_at			$8.48 \cdot 10^{-2}$	1.55
1422256_at	Sstr2	somatostatin receptor 2	$2.19 \cdot 10^{-2}$	1.54
1435935_at	2410131K14Rik	RIKEN cDNA 2410131K14 gene	$2.77 \cdot 10^{-2}$	1.54
1436018_at	Mex3a	mex3 homolog A (C. elegans)	$6.06 \cdot 10^{-2}$	1.54
1424581_at	Stac2	SH3 and cysteine rich domain 2	$5.52 \cdot 10^{-2}$	1.5
1451463_at	Prr5	proline rich 5 (renal)	$8.28 \cdot 10^{-2}$	1.49
1417262_at	Ptgs2	prostaglandin-endoperoxide synthase 2	$5.24 \cdot 10^{-2}$	1.48
1416554_at	Pdlim1	PDZ and LIM domain 1 (elfin)	$1.46 \cdot 10^{-2}$	1.46
1440001_at	Rian	RNA imprinted and accumulated in nucleus	$8.77 \cdot 10^{-2}$	1.46

Table 4.19: Top 50 genes up-regulated in D2 iSPNs with chronic high-dose levodopa treatment

Set Name	Matches	Size of Set	p-value	Bonf adj p-val	b-h FDR adj p-val
Myometrial Relaxation and Contraction Pathways	22	164	$4.79 \cdot 10^{-8}$	$4.98 \cdot 10^{-6}$	$4.98 \cdot 10^{-6}$
Kit Receptor Signaling Pathway	10	68	$1.02 \cdot 10^{-4}$	$1.06 \cdot 10^{-2}$	$5.28 \cdot 10^{-3}$
IL-3 Signaling Pathway	11	102	$7.47 \cdot 10^{-4}$	$7.77 \cdot 10^{-2}$	$2.59 \cdot 10^{-2}$
Calcium Regulation in the Cardiac Cell	13	154	$2.53 \cdot 10^{-3}$	0.26	$6.58 \cdot 10^{-2}$
ErbB signaling pathway	6	46	$4.63 \cdot 10^{-3}$	0.48	$9.64 \cdot 10^{-2}$
TGF-beta Receptor Signaling Pathway	12	155	$7.23 \cdot 10^{-3}$	0.75	0.11
EGFR1 Signaling Pathway	13	176	$7.79 \cdot 10^{-3}$	0.81	0.11
PluriNetWork	19	303	$8.7 \cdot 10^{-3}$	0.9	0.11
Splicing factor NOVA regulated synpatic proteins	5	42	$1.39 \cdot 10^{-2}$	1	0.16
T Cell Receptor Signaling Pathway	10	134	$1.71 \cdot 10^{-2}$	1	0.18
Leptin Insulin Overlap	3	17	$1.91 \cdot 10^{-2}$	1	0.18
Apoptosis	7	83	$2.37 \cdot 10^{-2}$	1	0.21
Signaling of Hepatocyte Growth Factor Receptor	4	34	$2.81 \cdot 10^{-2}$	1	0.21
Glycogen Metabolism	4	34	$2.81 \cdot 10^{-2}$	1	0.21
TGF Beta Signaling Pathway	5	52	$3.23 \cdot 10^{-2}$	1	0.21
IL-5 Signaling Pathway	6	70	$3.27 \cdot 10^{-2}$	1	0.21
Diurnally regulated genes with circadian orthologs	5	55	$3.98 \cdot 10^{-2}$	1	0.24
Insulin Signaling	10	157	$4.43 \cdot 10^{-2}$	1	0.26
Androgen Receptor Signaling Pathway	8	118	$5.02 \cdot 10^{-2}$	1	0.26
mRNA processing	24	494	$5.58 \cdot 10^{-2}$	1	0.26

Table 4.20: Wikipathways pathways over-represented among genes changed in D2 dSPNs upon dopamine depletion and chronic high-dose levodopa treatment

4.5.3 Levodopa dose-dependent genes

To prioritize genes most likely to be relevant to emergence of levodopa-induced dyskinesia, we applied two complementary procedures. First, we directly compared expression between the high-dose group (which develop more severe dyskinesias) to the low-dose group. In addition, we used the AIM scores that quantify dyskinesia phenotype of individual mice, and fit linear models relating probe-set expression, dose, and AIM score, and compared those models to assess the signficance of the relationship between expression, dose, and dyskinesia (Figure 4.4).

In dSPNs, 298 genes (409 probe sets) had significant positive dose-responses and were associated with more severe AIMs, whereas 192 genes (230 probe sets) were anti-correlated with dose and AIMs. No genes had statistically significant excess correlation with AIMs, after accounting for the effect of dose and multiple testing adjustments; the variability of AIMs across animals was much smaller than the effect of dose. In iSPNs, no genes had significant correlations with either dose or AIMs after multiple testing adjustment.

The genes in dSPNs that had the most significant correlations between expression and dose (Table 4.21 and Figure 4.8) were *Gpr39*, *Fndx9*, *Cstb*, *Trh*, *Srxn1*, *Ier3*, *Tinf2*, *Cdk11b*, *Nr4a2* (Nurr), *Itch*, *Scp*, and *Fos1* (Fra-1). Of these, only *Trh* had been previously linked to levodopa-induced dyskinesia [17]. *Fosb* has previously been implicated in dyskinesia [**Andersson1999**], and Fosl may have a similar role. *Ier3* (positively correlated) encodes an inhibitor of protein phosphatase 2A-dependent ERK dephosphorylation, and thus may enhance ERK signaling. Itch, whose expression is anticorrelated to AIMs, is an E3 ubiquitin ligase that regulates c-Jun (Jun) levels. Surprisingly, no genes linked to CREB signaling were among those with the most significant correlation between expression and dose.

Probeset	Gene Symbol	Gene Description	D1 high vs low bh pval	D1 high vs low FC
1432260_at	Gpr39	G protein-coupled receptor 39	$8.04 \cdot 10^{-5}$	0.83
1436484_at	Fndc9	fibronectin type III domain containing 9	$1.95 \cdot 10^{-4}$	0.79
1422507_at	Cstb	cystatin B	$1.95 \cdot 10^{-4}$	1.08
1418756_at	Trh	thyrotropin releasing hormone	$3.49 \cdot 10^{-4}$	1.92
1451680_at	Srxn1	sulfiredoxin 1 homolog (S. cerevisiae)	$3.49 \cdot 10^{-4}$	0.67
1452418_at			$3.55 \cdot 10^{-4}$	1.24
1419647_a_at	Ier3	immediate early response 3	$3.55 \cdot 10^{-4}$	1.12
1437166_at	Tinf2	Terf1 (TRF1)-interacting nuclear factor 2	$3.71 \cdot 10^{-4}$	1.78
1442887_at			$5.82 \cdot 10^{-4}$	1.51
1418841_s_at	Cdk11b	cyclin-dependent kinase 11B	$6.11 \cdot 10^{-4}$	0.86
1450750_a_at	Nr4a2	nuclear receptor subfamily 4, group A, member 2	$6.88 \cdot 10^{-4}$	1.92
1415769_at	Itch	itchy, E3 ubiquitin protein ligase	$6.88 \cdot 10^{-4}$	−0.59
1447863_s_at	Nr4a2	nuclear receptor subfamily 4, group A, member 2	$6.88 \cdot 10^{-4}$	1.55
1449686_s_at	Scp2	sterol carrier protein 2, liver	$6.88 \cdot 10^{-4}$	−0.67
1417488_at	Fosl1	fos-like antigen 1	$6.88 \cdot 10^{-4}$	1.08
1418350_at	Hbegf	heparin-binding EGF-like growth factor	$6.94 \cdot 10^{-4}$	0.79
1460373_a_at	Setd4	SET domain containing 4	$7.46 \cdot 10^{-4}$	1.15
1422256_at	Sstr2	somatostatin receptor 2	$7.46 \cdot 10^{-4}$	1.04
1456886_at	Zfp839	zinc finger protein 839	$7.46 \cdot 10^{-4}$	1.3
1417262_at	Ptgs2	prostaglandin-endoperoxide synthase 2	$7.57 \cdot 10^{-4}$	2.05
1452417_x_at			$7.64 \cdot 10^{-4}$	3.08
1419069_at	Rabgef1	RAB guanine nucleotide exchange factor (GEF) 1	$7.87 \cdot 10^{-4}$	0.74
1451163_at	Tinf2	Terf1 (TRF1)-interacting nuclear factor 2	$8.43 \cdot 10^{-4}$	1.13
1451039_at	Nop9	NOP9 nucleolar protein	$8.9 \cdot 10^{-4}$	0.85
1437111_at	Zc3h12c	zinc finger CCCH type containing 12C	$8.9 \cdot 10^{-4}$	0.84
1456819_at	Nrn1l	neuritin 1-like	$8.99 \cdot 10^{-4}$	1.12
1417947_at	Pcna	proliferating cell nuclear antigen	$9.46 \cdot 10^{-4}$	0.73
1435625_at	Entpd7	ectonucleoside triphosphate diphosphohydrolase 7	$9.48 \cdot 10^{-4}$	0.83
1450749_a_at	Nr4a2	nuclear receptor subfamily 4, group A, member 2	$9.48 \cdot 10^{-4}$	1.76
1453029_s_at	Ubr3	ubiquitin protein ligase E3 component n-recognin 3	$1.08 \cdot 10^{-3}$	−0.69
1417487_at	Fosl1	fos-like antigen 1	$1.08 \cdot 10^{-3}$	1.82
1451534_at	Scgn	secretagogin, EF-hand calcium binding protein	$1.08 \cdot 10^{-3}$	0.94
1452463_x_at			$1.08 \cdot 10^{-3}$	2.15
1428301_at			$1.13 \cdot 10^{-3}$	1.27
1416271_at	Perp	PERP, TP53 apoptosis effector	$1.13 \cdot 10^{-3}$	1.31
1420858_at	Pkia	protein kinase inhibitor, alpha	$1.13 \cdot 10^{-3}$	−0.6
1455034_at	Nr4a2	nuclear receptor subfamily 4, group A, member 2	$1.13 \cdot 10^{-3}$	1.49
1460275_at	Gpr3	G-protein coupled receptor 3	$1.13 \cdot 10^{-3}$	0.91
1425661_at	Cdadc1	cytidine and dCMP deaminase domain containing 1	$1.13 \cdot 10^{-3}$	−0.89
1415855_at	Kitl	kit ligand	$1.41 \cdot 10^{-3}$	0.78
1427931_s_at	Pdxk	pyridoxal (pyridoxine, vitamin B6) kinase	$1.55 \cdot 10^{-3}$	0.73
1426208_x_at	Plagl1	pleiomorphic adenoma gene-like 1	$1.55 \cdot 10^{-3}$	0.69
1451714_a_at	Map2k3	mitogen-activated protein kinase kinase 3	$1.55 \cdot 10^{-3}$	0.93
1427660_x_at			$1.63 \cdot 10^{-3}$	2.96
1428157_at	Gng2	guanine nucleotide binding protein (G protein), gamma 2	$1.63 \cdot 10^{-3}$	−1.44
1427455_x_at			$1.68 \cdot 10^{-3}$	2.92
1418778_at	Ccdc109b	coiled-coil domain containing 109B	$1.7 \cdot 10^{-3}$	1.2
1430029_a_at	Tspan31	tetraspanin 31	$1.7 \cdot 10^{-3}$	0.62
1419283_s_at	Tns1	tensin 1	$1.78 \cdot 10^{-3}$	1.03
1460455_at	Ubr3	ubiquitin protein ligase E3 component n-recognin 3	$1.78 \cdot 10^{-3}$	−0.82
1441165_s_at	Clstn2	calsyntenin 2	$1.8 \cdot 10^{-3}$	0.63
1434496_at	Plk3	polo-like kinase 3	$1.8 \cdot 10^{-3}$	1.01

Table 4.21: Probesets with greatest dependence on levodopa dose in D1 dSPNs, sorted by significance of difference betwen high- and low-dose groups.

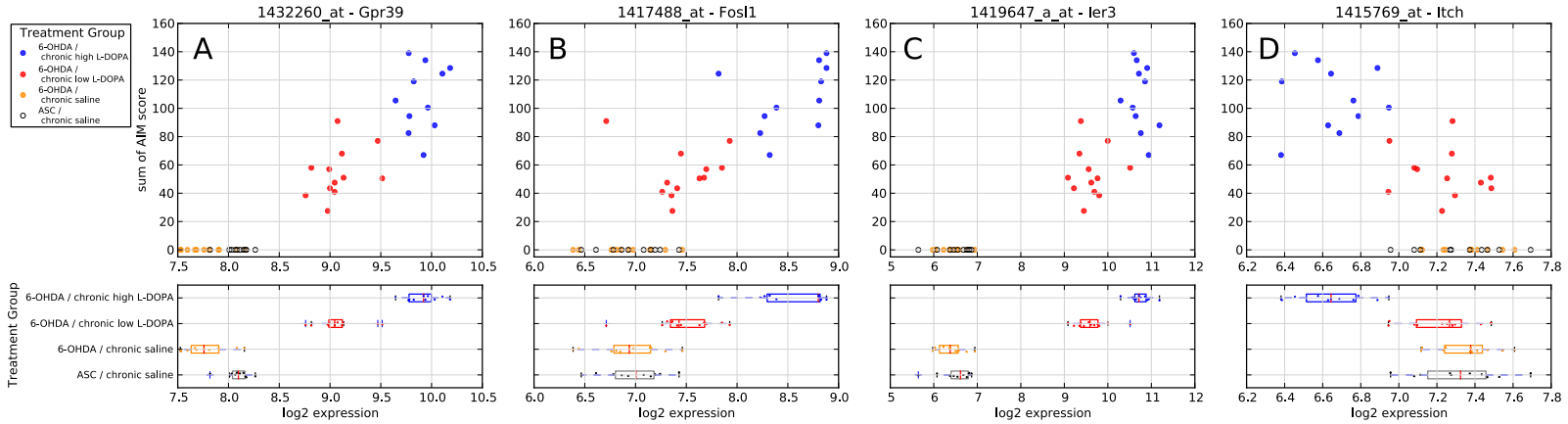


Figure 4.8: Representative examples of dose-dependent, AIM-correlated probe-sets with different patterns of responses to dopamine depletion and levodopa treatments in Drd1a dSPNs. (Upper) Scatterplots of total AIM score vs. log2 gene expression. Each point represents the gene expression measurement and AIM score from a single mouse. Colors indicate treatment groups (see key, Upper Left). (Lower) Box-plots summarizing gene expression across treatment groups. (A) Gpr39 (1432260_at): Expression decreases with dopamine depletion, increases significantly with chronic levodopa treatment, and expression depends on levodopa dose. (B) Fosl1 (1417488_at): Expression is unchanged by dopamine depletion, increases significantly with chronic levodopa, and expression depends on levodopa dose. (C) Ier3 (1419647_a_at): Expression is unchanged by dopamine depletion, increases dramatically with chronic levodopa treatment, and depends on levodopa dose. (D) Itch (1415769_at): Expression is unchanged by dopamine depletion, decreases significantly with levodopa treatment, and depends on levodopa dose

MAPK signaling was among the pathways most significantly enriched within the genes with expression correlated with dose in dSPNs (Table 4.22), while cell-cycle control, DNA replication, and B-cell receptor signaling were over-represented within anti-correlated genes (Table 4.23). Motifs overrepresented in promoters of dose-correlated genes included CREB, AP-1, and ERK-dependent motifs (Table **??** and 4.24).

Set Name	Matches	Genes in Overlap	Size of Set	p-value	Bonf adj p-val	b-h FDR adj p-val
MAPK signaling pathway	22	TGFB1, NGF, TGFB3, ACVR1C, RAP1B, IL1A, HSPA5, GADD45A, JUN, MAP3K4, MAPK4, DUSP5, DUSP4, DUSP1, SRF, NR4A1, JUND, NFKB1, DUSP10, RPS6KA3, PRKACA, ATF4	160	$1.52 \cdot 10^{-6}$	$1.78 \cdot 10^{-4}$	$1.78 \cdot 10^{-4}$
TGF-beta Receptor Signaling Pathway	18	XPO1, TGFB1, ACVRL1, SDC2, FOSB, STK11, JUN, JUND, AP2B1, MAP2K3, FOXO1, TFDP2, JUNB, FOXO4, SMAD7, TGFB3, CDKN1A, CTNNB1	155	$1.29 \cdot 10^{-4}$	$1.51 \cdot 10^{-2}$	$5.68 \cdot 10^{-3}$
Hypertrophy Model	6	NR4A3, JUND, HBEGF, VEGFA, DUSP14, IL1A	20	$1.46 \cdot 10^{-4}$	$1.7 \cdot 10^{-2}$	$5.68 \cdot 10^{-3}$
Senescence and Autophagy	10	TGFB1, FN1, CDKN1B, CDKN1A, SH3GLB1, IL6ST, MAP2K3, MAP1LC3A, SERPINB2, ING1	60	$2.25 \cdot 10^{-4}$	$2.63 \cdot 10^{-2}$	$6.57 \cdot 10^{-3}$
Cholesterol Biosynthesis	5	IDI1, MVK, CYP51, PMVK, HMGCR	16	$4.33 \cdot 10^{-4}$	$5.06 \cdot 10^{-2}$	$1.01 \cdot 10^{-2}$
EGFR1 Signaling Pathway	17	ERRFI1, RPS6KA3, EPS8, DUSP1, JUN, JUND, MAP3K4, ARF4, KRT17, SPRY2, FOXO1, JAK2, PRKAR1A, ELK4, MAP2K3, PITPNA, STXBP1	176	$1.63 \cdot 10^{-3}$	0.19	$3.17 \cdot 10^{-2}$
Diurnally regulated genes with circadian orthologs	8	BTG1, PIGF, IDI1, ERC2, AZIN1, GSTM5, PER1, GFRA1	55	$2.34 \cdot 10^{-3}$	0.27	$3.9 \cdot 10^{-2}$
Integrin-mediated cell adhesion	11	ITGB4, ARHGEF7, RAP1B, ITGA6, MAP2K3, VASP, ITGAV, MAPK4, PAK6, TNS1, PAK3	97	$3.05 \cdot 10^{-3}$	0.36	$4.46 \cdot 10^{-2}$
Adipogenesis	13	TGFB1, KLF5, KLF6, GADD45A, GADD45B, NRIP1, PPARGC1A, HIF1A, IL6ST, CTNNB1, MEF2B, FOXO1, IRS2	135	$5.62 \cdot 10^{-3}$	0.66	$7.3 \cdot 10^{-2}$
Insulin Signaling	14	MAP2K3, RPS6KA3, SRF, RPS6KA6, JUN, ARF6, SLC2A1, STXBP1, FOXO1, MAP3K4, MAPK4, MAP4K5, PTPN1, RHEB	157	$8.1 \cdot 10^{-3}$	0.95	$9.11 \cdot 10^{-2}$
Dopminergic Neurogenesis	5	NR4A2, TGFB1, TH, RET, PITX3	30	$8.56 \cdot 10^{-3}$	1	$9.11 \cdot 10^{-2}$
MicroRNAs in cardiomyocyte hypertrophy	9	TGFB1, FZD2, IL6ST, CTNNB1, MAP2K3, MAPK4, IKBKE, NFKB1, CDK9	85	$1.09 \cdot 10^{-2}$	1	0.11
PluriNetWork	22	TGFB1, ACVR1C, CDKN1A, SMAD7, JARID2, ERCC5, HIF1A, IL6ST, CTNNB1, EP400, SMARCA5, STK40, GADD45A, CDK2AP1, KLF5, PERP, KDM6B, NFKB1, PIM3, PIM1, PRKACA, CTBP2	303	$1.23 \cdot 10^{-2}$	1	0.11
Alanine and aspartate metabolism	3	PCX, GAD1, GAD2	12	$1.31 \cdot 10^{-2}$	1	0.11
Circadian Exercise	7	BTG1, PIGF, IDI1, ERC2, AZIN1, PER1, GFRA1	61	$1.56 \cdot 10^{-2}$	1	0.12
G1 to S cell cycle control	7	PCNA, CDKN2C, CDKN1B, CDKN1A, MCM6, GADD45A, TFDP2	62	$1.7 \cdot 10^{-2}$	1	0.12
Wnt Signaling Pathway NetPath	10	FZD2, CSNK1A1, FZD8, TBP, JUN, FRAT2, CTNNB1, CSNK1D, BCL9, CTBP2	109	$1.93 \cdot 10^{-2}$	1	0.12
Biogenic Amine Synthesis	3	GAD2, GAD1, TH	14	$2.04 \cdot 10^{-2}$	1	0.12
FAS pathway and Stress induction of HSP regulation	5	IL1A, ARHGDIB, JUN, MAPKAPK3, CFLAR	37	$2.05 \cdot 10^{-2}$	1	0.12
IL-1 Signaling Pathway	5	IL1A, SQSTM1, PELI1, IRAK3, NFKB1	37	$2.05 \cdot 10^{-2}$	1	0.12

Table 4.22: Wikipathways pathways over-represented among dose-dependent *positively* correlated genes in D1

Set Name	Overlap	Genes in overlap	Genes in group	p-value	bonferroni	b-h fdr adj pval
Cell cycle	11	HDAC2, SMAD4, CCND2, ORC2, ORC3, GSK3B, MCM7, ORC6, PTTG1, WEE1, CCNH	89	$1.84 \cdot 10^{-4}$	$1.78 \cdot 10^{-2}$	$9.97 \cdot 10^{-3}$
G1 to S cell cycle control	9	ORC6, CCND2, ORC2, ORC3, RPA2, MCM7, CDK7, WEE1, CCNH	62	$2.05 \cdot 10^{-4}$	$1.99 \cdot 10^{-2}$	$9.97 \cdot 10^{-3}$
DNA Replication	7	ORC6, RFC1, ORC2, ORC3, RPA2, MCM7, POLD3	42	$4.4 \cdot 10^{-4}$	$4.27 \cdot 10^{-2}$	$1.42 \cdot 10^{-2}$
B Cell Receptor Signaling Pathway	14	PTK2, PIP5K1C, CCND2, CREB1, GSK3B, IKBKB, PDK2, PLEKHA1, PIK3R1, CDK7, PPP3CA, WAS, RASA1, PIP4K2C	163	$1.18 \cdot 10^{-3}$	0.11	$2.85 \cdot 10^{-2}$
Diurnally regulated genes with circadian orthologs	6	ARNTL, CLOCK, SUMO3, CRY1, VAPA, UGP2	55	$9.85 \cdot 10^{-3}$	0.96	0.17
Regulation of Actin Cytoskeleton	11	PTK2, CYFIP2, PIP5K1C, NRAS, MRAS, CHRM1, MAPK6, PIK3R1, FGF11, WAS, PIP4K2C	149	$1.16 \cdot 10^{-2}$	1	0.17
Cardiac Hypertrophy: miR-208	2	MED13, MYH7	6	$1.53 \cdot 10^{-2}$	1	0.17
Circadian Exercise	6	ARNTL, CLOCK, SUMO3, CRY1, VAPA, UGP2	61	$1.6 \cdot 10^{-2}$	1	0.17
TNF-alpha NF-kB Signaling Pathway	13	NR2C2, HDAC2, TRAF3, MCM7, GSK3B, IKBKB, PTK2, KPNA3, SMARCE1, PML, CYLD, CRADD, BCL7A	198	$1.6 \cdot 10^{-2}$	1	0.17
Toll Like Receptor signaling	4	IKBKB, NR2C2, TRAF3, RALBP1	33	$2.35 \cdot 10^{-2}$	1	0.21
Delta-Notch Signaling Pathway	7	HDAC2, ITCH, SMAD4, GSK3B, SNW1, PIK3R1, HIVEP3	85	$2.36 \cdot 10^{-2}$	1	0.21
Circadian clock tutorial_CarlosBorroto	1	CLOCK	1	$3.34 \cdot 10^{-2}$	1	0.26
methylation	2	COMT, MAT2B	9	$3.44 \cdot 10^{-2}$	1	0.26
Translation Factors	5	EIF4E, EEF1A1, EEF1D, EIF6, EIF2S2	57	$4.14 \cdot 10^{-2}$	1	0.29
PluriNetWork	16	ZMYM2, GATAD2B, HDAC2, ETV5, LEO1, SMAD4, NR2F1, CREB1, GSK3B, KDM4C, NEDD4L, CTR9, PHF17, IPO9, PML, CTCF	303	$4.89 \cdot 10^{-2}$	1	0.32
estrogen signalling	6	HDAC2, CREB1, IKBKB, POLR2G, CDK7, CCNH	81	$5.38 \cdot 10^{-2}$	1	0.33
EGFR1 Signaling Pathway	10	NCK2, RALBP1, ITCH, EEF1A1, NRAS, CREB1, PIK3R1, SH3BGRL, RASA1, SH3GL2	176	$7.18 \cdot 10^{-2}$	1	0.39
Calcium Regulation in the Cardiac Cell	9	RGS20, GNB4, CAMK2G, PKIA, GNAZ, GNG2, CHRM1, RGS2, GNA11	154	$7.39 \cdot 10^{-2}$	1	0.39
TGF-beta Receptor Signaling Pathway	9	SNW1, SMAD4, CAMK2G, NFYA, NFYC, MEF2A, PIK3R1, ZEB1, CTCF	155	$7.63 \cdot 10^{-2}$	1	0.39
Alpha6-Beta4 Integrin Signaling Pathway	5	EIF4E, LAMB1, PTK2, EIF6, PIK3R1	69	$8.07 \cdot 10^{-2}$	1	0.39

Table 4.23: Wikipathways pathways over-represented among dose-dependent *negatively* correlated genes in D1 dSPNs

Overlap	Motif	Genes in group	P-value	Total motif occurrences	BH adj p-val
75	ATF5_CREB3.p2	298	$1.33 \cdot 10^{-12}$	2,370	$2.31 \cdot 10^{-10}$
63	JUN.p2	298	$2.54 \cdot 10^{-12}$	1,812	$4.4 \cdot 10^{-10}$
246	SP1.p2	298	$1.34 \cdot 10^{-11}$	14,315	$2.3 \cdot 10^{-9}$
189	TFDP1.p2	298	$2.08 \cdot 10^{-9}$	10,211	$3.55 \cdot 10^{-7}$
225	KLF4.p3	298	$3.24 \cdot 10^{-9}$	13,075	$5.51 \cdot 10^{-7}$
221	PATZ1.p2	298	$7.31 \cdot 10^{-9}$	12,837	$1.24 \cdot 10^{-6}$
199	TFAP2A,C.p2	298	$1.9 \cdot 10^{-8}$	11,209	$3.19 \cdot 10^{-6}$
56	EP300.p2	298	$6.83 \cdot 10^{-8}$	1,963	$1.14 \cdot 10^{-5}$
225	MAZ.p2	298	$1.83 \cdot 10^{-7}$	13,533	$3.03 \cdot 10^{-5}$
160	TFAP2B.p2	298	$5.29 \cdot 10^{-7}$	8,716	$8.73 \cdot 10^{-5}$
37	ATF4.p2	298	$6.52 \cdot 10^{-7}$	1,130	$1.07 \cdot 10^{-4}$
131	HIC1.p2	298	$1.32 \cdot 10^{-6}$	6,807	$2.15 \cdot 10^{-4}$
140	PAX5.p2	298	$1.05 \cdot 10^{-5}$	7,682	$1.69 \cdot 10^{-3}$
32	FOS_FOSB,L1_JUNB,D.p2	298	$1.45 \cdot 10^{-5}$	1,042	$2.33 \cdot 10^{-3}$
51	PAX2.p2	298	$4.06 \cdot 10^{-5}$	2,127	$6.48 \cdot 10^{-3}$
27	ATF2.p2	298	$4.62 \cdot 10^{-5}$	858	$7.32 \cdot 10^{-3}$
142	TCF4_dimer.p2	298	$7.08 \cdot 10^{-5}$	8,090	$1.11 \cdot 10^{-2}$
101	MAFB.p2	298	$7.36 \cdot 10^{-5}$	5,292	$1.15 \cdot 10^{-2}$
35	CREB1.p2	298	$1.02 \cdot 10^{-4}$	1,312	$1.58 \cdot 10^{-2}$
136	SPI1.p2	298	$1.19 \cdot 10^{-4}$	7,746	$1.83 \cdot 10^{-2}$
156	GTF2I.p2	298	$1.81 \cdot 10^{-4}$	9,237	$2.75 \cdot 10^{-2}$
90	NHLH1,2.p2	298	$3.1 \cdot 10^{-4}$	4,758	$4.64 \cdot 10^{-2}$

Table 4.24: Motifs over-represented among dose-dependent up-regulated genes in dSPNs

Overlap	Motif	Genes in group	P-value	Total motif occurrences	BH adj p-val
90	ELF1,2,4.p2	192	$6.01 \cdot 10^{-5}$	7,322	$9.93 \cdot 10^{-3}$
115	TFDP1.p2	192	$1.13 \cdot 10^{-4}$	10,211	$1.85 \cdot 10^{-2}$
90	PAX5.p2	192	$3.99 \cdot 10^{-4}$	7,682	$6.34 \cdot 10^{-2}$

Table 4.25: Motifs over-represented among dose-dependent down-regulated genes in dSPNs

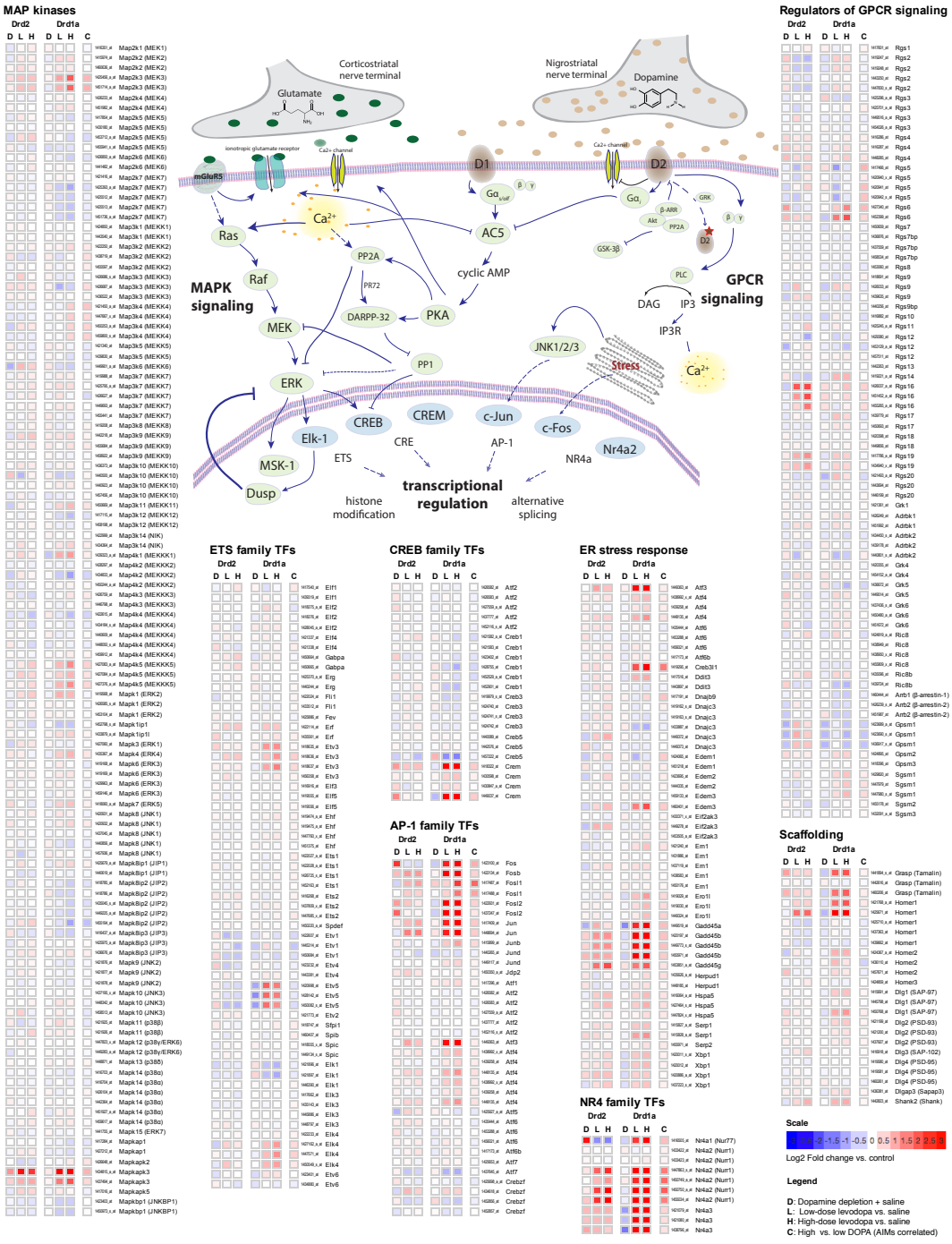


Figure 4.9: Expression of Genes in Key Pathways involved in LID and MSN function. Pathway cartoon was adapted from [18]

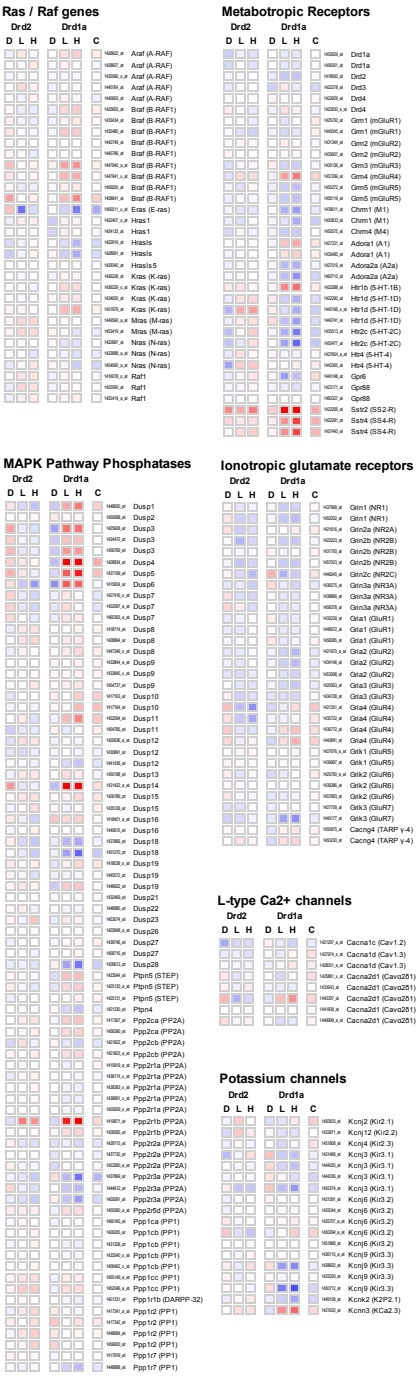


Figure 4.9: continued

Dataset 1	Probe-sets with significant changes (Benjamini–Hochberg adjusted P values < 0.10) of at least 1.5-fold up or down in dSPNs following dopamine depletion and chronic saline treatment
Dataset 2	Probesets for genes with significant differential expression upon dopamine depletion in both dSPNs and iSPNs
Dataset 3	Probe-sets with significant changes (Benjamini–Hochberg adjusted P values < 0.10) of at least 1.5-fold up or down in iSPNs following dopamine depletion and chronic saline treatment
Dataset 4	Enriched pathways from Wikipathways in dopamine-depleted dSPNs
Dataset 5	Enriched pathways from Wikipathways in dopamine-depleted iSPNs
Dataset 6	Probe-sets with significant (Benjamini–Hochberg adjust P value < 0.10) changes of at least 1.5-fold up or down in dSPNs with low-dose chronic levodopa following dopamine depletion
Dataset 7	Enriched Wikipathways pathways among genes differentially expressed with low-dose levodopa in dSPNs
Dataset 8	Probe-sets with significant changes (Benjamini–Hochberg adjusted P value < 0.10) of at least 1.5-fold up or down in dSPNs with high-dose chronic levodopa treatment following dopamine depletion
Dataset 9	Probe-sets with significant (Benjamini–Hochberg adjustd P value < 0.10) changes of at least 1.5-fold up or down in dSPNs with high-dose levodopa that also had any statistically significant changes with low-dose levodopa
Dataset 10	Wikipathways pathways enriched among genes altered by high-dose levodopa in dSPNs
Dataset 11	Motifs enriched within promoter regions (as predicted by SwissRegulon) of genes with altered expression in dSPNs upon chronic high-dose levodopa treatment
Dataset 12	Differentially expressed probe-sets in iSPNs with low-dose chronic levodopa
Dataset 13	Differentially expressed probe-sets (Benjamini–Hochberg adjusted P value < 0.10, at least 1.5-fold up or down) in iSPNs with chronic high-dose Levodopa
Dataset 14	Probe-sets significantly altered (≥ 1.5 fold-change up or down) by high-dose levodopa in both dSPNs and iSPNs
Dataset 15	Wikipathways pathways enriched in genes altered by high-dose levodopa in iSPNs
Dataset 16	Probe-sets differentially expressed between high- and low-dose levodopa
Dataset 17	Wikipathways pathways enriched among dose-dependent positively correlated genes in dSPNs
Dataset 18	Wikipathways pathways enriched among dose-dependent negatively correlated genes in dSPNs
Dataset 19	Motifs enriched within promoter regions of genes with significantly altered expression in dSPNs between chronic high-dose and chronic low-dose levodopa treatment
Dataset 20	Complete table of statistics for all probe-sets and comparisons performed

Table 4.26: List of data tables containing major results of PD LID analysis

4.6 Discussion

4.6.1 Homeostatic regulation of signaling

Many of the expression changes observed can be interpreted as homeostatic changes in response to activation of pathways previously known to be activated by the dopamine receptors (see Figure 4.9). Although there was no evidence of of direct modulation of levels of the dopamine receptors themselves, several genes likely to directly regulate G-protein signaling were up-regulated upon chronic L-DOPA treatment, including *Rgs16* in iSPNs and *Rgs6* in dSPNs. *Rgs6* has a GTPase-activating function towards $G_{\alpha/olf}$ (the principal G-protein coupled to Drd1a receptors in dSPNs) and so would be exerting a homeostatic down-regulation of Drd1a signaling in dSPNs. *Rgs16* has a GTPase-activating function towards $G_{i/q}$, and therefore would negatively regulate Drd2 signaling.

4.6.2 Changes to other receptors

Among the metabotrophic receptors, *Sstr2* and *Sstr4* were up-regulated in dSPNs. The somatostatin receptors are G_i coupled, and their signaling is likely antagonistic to the increased Drd1a signaling in dSPNs. Since there are already somatostatin receptor ligands in clincal use and development for other indications, these receptors may present an attractive potential target for modulating SPN activities.

4.6.3 Genes most associated with dose and development of dyskinesias

Many of the genes with the greatest correlations to dose - *Gpr39*, *Fndx9*, *Cstb*, *Trh*, *Srxn1*, *Ier3*, *Tinf2*, *Cdk11b*, *Nr4a2* (Nurr), *Itch*, *Scp*, and *Fos1* (Fra-1) - had not been previously connected to dyskinesia. Of these, *Gpr39*, being a GPCR in the grelin family, may be a potential therapeutic target. *Nr4a2* (Nurr1), an orphan nuclear receptor known to be involved in development and maintenance of dopaminergic neurons [96], has dramatic expression changes in both cell types and may be regulating some of the transcriptional changes observed. Because of its role in dopaminergic neuron development, Nr4a2 has already been considered as a potential target in PD [28], and this role is reinforced by our observations.

Part III

Transcriptional Dysregulation in Huntington's Disease

Chapter 5

Background: Transcriptional Dysregulation in Huntington Disease

This chapter provides a brief review of the biology of Huntington Disease, and the mouse models of HD that have been developed. We discuss the importance and potential mechanisms of transcriptional dysregulation in HD, and define the major questions to be investigated in chapters 6, 7, and 8.

5.1 Huntington Disease

5.1.1 Clinical description and incidence

Huntington Disease (HD) is an autosomal dominant inherited neurodegenerative disorder that causes progressive motor, cognitive and psychiatric symptoms, and eventually leads to death [11]. Symptoms typically only become clinically evident in middle age, which makes HD particularly devastating and contributes to its transmission to successive generations. The prevalence of HD in Western populations may be as high as 1 in 7,300, making it the most common monogenic neurological disease [11].

5.1.2 Genetics of Huntington's Disease and Huntingtin

The genetic defect responsible for HD was mapped in 1993 [130], and found to be an expansion in a CAG trinucleotide repeat in exon 1 of a gene, *HTT*, located at 4p16.3. *HTT* encodes a large, 348kDA protein, Huntingtin, with pleiotropic and still poorly understood function.

The CAG repeat in HTT codes for a polyglutamine tract near the N-terminus of the protein. The length of this repeat varies, and only causes disease when longer than 35

CAG repeats. The disease is incompletely penetrant for alleles with 36-39 repeats, but is fully penetrant when repeat length exceeds 40.

Two genetic characteristics of Huntington disease are notable and should be accounted for by potential explanations of the disease mechanisms [44]. First, the Huntington disease allele is dominant, its effects are independent of gene dose, and beyond the critical CAG threshold of 40 repeats, it is fully penetrant. This implies that the pathophysiological mechanism is a toxic gain of function. Second, the timing and severity of disease is correlated with the number of CAG repeats. Soon after the discovery of the HTT gene, it was recognized that the number of repeats is both highly variable and unstable through parental transmission [33]. Alleles with more than 45 repeat units often cause a juvenile form of HD, and age of onset is inversely correlated with repeat lengths extending to ~80 repeats, while the more common disease alleles with ~40 repeats lead to the classic presention of HD symptoms in middle age.

5.1.3 Normal functions of Huntingtin

The Huntingtin gene is highly conserved over evolution. Loss of Huntingtin is lethal in mice, and the human *HTT* gene can compensate for loss of the mouse gene to rescue the embryonic lethal phenotype [54]. In addition to the polyglutamine tract, Huntingtin contains several HEAT-repeat domains, which are thought to be involved in protein-protein interactions. Huntingtin is found in many cellular compartments, and translocates between the cytoplasm and nucleus.

Huntingtin protein interactions and complexes have been studied systematically by yeast two-hybrid screens and by immunoprecipitation and mass spectrometry [63, 111, 132]. Huntington seems to interact with hundreds of proteins having diverse functions, including 14-3-3 signaling, presynaptic and post-synaptic organization, calcium signaling, cytoskeletal organization, and mitochondrial function. Huntingtin also interacts with many transcriptional regulators, including CBP [121], TBP[57], SP1[31], REST[145], MeCP2 [87], and the PRC2 complex subunits Ezh2 and Suz12 [109].

5.1.4 Anatomical specificity and selective vulnerability in HD

HD patients exhibit atrophy of brain tissue by MRI and pathological examination, particularly of the caudate and putamen, years before the manifestation of motor symptoms. Later in the course of disease, degeneration of the cortex and other brain regions is apparent, but the striatum is most severely affected. Given the ubuiquitous expression of the huntingtin protein, both in neurons and other cell types, the selective vulnerability of striatal medium-spiny neurons (MSNs) to the effects of mutant huntingtin remains mysterious.

5.1.5 Other polyglutamine repeat diseases

In addition to HD, eight other dominant gain-of-function neurological disesases are caused by expansions of polyglutamine repeats (listed in Table 5.1). Each tends to selectively affect specific brain structures and cell types [93]. The existence of these diseases strongly suggests some common pathophysiological mechanism involving the polyglutamine domain, and the diverse structures affected offer examples to motivate hypotheses about mechanisms of cell-type specificity.

5.2 Huntington Pathophysiology and Polyglutamine Toxicity

Multiple molecular mechanisms contribute to Huntingtin pathogenicity. Labbadia and Morimoto provide a recent review [70], in which they propose that Huntingtin contributes to five major pathogenic processes: impaired protein degradation; altered protein folding; disrupted neuronal circuitry; mitochondrial dysfunction; and transcriptional dysregulation.

Mutant Huntingtin protein tends to form insoluble aggregates and inclusions, which can be found in both the cytoplasm and nucleus [101]. The presence of these aggregates might overwhelm normal cellular capacities for facilitating protein folding and degradation. Other proteins, such as transcription factors that interact with Huntingtin physiologically, may also be sequestered within aggregates and thus inhibited from performing their normal functions.

Disease	Protein	Symbol	Selectively Vulnerable Regions	Normal Repeat Length	Expanded Repeat Len
HD	Huntingtin	HTT	Caudate and Putamen (GABAergic MSNs); Cerebral cortex	6-34	36-121
SCA1	Ataxin1	ATXN1	Cerebellum; brain stem (inf. olive)	6-44	39-82
SCA1	Ataxin2	ATXN2	Cerebellum; brain stem (inf. olive)	15-24	32-200
SCA3	Ataxin3	ATXN3	Cerebellum; basal ganglia; brain stem and SC (but not inf. olive)	13-36	61-84
SCA6	CACNA1A	CACNA1A	Cerebellum; brain stem (inf. olive)	4-19	10-33
SCA7	Ataxin7	ATXN7	Cerebellum; brain stem (inf. olive); photoreceptors	4-35	37-306
SCA17	TATA-binding protein	TBP	Cerebellum; brain stem (inf. olive)	25-42	47-63
SBMA	Androgen Receptor	AR	Anterior horn cell, bulbar neuron, dorsal root gangliion	9-36	38-62
DRPLA	Atrophin-1	ATN1	Cerebellum; cortex; globus pallidus; striatum	7-34	49-88

Table 5.1: Human diseases caused by polyglutamine expansions

mHTT may interfere with mitochondrial function, both through direct interactions with mitochondrial membrane proteins and effects on mitochondrial trafficking, leading to impaired energy metabolism and increases in oxidative stress [70].

mHTT can potentially affect synaptic signaling, which is highly dependent on protein trafficking and turnover, through indirect effects on energy metabolism, proteostatis, and transcription. Large aggregates could also physically interfere with microtubule-mediated transport of materials to the synapse.

5.3 Mouse Models of Huntington Disease

Since human brain tissue cannot be studied directly as Huntington disease progresses, several transgenic mouse models have been developed [89]. Expression of HTT proteins with expanded polyglutamine alleles generate behavioural and motor phenotypes in mice which model key aspects of human HD.

5.3.1 R6/2 N-terminal mHTT model

The R6/2 model was the first transgenic mouse model of HD to be developed. Mangiarini et al. found that expression of only Exon 1 of the human HD gene, containing the expanded CAG repeat, was sufficient to produce a neurological phenotype [83] in C57BL/6 mice. The model was generated by integration of a very small 1.9kb fragment consisting of the endogenous human promoter and first exon of HTT. The number of CAG repeats in these mice is much larger (>150) than that needed to cause human disease (40); also, it was found that the number of repeats is unstable and increases over time as the gene is propagated through the germline. The progression of disease is rapid, with onset of phenotype in the original R6-2 line at 9-11 weeks, and death at 10-13 weeks. The brain (and striatum) in R6-2 mice is smaller than in controls, although the progressive atrophy and neurodegeneration characteristic of human HD is not observed, perhaps because the disease develops so quickly in the model.

5.3.2 YAC128 full-length mHTT model

Slow and colleagues generated transgenic mice which integrated YAC constructs containing a full-length version of the human Htt gene with 128 CAG repeats [115] into the FVB/N background strain. Earlier YAC-based full-length models constructed by the same group with 46 or 72 repeats had a more variable and mild motor phenotype that developed too slowly to be studied efficiently. Slow's YAC128 model develops a phenotype slower than R6-2 but faster than YAC46 and YAC72, and is more consistent and easily measurable. YAC128 mice, as originally described, have a hyperkinetic phenotype that emerges at 3 months, followed by a motor deficit measurable by rotarod assay by 12 months. Striatal and cortial atrophy and quantitative loss in the number of striatal neurons can also also be detected by 12 months. This combination of phenotypes recapitulates many of the key features, including the progressive neurodegeneration, of human HD. Slow suggested that the background strain used for the YAC128 model, FVB/N, may be more vulnerable to excitotoxicity which may contribute to the progressive neuron loss seen in YAC128 that was not detected in the R6/2 model.

5.3.3 Knock-in models of HD

In contrast to the transgenic models, in which an additional copy of the Human *mHTT* gene (or exon 1) is inserted randomly into the mouse genome, knock-in models have an expanded CAG tract inserted into one or both of the endogenous mouse huntingtin genes [110, 138, 88]. Since the number of copies of the Htt gene remains normal, and the gene is expressed under the control of its endogenous promoter, knock-ins should in principle provide a more physiologically realistic model of Huntington's. The first knock-in models, however, with 50-80 CAG repeats, did not exhibit the severe motor phenotypes seen in HD and some of the transgenic mouse models. Knock-ins with longer repeats, such as Q90, Q111, and Q150, do develop rather slowly progressing motor deficits, as well as neuropathological phenotypes including nuclear aggregates. Although less convenient to study, the slow progression of disease in these knock-in models may better mimic the late onset of clinical symptoms characteristic of adult-onset Huntington disease in humans.

5.4 Transcriptonal profiling and dysregulation in HD

It has been well established that prior to cell loss and neurodegeneration in HD, there are changes to neuronal function reflected in altered levels of neurotransmitters and neurotranmitter receptors [20]. *In situ* hybridization of human brains from early stages of HD showed changes in levels of D1 and D2 receptor mRNA and several other transcripts involved in neuronal signaling. These initial observations motivated efforts to profile gene expression systematically in cell-based systems, postmortem human brains and in the transgenic mouse models of HD.

5.4.1 Human brains

Hodges and colleagues measured mRNA expression in the caudate nucleus, cerebellum, prefrontal association area (BA9), and motor cortex (BA4), comparing 44 human brains at early stages of disease (grades 0-2) to 36 unaffected controls [53]. As was expected based on the selective regional vulnerability, a majority of expression changes (nearly 10,000 nominally significant genes) were seen in the caudate, and there were many fewer changes in the cerebellum, which is less vulnerable to neurodegeneration. In the cortex, BA4 had more changes than BA9. The very large numer of genes with changes in affected parts of the HD brain makes functional interpretation difficult, but suggests that transcriptional dysregulation is a central feature of the HD disease process.

5.4.2 Mouse Models

Using microarrays, Kuhn et. al [69] profiled striatal mRNA expression in seven different transgenic mouse models of HD, including the R6-2, YAC128, and knock-in models described above. As in the human studies, thousands of genes have differential expression when compared to wild-type controls. The various models generally exhibit similar changes, and there was not a clear distinction that could be made between the transcriptional impacts of full-length vs. exon-1-only models. Kuhn also noted the statistically significant concordance between these mouse models and expression changes in human HD as reported by Hodges, especially, though not exclusively, among downregulated

genes.

5.4.3 Cell-based Models

Several studies have looked at transcriptional changes caused by *mHTT* in cell-based models. While cells are potentially less faithful to actual disease processes, they can be much easier to manipulate and study. For example, Sipone et al. used an inducible system to profile the effects of expressing mutant Huntingtin fragments of various CAG repeat lengths in the ST14A rat striatal-derived cell line [114]. Expression changes to genes involved in signaling, vesicle trafficking, RNA processing, and lipid metabolism were seen within 12 hours of inducing expression of mutant huntingtin, well before the formation of protein aggregates or cytotoxicity.

5.4.4 Challenges in interpreting HD transcriptional dysregulation

There has been meager progress in interpreting the many transcriptional changes seen in the human HD brains or the mouse models in terms of mechanisms that explain either the development or the consequences of Huntington disease. With relatively few time points (especially from human data), it is difficult to distinguish changes directly caused by mHTT and contributing to the disease process from homeostatic or adaptive changes generated in response to the disease. There are also likely both HD-specific processes and more general responses to cellular stress (eg. from aggregation of toxic proteins and excitotoxicity) represented in any snapshot of transcription. Stresses on neurons may be due either to intrinsic processes or to abnormal signaling from other affected neurons, and this is also difficult to untangle. Finally, studies of transcription from whole tissue measure mixtures of many different neuronal and non-neuronal cell types. Although the studies described above have demonstrated that the number and magnitude of expression changes observed does tend to correlate with the selective regional vulnerability of the striatum, it is likely that many relevant cell-type specific changes within the striatum and other structures cannot be detected using measurements from homogenized tissue.

5.5 Potential mechanisms of transcriptional dysregulation in HD

Nearly every known transcriptional regulatory mechanism has been proposed to be involved in how mHTT leads to transcriptional changes. Moumne et al. [91] provide a recent review of transcriptional regulatory processes that have been associated with Huntington's.

5.5.1 Interactions with transcription factors, co-activators, and repressors

Huntingtin interacts with several transcription factors essential to neuron development and survival. Zuccato et al. found that Huntingtin interacts with REST [145], a repressive factor critical to survival of neurons in old age [79]. Soluble Htt also interacts with and inhibits the function of Sp1 [31], and this interaction is polyglutamine-length dependent. Sp1 is a widely expressed transcriptional activator involved in assembly of TFIID, one of the general transcription factor complexes involved in initiation of RNA polymerase II transcription, with particular importance in neurons and a role in regulation of D2 dopamine receptor and NGFR [76] transcription.

5.5.2 Effects on miRNA

After years of neglect, the functional and regulatory importance of non-coding RNA has recently been recognized. One important class of such RNAs with important functions in neural development are the miRNAs [125]. Hoss et al. [56] quantified both miRNA and mRNA expression using sequencing, in 12 human brains with HD and 9 control brains, and found HD-specific changes to several miRNAs. Several of the most differentially expressed miRNAs were encoded within Hox gene clusters, which are canonical targets of H3K27me3 deposition and regulation by the PRC2 complex [85]. Moreover, several adjacent *Hox* genes were also upregulated.

5.5.3 Epigenetic mechanisms

With the emergence of ChIP-seq and related technologies, there has been increasing recognition of the epigenetic regulation of transcription, through DNA methylation and

the many histone modifications that modulate chromatin structure and accessibility [128]. A wide variety of epignetic changes have been observed in HD models.

DNA Methylation Using bisulfite sequencing, Ng and colleagues found that mHTT caused significant changes to DNA methylation in a strial-derived cell line [92]. These changes were locus specific, and associated with the presence of sequence motifs recognized by the CREB, AP-1, SOX, and ETS families of TFs.

Histone acetylation Valor et al. observed significant hypoacetylation of H3K9,14, and H4K12 in HD82Q knock-in mice, though only a small subset of these acetylation changes were correlated with changes in gene expression [135].

Histone methylation The methylation states of lysine 4 and lysine 27 of Histone 3 are thought to have a particularly important role in defining active vs. repressed loci over development and differentiation [90].

Vashistha et al. observed changes in H3K4me3 marks in both R6/2 mice and human HD brains [136]. H3K4me3 tends to be associated with a transcriptionally active state, and promoters with reduced H3K4me3 were associated with decreased expression. Vashishtha also noted that a specific spatial distribution of the H3K4me3 signal, extending further into the coding region, was particularly prevalent among down-regulated genes. Furthermore, the expression changes due to H3K4me3 dysregulation appeared to be relevant to HD pathology. Knocking down Jarid1, the H3K4 demethylase, restored the expression of genes such as *Bdnf* in cultured neurons from the BACHD mouse, and was protective against degeneration of neurons in a *Drosophila* model.

A potential connection between Huntingtin and H3K27 methylation was first proposed by Seong et al. [109]. Huntingtin is essential for proper embryonic development, and Seong observed that huntingtin-null mouse embryos had phenotypes similar to those lacking Ezh2, Suz12, or Eed, constituents of the Polycomb Repressive Complex (PRC2) responsible for trimethylation of H3K27. PRC2 function was also impaired in the absence of Huntingtin, and the Huntingtin protein formed a complex with PRC2 and increased its methyltransferase activity *in vitro*.

Motivated by these observations, Biagioli et al. [15] studied the effects of Huntingtin knockout and expanded CAG repeats on histone modifications and mRNA expression in a panel of isogenic mouse embryonic stem cell lines (ESC) and ES-derived neural progenitor cells (NPC). They observed that changes in Htt (knockout or expanded CAG repeats) did not generate obvious phenotypes in either the ES or NPC cells.

Htt knockout decreased the number of H3K27me3 enriched transcription start sites, but but did not have a significant effect on the other histone marks assayed, H3K36me3 or H3K4me3. The majority of the H3K27me3 sites that were affected were the so-called bivalent loci, sites marked by both H3K27me3 and H3K4me3 [90]. Such sites in wild-type ES cells lost their H3K27me3 marks and had only H3K4me3 (classified as 'active' loci) in the Htt-null ESCs, while in NPCs, there was an overall increase in bivalent loci. However, some loci switched from being bivalent to active in Htt-null NPC, and other loci that normally become active in WT NPCs retained the H3K27me3 mark and incorrectly remained bivalent in the Htt-null NPCs. This suggested that Htt can affect both H3K27me3 deposition and removal.

Given the importance of repeat-length dependence to HD pathogenesis, Biagioli also investigated the effects of Htt with different numbers of CAG repeats, identifying loci with changes in histone marks correlated with increasing CAG repeat length. Across both ESCs and NPCs, several loci had such repeat-length correlated changes in both H3K27me3 and H3K4me4. The majority of affected loci were different to those changing in the *Htt*-null condition, consistent with the prevailing genetic understanding of HD as a gain of function. The affected loci appeared to be cell-type dependent, and defied any simple interpretation as being a result of either facilitation or inhibition of PRC2 function by mHTT.

5.6 mHTT transcriptional dysregulation, toxicity, and selective vulnerabilty

It is useful to distinguish between several alternative hypotheses about the connection between expression changes and neurodegeneration, as well as related hypotheses about the mechanisms responsible for selective vulnerability in HD.

First, mHTT might cause transcriptional changes (in either direction) that are directly toxic. The development of these changes may be conditional on the transcriptional-

regulatory state, or the ensuing toxicity may depend on the the neurophysiology of vulnerable cell types.

An alternative hypothesis is that mHTT inhibits transcriptional changes normally required for survival over aging in the affected cell types. In Alzheimer's disease, for example, it has been suggested that activity of the REST transcription factor is essential to protect against oxidative stress and amyloid β-protein toxicity [79]. mHTT might have an opposite effect, inhibiting normal transcriptional changes necessary for survival of neurons that are subjected to stress as they age. If this were the case, genes whose expression changes over time in affected cell types and that are also modulated by mHTT might be of particular relevance.

A third hypothesis is that mHTT is toxic mainly through mechanisms other than direct effects on transcription. In this scenario, expression changes are a symptom of cellular dysfunction, but are not necessary for mHTT to be toxic to affected neurons.

Independent of expression, the selective vulnerability of HD could result from cell-autonomous differences directly affecting the toxicity of mHTT, such as a nuclear or cytoplasmic environment in which mHTT forms toxic aggregates faster. Alternatively, specific cell types may depend on some process or function, such as the activity of particular transcription factors, specifically impaired by mutant Huntingtin. Vulnerability might be due to neurophysiological differences, not directly connected to Huntingtin function, such as increased sensitivity to glutamate excitotoxicity or a higher and more sustained level of metabolic activity and oxidative stress. Such neurons could be vulnerable to defects in homeostasis which might otherwise be tolerated by cells operating in a more relaxed neurophysiological and metabolic regime.

5.7 Categorizing and explaining HD-dysregulated genes

Given the extensive gene expression changes observed in HD, it seems reasonable to assume that at least a subset of such changes are relevant to HD pathophysiology. It is therefore important to prioritize the most relevant genes, understand their pathological effects, and explain how their dysregulation is a consequence of mutant Htt.

Several distinct approaches to classifying these genes may be useful. Expression

changes can be organized in terms of biochemical processes that might affect neuronal function and survival. Alternatively, one can classify expression changes by their dependencies on factors such as cell type, tissue type, time, models, and repeat length. Finally, such dependencies, along with prior knowledge about potential regulatory interactions, may be used to generate hypotheses about the mechanisms and molecules causing the observed expression changes, and about how those regulatory processes are related to the proximal effects of mutant Huntingtin protein.

Functional activities and physiological consequences

The most obvious way to categorize gene expression changes is by neurophysiological function or pathological effects. Affected genes can first be organized by the molecular function and biological processes in which they are involved, using the many stanard databases and methods available for assigning function and assessing over-representation of annotations. Such categorizations and functional abstractions can be further refined by their predicted pathological impact. Affected genes might be contributing directly to pathological mechanisms, might be compensatory, or might be completely irrelevant to the course of disease and outcomes of interest.

Dependencies

A second approach to categorize expression changes is by their dependence on other variables relevant to HD, which include:

- Differential expression as a function of mHtt status
- Anatomical and cell-type specificity of WT expression
- Anatomical and cell-type specificity of mHtt dependent differential expression
- Time-course of expression in WT, vs. time-course of (differential) expression in the presence of mHtt
- CAG-repeat length

Identifying such dependencies depends on larger experiments to test the many possible conditions and their interactions, and such experiments have only recently become technically and economically practical.

Inferred regulatory mechanisms

A third way to organize expression changes is by the mechanisms responsible for their regulation. Given a set of coördinately regulated genes, one can predict potential transcriptional regulatory processes and proteins based on the *cis*-regulatory motifs in promoter regions, biochemical interactions between chromatin-binding proteins and the relevant loci, and co-expression observed in other contexts.

5.8 Objectives

Given the enduring state of confusion about the context-dependent connections between *mHtt* and gene expression, we sought to take a systematic, computational approach to interpreting available expression data using the frameworks outlined above.

We hypothesized that given cell-type-specific TRAP data and time-resolved, tissue-specific allelic series data that has been generated recently, combined with the large amounts of prior data about gene regulatory mechanisms now available, we might now be in a better position to consider three questions:

1. What are the most likely regulatory mechanisms contributing to transcriptional changes downstream of mHTT?

2. Which expression changes are tissue- and cell-type specific, what mechanisms contribute to that specificity, and how are specific expression changes related to the selective vulnerability of different cell types?

3. Which expression changes are time-dependent, and how do such dependencies inform hyptheses about possible regulatory and pathological mechanisms?

In chapter 6, we re-analyze published expression data to predict possible regulators of HD-dependent expression changes. In chapter 7, we analyze a cell-type-specific

dataset generated by Fenster and Heiman to also consider cell-type dependecies of HD-associated expression changes, along with their potential regulatory mechanisms. Finally, in chapter 8, we analyze data from a large experiment performed by the CHDI Foundation [22], which used HD knock-in models to assess the dependence of expression changes on time, CAG-repeat length, and additional diverse tissue types.

Chapter 6

Over-representation of PRC2 targets among HD-dysregulated genes

6.1 Introduction

Technologies for expression profiling have been applied to characterize transcriptional dysregulation in Huntington mouse models and human HD brains. Here, we review and re-analyze several of these published HD expression profiling datasets to generate hypotheses about putative regulators of the transcriptional changes in Huntington's.

Hodges et al. [53] conducted one of the largest studies to date using Human HD brain tissue (from caudate, cerebellum, and the BA4 and BA9 regions of the frontal cortex), comparing gene expression in 44 HD brains to 36 controls using Affymetrix HG-U133A and HG-U133B arrays. As expected based on HD pathology, the greatest number of HD-associated expression changes were observed in the caudate (Table 6.1a).

Kuhn et al. [69] (which also incorporated data from Becanovic et al. [12]) profiled expression in seven mouse models (R6/2, R6/1, CHL2, HdhQ92, Hdh4/80Q, HD46, YAC128) using Affymetrix Mouse 430 2.0 microarrays. Transgenic models based on the short, exon-1 only models (R6/2) develop phenotypes most quickly and have the greatest number of expression changes, while knock-in and full-length models develop more slowly (Table 6.1b). Kuhn observed that there is statistically significant concordance among most of the mouse models, and between the mouse and human data, particularly for down-regulated genes.

Since these expression profiling experiments were first published, accumulated knowledge of transcriptional regulators and their potential targets has grown. We hypothesized that by comparing the sets of HD-dysregulated genes to the large number of ChIP-ChIP

Sample	Number of probe sets (at p < 0.001) down	up
Caudate	4432	5331
Cerebellum	382	131
BA4 cortex	958	1482
BA9 cortex	5	6

(a) Numbers of up- and down-regulated probe sets in human HD brain tissues in Grades 0-2 samples, from Table 2 of Hodges et al, 2006. [53]

nominal p-value threshold	0.01				0.05			
log2 fold change threshold	0.00		0.58		0.00		0.58	
direction	down	up	down	up	down	up	down	up
dataset								
Becanovic YAC128 24mo	187	148	7	28	817	644	18	54
Kuhn GSE10202 CHL2	1314	1444	595	277	2527	3124	683	356
Kuhn GSE7958 HdhQ92/Q92 18 mo	172	152	37	23	895	923	117	127
Kuhn GSE7958 HdhQ92/Q92 3 mo	66	45	1	1	491	274	3	8
Kuhn GSE9303 R6/2 group 1	1176	1625	519	261	2690	3577	642	368
Kuhn GSE9304 R6/2 group 2	1244	1516	530	336	2972	3237	656	514

(b) Numbers of up- and down-regulated genes in Kuhn's mouse HD models at some representative fold-change and nominal p-value thresholds, using Welch's t-test.

Table 6.1: Numbers of expression changes in Human HD samples and mouse models

and ChIP-seq datasets now available, as well as with additional sources of regulatory information such as motif occurrences and epigenetic marks, we might implicate potential regulatory mechanisms that had not been previously recognized.

6.2 Experimental Designs and Data

Kuhn Mouse Model Data

Processed, normalized expression data from the Kuhn and Becanovic studies was downloaded as GEO Series Matrix files from the NCBI Gene Expression Omnibus (`http://www.ncbi.nlm.nih.gov/geo`). We focused on data acquired using the Affymetrix 430 2.0 platform to facilitate comparisons between datasets, as well as those from later time-points at which the HD model phenotypes are most evident.

GEO Accession	Model	Age
GSE10202	CHL2(HdhQ15)	22 months
GSE7958	HdhQ92	18 months
GSE9375	Hdh4Q80	12 months
GSE9803	R6/2	12 weeks (set 1)
GSE9804	R6/2	12 weeks (set 2)
GSE19677	YAC128	12 months
GSE19677	YAC128	24 months

Table 6.2: Datasets from Kuhn et al. used in analysis

Hodges Human HD Data

Differentially expressed genes and statistics from the Hodges et al. study were extracted from table S1 of their supplementary material.

6.3 Methods

6.3.1 Differential expression testing

For the mouse model data, differential expression was tested using Welch's t-test, comparing each model to its wild-type control.

For the human data, differential expression was assessed using the p-values from moderated t-tests reported in the supplementary table S1 of Hodges et al., which had been calculated using *limma* [116].

The relatively small number of samples limited power, particularly in the mouse studies, and as in the initial reports we worked with mainly with nominal p-values.

6.3.2 Over-representation analyses

Over-representation was assessed using the hypergeometric test, assuming a null model drawing from the set of all gene symbols. Given the small number of replicates available in most of the experiments analyzed, it was not practical or beneficial to use non-parametric methods that rely on sample permutation.

Regulatory motif occurrences were extracted from the SwissRegulon database [94]. These annotations are based on recognition of motifs compiled from the JASPAR and

TRANSFAC databases. Since short motifs can occur frequently by chance, the observation of a high-scoring match is not sufficient to predict that a motif is functional. SwissRegulon uses a Bayesian model to integrate information about the location of a motif in the promoter relative to the transcription start site with information about conservation in orthologous genes from human, mouse, macaque, dog, cow, horse, and opossum [5], and assigns a probability for each motif that it predicts in promoter regions.

Targets of chromatin-binding regulatory proteins were obtained from the ChEA database [71].

P-values from all tests were adjusted using the Benjamini–Hochberg procedure with "multicomp.multipletests" in python statsmodels [108] to control false-discovery rate over all probe-sets. Bonferroni-adjusted and nominal P values are also reported.

6.4 Results

6.4.1 Transcriptional regulators with altered expression in HD

Changes to the activities of transcriptional regulators offer one potential mechanism by which the extensive transcriptional changes of HD might be produced. Although functional activities can be (and often are) modulated independently from expression, as a first step we decided to investigate the largest expression changes among transcriptional regulators (incuding both transcription factors, their interactors, and epigenetic regulators and modifiers). The set of such genes was defined based on having gene ontology molecular function annotations containing any of the following terms: "DNA binding", "transcriptional", "transcription factor", "regulation of transcription", "sequence-specific DNA binding", "chromatin modification", "enhancer binding", "nucleic acid binding", or "chromatin binding".

We focused initially on the R6-2 model, which has the largest number of transcriptional changes. Transcriptional regulators with the biggest decreases in expresssion are listed in Table 6.3. Among these are Egr2, a zinc-finger transcription factor in the immediate-early genes (IEG) family. Decreased expression of immediate-early genes in the HD striatum is well established [119], and thought to be a possible consequence of decreased dopamine

and adenosine receptor signaling. Expression of the similar Egr3 gene also decreased.

Npas4 is also classified as an IEG, and is thought to be important to plasticity and learning [86]. Claspin is a DNA-binding protein involved in formation of replication forks as well as DNA-damage responses and checkpoint regulation. Lmo2 is an adaptor protein involved in assembly of transcription factor complexes, more famously known for its role as an oncogene in T-cell leukemia. It is expressed throughout the brain, but its relevance to neurons is not well understood. Among the many proteins with which Lmo2 interacts is Kdm5a (Jarid1a) [84], the lysine-specific demethylase responsible for removal of H3K4me3 marks. Tcf7 is a transcription factor activated by Wnt signaling, which is thought to protect against neurodegeneration [7].

Transcriptional regulators whose expression increased included Nfxl1, a zinc-finger transcription factor whose function is not well annotated. The function of Crebzf in the brain is also not completely understood, but it may regulate the unfolded protein response [142], and may also modulate BMP signaling by binding to SMADs [72]. Sox11, which is upregulated in HD in both the R6-2 and YAC128 models, is a transcription factor thought to promote neuronal survival and neurogenesis, possibly through induction of BDNF [105].

	[["dataset", "Becanovic YAC128 24mo"], ["mc", "nominal"], ["st", "pval"], ["tt", "welch ttest"]]	[["dataset", "Kuhn GSE9303 R6/2 group 1"], ["mc", "nominal"], ["st", "pval"], ["tt", "welch ttest"]]	[["dataset", "Kuhn GSE9304 R6/2 group 2"], ["mc", "nominal"], ["st", "pval"], ["tt", "welch ttest"]]	[["dataset", "Becanovic YAC128 24mo"], ["st", "fc_means"]]	[["dataset", "Kuhn GSE9303 R6/2 group 1"], ["st", "fc_means"]]	[["dataset", "Kuhn GSE9304 R6/2 group 2"], ["st", "fc_means"]]	symbol	gene_name
1427683_at	0.49	0.00013	0.0014	-0.44	-2.3	-2.8	Egr2	early growth response 2
1459372_at	0.85	0.00044	0.00011	0.15	-2.1	-2.5	Npas4	neuronal PAS domain protein 4
1427682_a_at	0.58	0.00085	0.017	-0.34	-1.7	-2.3	Egr2	early growth response 2
1456280_at	0.41	0.00053	4.9e-05	-0.4	-1.5	-1.9	Clspn	claspin
1421037_at	0.00088	1.2e-06	2e-05	-0.37	-1.2	-1.3	Npas2	neuronal PAS domain protein 2
1454086_a_at	0.34	4.8e-05	0.00015	-0.22	-1.1	-1.4	Lmo2	LIM domain only 2
1451280_at	0.64	0.00014	0.00037	-0.16	-1.1	-1.6	Arpp21	cyclic AMP-regulated phosphoprotein, 21
1433959_at	0.14	0.0029	0.39	-0.13	-1	-0.14	Zmat4	zinc finger, matrin type 4
1419665_a_at	0.92	0.0012	0.12	0.028	-1	-0.74	Nupr1	nuclear protein transcription regulator 1
1429779_at	0.18	0.0026	0.001	-0.24	-0.98	-0.91	Ago4	argonaute RISC catalytic subunit 4
1433471_at	0.89	0.0036	7.9e-05	-0.059	-0.89	-1.3	Tcf7	transcription factor 7, T cell specific
1444152_at	0.83	0.0033	0.85	-0.072	-0.89	0.03	Celf2	CUGBP, Elav-like family member 2
1424248_at	0.26	7.1e-06	8.6e-06	-0.13	-0.88	-1.3	Arpp21	cyclic AMP-regulated phosphoprotein, 21
1453289_at	0.19	0.003	0.00012	-0.25	-0.85	-1	Ago4	argonaute RISC catalytic subunit 4
1451046_at	0.098	0.00054	0.0085	-0.34	-0.85	-0.66	Zfpm1	zinc finger protein, multitype 1
1421175_at	0.92	1e-05	0.021	-0.018	-0.83	-0.44	Myt1l	myelin transcription factor 1-like
1436329_at	0.064	0.0037	0.00095	-0.2	-0.82	-1.1	Egr3	early growth response 3
1418317_at	0.34	0.0025	0.034	-0.21	-0.8	-0.4	Lhx2	LIM homeobox protein 2
1423478_at	0.099	1e-05	0.00052	-0.15	-0.79	-0.78	Prkcb	protein kinase C, beta
1453287_at	0.6	0.0025	0.74	-0.19	-0.78	-0.098	Ankrd33b	ankyrin repeat domain 33B

Table 6.3: Transcriptional regulatory genes down-regulated in R6/2

	[["dataset", "Becanovic YAC128 24mo"], ["mc", "nominal"], ["st", "pval"], ["tt", "welch ttest"]]	[["dataset", "Kuhn GSE9303 R6/2 group 1"], ["mc", "nominal"], ["st", "pval"], ["tt", "welch ttest"]]	[["dataset", "Kuhn GSE9304 R6/2 group 2"], ["mc", "nominal"], ["st", "pval"], ["tt", "welch ttest"]]	[["dataset", "Becanovic YAC128 24mo"], ["st", "fc_means"]]	[["dataset", "Kuhn GSE9303 R6/2 group 1"], ["st", "fc_means"]]	[["dataset", "Kuhn GSE9304 R6/2 group 2"], ["st", "fc_means"]]	symbol	gene_name
1418152_at	0.8	0.0055	1.8e-05	-0.045	0.58	0.54	Hmgn5	high-mobility group nucleosome binding domain 5
1434618_at	0.43	0.00061	0.0054	-0.21	0.59	0.75	Crebzf	CREB/ATF bZIP transcription factor
1455634_at	0.41	0.0019	0.021	0.11	0.59	0.33	Son	Son DNA binding protein
1460725_at	0.21	0.00045	0.027	0.24	0.6	0.21	Xpa	xeroderma pigmentosum, complementation group A
1429170_a_at	0.83	0.0043	0.0085	0.06	0.6	0.28	Mtf1	metal response element binding transcription f...
1418046_at	0.7	0.00016	0.011	0.077	0.6	0.44	Nap1l2	nucleosome assembly protein 1-like 2
1454976_at	0.16	0.002	0.093	0.2	0.61	0.22	Sod2	superoxide dismutase 2, mitochondrial
1418640_at	0.72	0.0015	0.016	0.072	0.61	0.33	Sirt1	sirtuin 1 (silent mating type information regu...
1448733_at	0.31	0.0022	0.012	0.21	0.61	0.25	Bmi1	Bmi1 polycomb ring finger oncogene
1448454_at	0.59	1.4e-05	0.00066	0.15	0.61	0.45	Srsf6	serine/arginine-rich splicing factor 6
1436191_at	0.13	0.0085	0.053	0.24	0.61	0.23	Arid4a	AT rich interactive domain 4A (RBP1-like)
1448497_at	0.15	0.00074	6e-05	0.15	0.62	0.61	Ercc3	excision repair cross-complementing rodent rep...
1436241_s_at	0.12	6.2e-06	0.015	-0.14	0.62	0.49	Hira	histone cell cycle regulation defective homolo...
1449121_at	0.46	2.8e-05	0.0037	0.16	0.63	0.43	Srsf10	serine/arginine-rich splicing factor 10
1416433_at	0.076	0.0067	0.14	0.26	0.63	0.24	Rpa2	replication protein A2
1429051_s_at	0.19	0.00062	0.22	0.49	0.63	0.28	Sox11	SRY-box containing gene 11
1435302_at	0.64	0.0064	0.067	-0.064	0.63	0.29	Taf4b	TAF4B RNA polymerase II, TATA box binding prot...
1456651_a_at	0.58	0.0049	0.21	0.068	0.64	0.32	Tpr	translocated promoter region
1422741_a_at	0.77	0.0012	0.085	-0.063	0.64	0.27	Bbx	bobby sox homolog (Drosophila)
1417145_at	0.42	0.0049	0.00014	0.09	0.64	0.72	Nfxl1	nuclear transcription factor, X-box binding-li...

Table 6.4: Transcriptional regulatory genes up-regulated in R6/2

6.4.2 Over-representation analysis of targets of chromatin-binding factors

Examination of expression changes for individual transcription factors and related genes hints at the diverse regulatory processes affected, but there are many more genes changing in HD models. These cannot be easily explained from changes in transcription factor expression, especially since many regulatory targets remain unknown and expression regulation occurs by many mechanisms other than by changes in transcription factor expression levels. To more systematically search for potential common regulators of the genes dysregulated in HD, we computed overlaps between the sets of genes changing in each model (using a nominal p-value cut-off of 0.01, and a fold-change cut-off of 1.5-fold) and sets of genes associated with chromatin binding factors from the ChEA database.

The top 10 over-represented sets for each HD model are shown in Tables 6.5, 6.6, 6.7, 6.8, and 6.10, and summarized in Figure 6.4.2.

Among the regulators whose targets were most significantly over-represented in the sets of HD-dysregulated genes are members of the PRC2 complex. The core proteins of the PRC2 complex are Suz12, EED, and Ezh1/2 [61]. Targets of both Suz12 and Ezh2 were significantly over-represented among down-regulated genes in the R6/2, CHL2, and HqhQ92 models. Also over-represented were targets of Jarid2, a regulator of the PRC2 complex; Mtf2, a transcription factor involved in recruiting the PRC2 complex to sites marked with H3K36me3; and Rnf2, another Polycomb group protein that interacts with PRC2 and which has also been reported to interact with Hip2, a Huntingtin-interacting protein [73].

Other highly-ranked potential regulators suggested by this analysis include Rcor3, a transcriptional co-repressor thought to operate in a complex with Lsd1 and Kdm1a. Also notable is Wt1, a zinc-finger transcription factor that is itself differentially expressed in many HD models.

There are fewer genes up-regulated in the HD models, which limits power to detect over-representation. However, among the top-ranked regulators is Kdm5b, the lysine-specific demethylase responsible for demethylation of H3K4 sites. This observation is interesting given the previously reported connection between H3K4me3 patterns and transcriptional dysregulation in HD [136].

The YAC128 model, as previously discussed, develops its phenotype much more slowly, and there are comparatively fewer expression changes observed at the same statistical thresholds. Curiously, several PRC2 targets are over-represented among the *up-regulated* genes at the late stages of disease in the YAC128 model, in contrast to the other models.

Finally, we performed the analagous analysis of genes differentially expressed in Huntington disease in the Hodges human caudate samples (Figure 6.4.2 and Tables 6.11 and 6.12). Components of the PRC2 complex and related proteins - Suz12, Mtf2, Ezh2, Rnf2, Eed, and Jarid2 - were again observed to have targets over-represented among genes significantly down-regulated in the HD caudate.

	Overlap	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
set_name						
SUZ12	277	5042	1.4e-46	1.8e-44	1.8e-44	0.42
MTF2	201	2981	2.1e-43	2.8e-41	1.4e-41	0.31
EZH2	105	1300	3.6e-27	4.8e-25	1.6e-25	0.16
JARID2	118	1639	3.3e-26	4.3e-24	1.1e-24	0.18
RNF2	121	1975	7.6e-21	1e-18	2e-19	0.19
RCOR3	121	2851	2.5e-09	3.4e-07	5.6e-08	0.19
WT1	80	1663	1.3e-08	1.7e-06	2.5e-07	0.12
EED	46	830	5e-07	6.6e-05	8.3e-06	0.07
PHC1	49	922	7.1e-07	9.4e-05	1e-05	0.075
TET1	133	3596	1.3e-06	0.00018	1.8e-05	0.2

(a) R6/2 downregulated

	Overlap	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
set_name						
CREM	106	5776	6.9e-06	0.00091	0.00091	0.36
KDM5B	75	3724	1.4e-05	0.0019	0.00093	0.26
GATA4	41	2039	0.002	0.26	0.086	0.14
ZFP42	31	1480	0.004	0.53	0.11	0.11
SIN3A	26	1186	0.0046	0.61	0.11	0.088
STAT5	26	1197	0.0052	0.69	0.11	0.088
ERG	38	1969	0.0057	0.76	0.11	0.13
PDX1	16	669	0.011	1	0.19	0.054
TCFCP2L1	36	1987	0.018	1	0.26	0.12
NKX2-5	28	1507	0.026	1	0.35	0.095

(b) R6/2 up regulated

Table 6.5: Over-representation of targets of chromatin-binding factors from ChEA among genes dysregulated in R/2 model of HD (Group 1)

	Overlap	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
set_name						
SUZ12	270	5042	1.5e-40	2.1e-38	2.1e-38	0.37
MTF2	184	2981	2e-32	2.6e-30	1.3e-30	0.25
EZH2	91	1300	1.3e-18	1.7e-16	5.7e-17	0.13
JARID2	102	1639	2.8e-17	3.8e-15	9.4e-16	0.14
RNF2	109	1975	8.5e-15	1.1e-12	2.3e-13	0.15
WT1	77	1663	3e-07	4e-05	6.6e-06	0.11
NRF2	52	1055	5.4e-06	0.00072	9e-05	0.072
NFE2L2	52	1055	5.4e-06	0.00072	9e-05	0.072
BMI1	73	1682	6.6e-06	0.00087	9.7e-05	0.1
OLIG2	84	2040	9.2e-06	0.0012	0.00012	0.12

(a) R6/2 downregulated group 2

	Overlap	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
set_name						
KDM5B	80	3724	0.0045	0.59	0.18	0.21
SUZ12	103	5042	0.0053	0.71	0.18	0.27
OLIG2	48	2040	0.006	0.8	0.18	0.13
MECOM	46	1951	0.0069	0.91	0.18	0.12
CREM	115	5776	0.0069	0.92	0.18	0.3
CEBPD	16	504	0.0089	1	0.2	0.042
YAP1	52	2329	0.011	1	0.22	0.14
TEAD4	50	2293	0.02	1	0.3	0.13
NR4A2	8	207	0.021	1	0.3	0.021
CRX	18	668	0.026	1	0.3	0.047

(b) R6/2 up regulated group 2

Table 6.6: Over-representation of targets of chromatin-binding factors from ChEA among genes dysregulated in R/2 model of HD (Group 2

	Overlap	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
set_name						
GFI1B	11	1871	0.0027	0.36	0.36	0.23
SALL4	9	1825	0.02	1	0.9	0.19
TEAD4	10	2293	0.031	1	0.9	0.21
YAP1	10	2329	0.034	1	0.9	0.21
TRIM28	12	3072	0.039	1	0.9	0.26
DMRT1	9	2144	0.05	1	0.9	0.19
NR3C1	5	918	0.057	1	0.9	0.11
CNOT3	7	1547	0.059	1	0.9	0.15
TCF3	13	3743	0.071	1	0.9	0.28
PAX6	5	1001	0.076	1	0.9	0.11

(a) YAC128 12mo downregulated

	Overlap	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
set_name						
DMRT1	5	2144	0.011	1	1	0.36
CHD1	3	843	0.018	1	1	0.21
TRIM28	5	3072	0.045	1	1	0.36
SOX2	5	4207	0.14	1	1	0.36
FOXP3	1	257	0.16	1	1	0.071
KLF1	2	1144	0.18	1	1	0.14
HCFC1	1	306	0.19	1	1	0.071
TEAD4	3	2293	0.2	1	1	0.21
RCOR1	3	2378	0.21	1	1	0.21
RARG	1	390	0.23	1	1	0.071

(b) YAC128 12mo up-regulated

Table 6.7: Over-representation of targets of chromatin-binding factors from ChEA among genes dysregulated in YAC128 at 12 mo

	Overlap	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
set_name						
TCFAP2C	4	2667	0.007	0.94	0.94	0.57
MEF2A	2	1048	0.046	1	1	0.29
PPARD	1	194	0.064	1	1	0.14
SUZ12	4	5042	0.065	1	1	0.57
ETS2	1	215	0.071	1	1	0.14
TCF7	2	1529	0.09	1	1	0.29
HCFC1	1	306	0.099	1	1	0.14
NR0B1	2	1691	0.11	1	1	0.29
SALL4	2	1825	0.12	1	1	0.29
TP53	3	3937	0.13	1	1	0.43

(a) YAC128 24mo downregulated

	Overlap	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
set_name						
MTF2	16	2981	1.9e-07	2.5e-05	2.5e-05	0.57
SUZ12	18	5042	9.1e-06	0.0012	0.0006	0.64
JARID2	10	1639	3.4e-05	0.0045	0.0015	0.36
EZH2	8	1300	0.00024	0.032	0.0081	0.29
EED	6	830	0.00073	0.098	0.019	0.21
RNF2	9	1975	0.00085	0.11	0.019	0.32
WT1	6	1663	0.022	1	0.38	0.21
BMI1	6	1682	0.023	1	0.38	0.21
EOMES	6	1744	0.027	1	0.4	0.21
YAP1	7	2329	0.032	1	0.42	0.25

(b) YAC128 24mo up-regulated

Table 6.8: Over-representation of targets of chromatin-binding factors from ChEA among genes dysregulated in YAC128 at 24 mo

	Overlap	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
set_name						
MTF2	202	2981	5.4e-34	7.2e-32	7.2e-32	0.27
SUZ12	278	5042	3.5e-33	4.7e-31	2.3e-31	0.37
JARID2	121	1639	2.2e-22	2.9e-20	9.8e-21	0.16
EZH2	98	1300	1.1e-18	1.5e-16	3.7e-17	0.13
RNF2	122	1975	2.6e-16	3.4e-14	6.8e-15	0.16
WT1	86	1663	7.6e-08	1e-05	1.7e-06	0.11
TP53	163	3937	3.5e-07	4.7e-05	6.7e-06	0.21
BMI1	78	1682	2e-05	0.0026	0.00033	0.1
YAP1	99	2329	4.7e-05	0.0062	0.00069	0.13
MYB	48	923	6.2e-05	0.0082	0.00076	0.063

(a) CHL2 downregulated

	Overlap	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
set_name						
KDM5B	70	3724	0.0016	0.21	0.14	0.22
SUZ12	89	5042	0.0022	0.29	0.14	0.28
MTF2	57	2981	0.0033	0.44	0.15	0.18
CEBPD	13	504	0.019	1	0.57	0.041
ZFP42	29	1480	0.026	1	0.57	0.092
BMI1	32	1682	0.029	1	0.57	0.1
IRF8	21	1004	0.03	1	0.57	0.067
YAP1	40	2329	0.062	1	0.88	0.13
CHD1	17	843	0.063	1	0.88	0.054
PHC1	18	922	0.073	1	0.88	0.057

(b) CHL2 up-regulated

Table 6.9: Over-representation of targets of chromatin-binding factors from ChEA among genes dysregulated in CHL2 model

	Overlap	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
set_name						
MTF2	16	2981	2.2e-05	0.0029	0.0029	0.38
SUZ12	19	5042	0.00037	0.049	0.025	0.45
JARID2	9	1639	0.002	0.27	0.088	0.21
RNF2	9	1975	0.007	0.92	0.23	0.21
PPARD	2	194	0.047	1	0.99	0.048
TCF3	11	3743	0.059	1	0.99	0.26
THAP11	4	864	0.068	1	0.99	0.095
BMI1	6	1682	0.077	1	0.99	0.14
EZH2	5	1300	0.081	1	0.99	0.12
DMRT1	7	2144	0.084	1	0.99	0.17

(a) Hdh Q92 model at 18 months downregulated

	Overlap	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
set_name						
OLIG2	6	2040	0.021	1	1	0.25
TBP	4	1057	0.028	1	1	0.17
SFPI1	6	2369	0.041	1	1	0.25
KDM5B	8	3724	0.042	1	1	0.33
ERG	5	1969	0.062	1	1	0.21
CHD1	3	843	0.066	1	1	0.12
SMARCA4	7	3481	0.079	1	1	0.29
YY1	2	464	0.094	1	1	0.083
MYCN	5	2261	0.1	1	1	0.21
MEF2A	3	1048	0.11	1	1	0.12

(b) Hdh Q92 model at 18 months up-regulated

Table 6.10: Over-representation of targets of chromatin-binding factors from ChEA among genes dysregulated in Hdh Q92 model at 18 months

	Overlap	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
set_name						
SMAD2	58	1936	3.6e-07	7.3e-05	7.3e-05	0.19
BACH1	38	1352	0.00018	0.037	0.014	0.12
SMAD3	73	3233	0.00024	0.048	0.014	0.23
FOXA2	68	2968	0.00027	0.055	0.014	0.22
EP300	64	2902	0.0012	0.24	0.047	0.2
CIITA	5	59	0.002	0.4	0.049	0.016
ATF3	50	2189	0.0021	0.43	0.049	0.16
WT1	43	1817	0.0023	0.47	0.049	0.14
RARG	14	390	0.0027	0.54	0.049	0.045
PPARD	79	3876	0.0027	0.55	0.049	0.25
PRDM14	45	1944	0.0028	0.57	0.049	0.14
NRF2	28	1055	0.003	0.61	0.049	0.089
PAX3-FKHR	28	1063	0.0033	0.67	0.049	0.089
CDX2	23	821	0.0037	0.74	0.049	0.073
HIF1A	12	321	0.0038	0.77	0.049	0.038

Table 6.11: Hodges Caudate Up Chea Overrepresentation

	Overlap	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
set_name						
SUZ12	224	5042	4.7e-17	9.6e-15	9.6e-15	0.4
MTF2	153	2981	3.8e-16	7.6e-14	3.8e-14	0.28
EZH2	79	1328	2.4e-11	4.8e-09	1.6e-09	0.14
RNF2	97	1975	3.8e-09	7.6e-07	1.9e-07	0.17
EED	49	830	2.5e-07	5e-05	1e-05	0.088
JARID2	76	1639	2.2e-06	0.00044	7.4e-05	0.14
BACH1	56	1352	0.00091	0.18	0.026	0.1
PHC1	41	922	0.0012	0.25	0.031	0.074
GBX2	17	286	0.0021	0.41	0.046	0.031
IKZF1	11	155	0.0033	0.66	0.065	0.02
WT1	68	1817	0.0035	0.71	0.065	0.12
ZFP281	80	2252	0.0061	1	0.096	0.14
RCOR3	98	2851	0.0062	1	0.096	0.18
YAP1	79	2329	0.019	1	0.27	0.14
PAX3-FKHR	40	1063	0.021	1	0.28	0.072

Table 6.12: Hodges Caudate Down Chea Overrepresentation

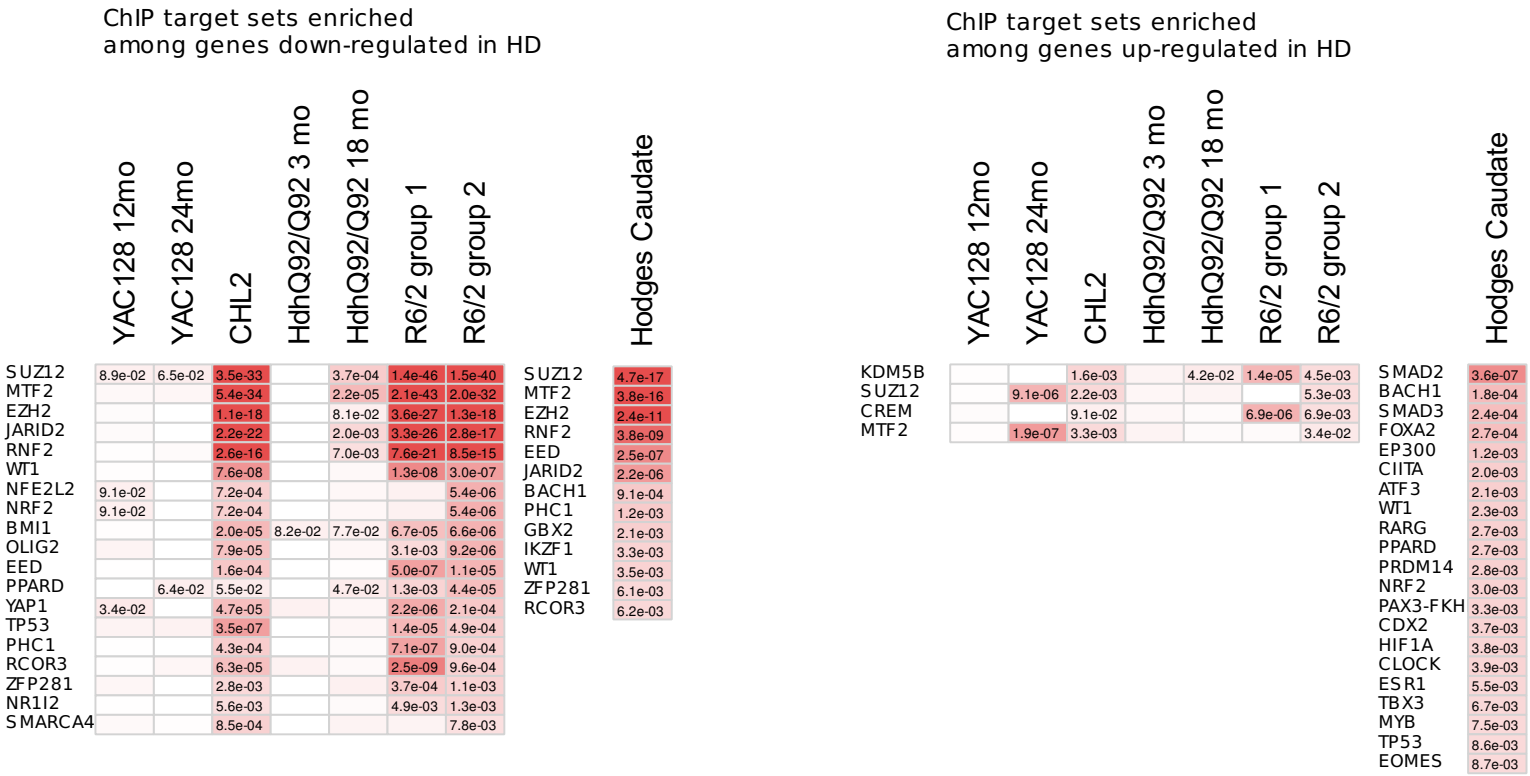


Figure 6.1: Summary of ChEA ChIP groups with targets overrepresented among differentiallly expressed genes in at least one of the models profiled in Kuhn et al.. [69], and in human Caudate profiled in Hodges et al. [53]

6.4.3 Over-representation analysis of regulatory motifs

We next considered whether any regulatory motifs were over-represented in the promoters of HD-dysregulated genes. We used the pre-computed Swissregulon database of motif occurrences [94], and for each motif extracted the set of genes for which that motif occurred within 1kb of the transcription start with a score of 0.7 or greater. Only the the R6/2 and CHL2 models had sufficient numbers of differentially expressed genes to yield significantly over-represented motifs in this analysis (Tables 6.13, 6.14, 6.15).

Among up-regulated genes, several AP-2 family motifs were over-represented. Among down-regulated genes, over-represented motifs included one recognized by Znf143, a factor thought to be involved in recruitment of distal regulatory elements to promoters [9]. A recent meta-analysis of expression data in neurodegeneration also identified Znf143 as a possible regulator of genes associated with neuropathology but not with normal aging [75].

6.4.4 A majority of genes changing in HD have striatal-specific expression

Our initial over-representation analysis of transcriptional regulators and potential regulatory motifs described above rely on the admittedly very naïve null model in which differentially expressed genes are drawn randomly. However, the genes expressed within neurons, and in specific neural cell types such as medium-spiny neurons are obviously not selected randomly. Some genes are *never* expressed in neurons, and will never exhibit differential expression in either direction. Genes with high levels of expression are more likely to be detected, regardless of their direction of change. Genes already expressed at maximal levels might not be able to show increased expression, and therefore more of such selectively highly expressed genes may be observed to decline. Similarly, down-regulation might not be detectable in genes with already very low levels of expression, while increases would be more apparent. Any of these situations could lead to confounding the properties of selectively expressed genes with characteristics of genes dysregulated by Huntington's.

It is already known that many genes with striatal-selective expression (including classic markers such as the dopamine receptors) are among the most severely dysregulated

	Overlap	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
set_name						
TFAP2{A,C}.p2	304	7721	5.8e-18	1.1e-15	1.1e-15	0.45
TFAP2B.p2	241	5916	1.3e-14	2.5e-12	1.2e-12	0.36
TFDP1.p2	297	7949	7e-14	1.3e-11	4.3e-12	0.44
MTF1.p2	116	2283	4e-12	7.4e-10	1.9e-10	0.17
PATZ1.p2	343	9909	5.4e-12	1e-09	2e-10	0.51
MAZ.p2	353	10505	1.4e-10	2.6e-08	4.4e-09	0.52
EGR1..3.p2	147	3378	3.2e-10	5.9e-08	8.5e-09	0.22
ZFP161.p2	126	2757	4.6e-10	8.5e-08	1.1e-08	0.19
HIC1.p2	152	3582	8.7e-10	1.6e-07	1.8e-08	0.23
GTF2I.p2	222	5964	3.7e-09	6.8e-07	6.8e-08	0.33

(a) R6/2 group 1 down

	Overlap	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
set_name						
TFDP1.p2	291	7949	5.7e-11	1.1e-08	1.1e-08	0.48
ZNF143.p2	82	1470	3.5e-10	6.5e-08	3.3e-08	0.14
ELK1,4_GABP{A,B1}.p3	175	4314	4.6e-09	8.6e-07	2.9e-07	0.29
ATF5_CREB3.p2	88	1775	2.3e-08	4.3e-06	1.1e-06	0.14
NRF1.p2	123	2810	3.6e-08	6.6e-06	1.3e-06	0.2
SP1.p2	380	11760	8.3e-08	1.6e-05	2.6e-06	0.63
PAX5.p2	190	5129	7.9e-07	0.00015	2.1e-05	0.31
ELF1,2,4.p2	189	5143	1.5e-06	0.00028	3.5e-05	0.31
TFAP2{A,C}.p2	260	7721	7.4e-06	0.0014	0.00015	0.43
MYB.p2	34	608	6.4e-05	0.012	0.0012	0.056

(b) R6/2 group 1 up

Table 6.13: Over-representation of motifs in promoter regions in R6/2

	Overlap	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
set_name						
TFDP1.p2	269	7949	5.4e-11	1e-08	1e-08	0.42
TFAP2{A,C}.p2	257	7721	1.7e-09	3.2e-07	1.6e-07	0.4
TFAP2B.p2	206	5916	9e-09	1.7e-06	5.5e-07	0.32
ZFP161.p2	113	2757	3.2e-08	5.9e-06	1.5e-06	0.17
PATZ1.p2	305	9909	7.4e-08	1.4e-05	2.8e-06	0.47
EGR1..3.p2	126	3378	8.7e-07	0.00016	2.7e-05	0.2
MAZ.p2	314	10505	1e-06	0.00019	2.7e-05	0.49
HIC1.p2	129	3582	4e-06	0.00074	9.2e-05	0.2
PAX5.p2	171	5129	7.1e-06	0.0013	0.00015	0.26
SP1.p2	339	11760	1e-05	0.0019	0.00019	0.52

(a) R6/2 group 2 down

	Overlap	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
set_name						
NRF1.p2	134	2810	1.8e-10	3.3e-08	1.8e-08	0.21
TFDP1.p2	296	7949	1.9e-10	3.5e-08	1.8e-08	0.46
SP1.p2	395	11760	6.8e-09	1.3e-06	4.2e-07	0.62
PAX5.p2	199	5129	1e-07	1.9e-05	4.7e-06	0.31
TFAP2B.p2	220	5916	4.4e-07	8.1e-05	1.6e-05	0.34
MAZ.p2	350	10505	1.2e-06	0.00022	3.6e-05	0.55
TFAP2{A,C}.p2	268	7721	4.4e-06	0.00082	0.00012	0.42
ZNF143.p2	68	1470	2.9e-05	0.0054	0.00059	0.11
ELK1,4_GABP{A,B1}.p3	161	4314	3.1e-05	0.0058	0.00059	0.25
HIC1.p2	138	3582	3.2e-05	0.0059	0.00059	0.21

(b) R6/2 group 2 up

Table 6.14: Over-representation of motifs in promoter regions in R6/2

	Overlap	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
set_name						
TFAP2B.p2	255	5916	9e-13	1.7e-10	1.7e-10	0.35
TFAP2{A,C}.p2	305	7721	5.8e-11	1.1e-08	5.4e-09	0.42
TFDP1.p2	311	7949	1e-10	1.9e-08	6.2e-09	0.43
HIC1.p2	164	3582	1.3e-09	2.5e-07	6.2e-08	0.23
ZFP161.p2	131	2757	1.2e-08	2.3e-06	4.5e-07	0.18
MAZ.p2	376	10505	2.5e-08	4.6e-06	7.6e-07	0.52
PAX5.p2	210	5129	3.9e-08	7.3e-06	1e-06	0.29
MTF1.p2	111	2283	6.5e-08	1.2e-05	1.5e-06	0.15
EGR1..3.p2	148	3378	2e-07	3.7e-05	4.1e-06	0.2
PATZ1.p2	351	9909	7.2e-07	0.00013	1.3e-05	0.48

(a) CHL2 down

	Overlap	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
set_name						
SP1.p2	357	11760	1.4e-09	2.6e-07	2.6e-07	0.66
ELK1,4_GABP{A,B1}.p3	160	4314	1.8e-08	3.4e-06	1.7e-06	0.29
MAZ.p2	319	10505	9.9e-08	1.8e-05	6.1e-06	0.59
TFDP1.p2	252	7949	4.2e-07	7.8e-05	1.9e-05	0.46
KLF4.p3	314	10533	1.4e-06	0.00026	5.2e-05	0.58
PAX5.p2	174	5129	1.8e-06	0.00033	5.6e-05	0.32
PATZ1.p2	296	9909	4.4e-06	0.00082	0.00012	0.54
TFAP2{A,C}.p2	239	7721	9.2e-06	0.0017	0.00021	0.44
TFAP2B.p2	190	5916	1.8e-05	0.0034	0.00037	0.35
ELF1,2,4.p2	169	5143	2e-05	0.0037	0.00037	0.31

(b) CHL2 up

Table 6.15: Over-representation of motifs in promoter regions in CHL2

log2 fold difference vs. median: Cell Type	direction	1	2	3
D1	higher	2326	987	444
	lower	5406	2253	864
D2	higher	2315	994	470
	lower	5475	2365	949

Table 6.16: Numbers of genes with selective expression in D1- and D2- medium spiny neurons based on Doyle data.

genes in HD and HD models [131]. We sought to examine whether this relationship holds more generally, for the full set of striatal-selective genes and the full set of HD-dysregulated genes. To assess this, we made use of two datasets describing wild-type expression across the mouse brain. The Brainstars project [66] profiled expression in 51 anatomically distinct regions, while Doyle et al. [30] used the TRAP approach [49] to measure cell-type expression in 20 cell types. Both studies measured expression using Affymetrix Mouse 430 2.0 microarrays.

To quantify selective expression in the relevant regions and cell types, we employed two metrics. The simplest approach was to compute the difference in expression to the median expression level in all regions. Alternatively, to capture genes with truly selective, rather than just relatively high expression, we computed the difference between each region or cell-type of interest and the second- or third-nearest other cell type or region. (The comparison was made to the second-nearest region to account for the presence of the two striatal MSN cell types in the Doyle data, which are known to be quite similar, and nearby anatomical regions of the striatum in Brainstars.)

Table 6.16 provides an overview of the numbers of selectively expressed genes, based on the distance-from-median metric in the Doyle data. Depending on the fold-difference threshold chosen, there are between several hundred and several thousand genes that might be considered to be selectively expressed.

Figures 6.2 and 6.3 show representative scatterplots comparing the difference-from-median selectivity to fold-change in the R6-2 and YAC128 models (for probe-sets having differential expression at a (nominal) alpha of 0.01). Among probesets that are differentially expressed in HD models, those down-regulated in HD have a strong tendency to

have have high relative expression in striatal cell types, and vice versa. Although this is not particularly surprising, it confirms that previous observations about the changes to the top striatal-specific marker genes in HD [131] apply across the entire striatal transcriptome.

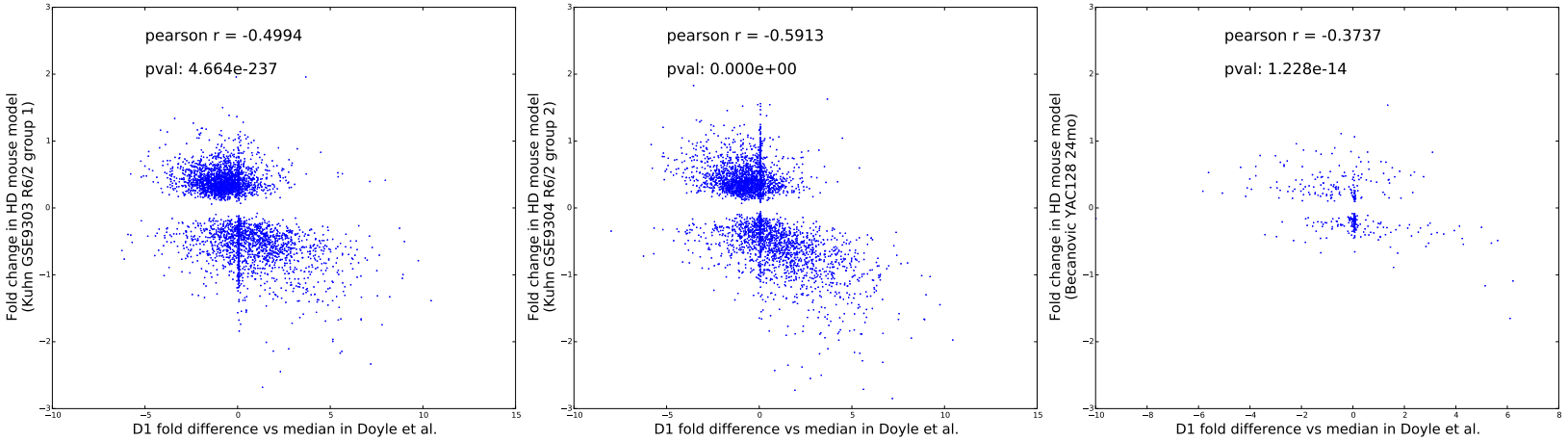


Figure 6.2: Scatterplots comparing selectivity of expression in D1 MSNs (Doyle et al.) to differential expression in HD models

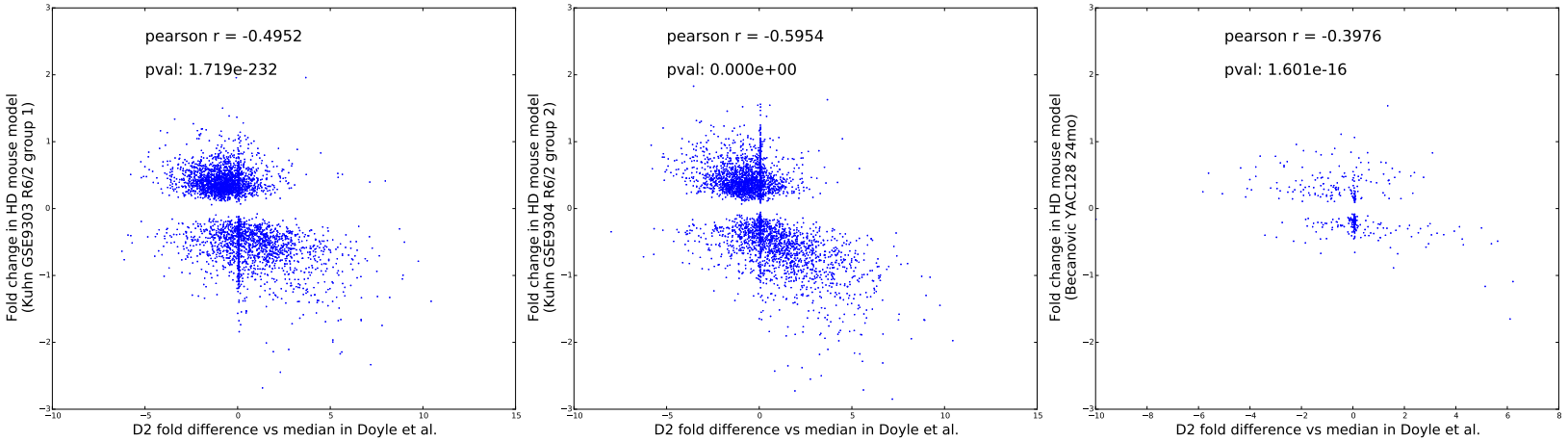


Figure 6.3: Scatterplots comparing selectivity of expression in D2 MSNs (Doyle et al.) to differential expression in HD models

6.4.5 Over-representation analyses of genes with selective neuronal and striatal expression

Given the relationship between differential expression in HD and the background set of genes with striatal-specific expression, we repeated our motif and regulator over-representation analyses against the sets of MSN-selective genes on their own. Tables 6.17 and 6.18 show the top potential regulators from the ChEA database over-represented within the set of genes with striatal-selective expression (2-fold difference from median cell type) in D1 and D2 MSNs, respectively. Most of the regulators with over-represented targets are the same as those observed in the analysis of genes differentially expressed in the HD models. In particular, targets of PRC2 components and associated proteins (Suz12, Ezh2, Jarid2, Ezh2, Rnf2) are over-represented among the set of genes with high expression in striatal cell-types.

6.5 Discussion

Dysregulated transcriptional regulators and motif over-representation

Our analysis of differentially expressed transcriptional regulators and over-represented motifs among HD-dysregulated genes identified a smattering of potentially relevant genes and transcription factors, but ultimately fails to explain much about the mechanisms or specificity of mHTT-dependent transcriptional dysregulation. One limitation is that the majority of significant expression changes were observed at a single time-point late in disease, which makes it impossible to even speculate about ordering and causality of changes. We will attempt to partially address this limitation using of time-resolved data in the next two chapters. Second, it remains difficult to integrate motif information with changes in transcription factor expression and activity. Many of the targets of transcription factors remain unknown. It is also unclear which motif occurrences are relevant in the cell-types and developmental stages of interest. This might eventually be improved with better models that account for interactions between multiple motifs, transcription factors, and effects on expression learned from systematic measurements across many cell types, as well as by experimentally measuring chromatin accessibility and DNA footprints, to

set_name	match_count	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
SUZ12	407	5042	4.6e-51	6.1e-49	6.1e-49	0.48
MTF2	278	2981	1.1e-42	1.4e-40	7e-41	0.33
JARID2	163	1639	1.2e-26	1.6e-24	5.4e-25	0.19
EZH2	131	1300	8.6e-22	1.1e-19	2.9e-20	0.15
RNF2	165	1975	5.8e-19	7.7e-17	1.5e-17	0.19
BMI1	126	1682	3.8e-11	5e-09	8.4e-10	0.15
WT1	121	1663	6e-10	8e-08	1.1e-08	0.14
YAP1	155	2329	1.1e-09	1.5e-07	1.8e-08	0.18
DMRT1	144	2144	2.7e-09	3.6e-07	4e-08	0.17
TP53	229	3937	1.2e-08	1.6e-06	1.6e-07	0.27

(a) ChEA over-representation in D1 selective (high-expression) genes

set_name	match_count	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
NRF2	166	1055	7e-08	9.3e-06	3.8e-06	0.076
NFE2L2	166	1055	7e-08	9.3e-06	3.8e-06	0.076
ASH2L	439	3336	1.1e-07	1.5e-05	3.8e-06	0.2
MYB	148	923	1.2e-07	1.5e-05	3.8e-06	0.068
E2F1	533	4172	1.8e-07	2.4e-05	4.5e-06	0.24
MTF2	396	2981	2e-07	2.7e-05	4.5e-06	0.18
SFPI1	315	2369	4.3e-06	0.00058	8.2e-05	0.14
PPARD	41	194	1.1e-05	0.0015	0.00018	0.019
TAL1	449	3578	1.6e-05	0.0022	0.00024	0.21
SMARCA4	435	3481	3.5e-05	0.0046	0.00042	0.2

(b) ChEA over-representation in D1 selective (low-expression) genes

Table 6.17: D1 ChEA

set_name	match_count	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
SUZ12	423	5042	1.1e-53	1.4e-51	1.4e-51	0.48
MTF2	300	2981	2.4e-50	3.2e-48	1.6e-48	0.34
JARID2	174	1639	4.3e-30	5.7e-28	1.9e-28	0.2
EZH2	141	1300	4.2e-25	5.6e-23	1.4e-23	0.16
RNF2	172	1975	4.6e-20	6.1e-18	1.2e-18	0.19
BMI1	133	1682	2.6e-12	3.4e-10	5.7e-11	0.15
WT1	127	1663	8.9e-11	1.2e-08	1.7e-09	0.14
YAP1	163	2329	1.2e-10	1.5e-08	1.9e-09	0.18
DMRT1	152	2144	2.3e-10	3.1e-08	3.4e-09	0.17
EED	74	830	2e-09	2.7e-07	2.7e-08	0.084

(a) ChEA over-representation in D1 selective (high-expression) genes

set_name	match_count	size of set	hypergeom p-val	bonferroni	b-h fdr adj pval	Fraction
MTF2	458	2981	3.8e-14	5.1e-12	5.1e-12	0.2
NRF2	185	1055	2.1e-10	2.8e-08	9.3e-09	0.08
NFE2L2	185	1055	2.1e-10	2.8e-08	9.3e-09	0.08
SUZ12	687	5042	3.9e-10	5.2e-08	1.3e-08	0.3
MYB	158	923	2.5e-08	3.4e-06	6.7e-07	0.068
SMARCA4	479	3481	1.6e-07	2.1e-05	3.6e-06	0.21
JARID2	242	1639	2.5e-06	0.00033	4.7e-05	0.1
E2F1	550	4172	4.3e-06	0.00057	7.1e-05	0.24
DMRT1	303	2144	5.7e-06	0.00076	8.4e-05	0.13
MECOM	278	1951	7.7e-06	0.001	0.0001	0.12

(b) ChEA over-representation in D1 selective (low-expression) genes

Table 6.18: D2 ChEA

identify accessible genomic loci and motifs actually bound by regulators within the cell types of interest.

Targets of the PRC2 complex are over-represented among genes dysregulated in HD models

As reviewed in Chapter 6, functional links between mHTT and the PRC2 complex and its activity have previously been proposed [109, 15]. These connections had been shown *in vitro*, and in mouse embryos, embryonic stem cells and developing neural stem cells. The analysis here provides additional evidence in support of the hypothesis that PRC2 activity (and H3K27 methylation state) is relevant to mHTT-dependent expression changes in both mouse models and human HD.

Our analysis has several limitations which will not be resolved with observational data alone. First, striatal-specfic expression, HD-dependent expression changes, and over-representation of PRC2 targets cannot be untangled. It remains possible that the observed over-representation of PRC2 targets in HD-dysregulated genes is merely a statistical artifact due to the high expression of these particular genes in the striatum. Alternatively, the common PRC2-target overrepresentation is compatible with the hypothesis that PRC2-dependent regulation is particularly important to controlling selective expression of genes in striatal neurons, and that this PRC2 regulation of these genes is affected by mHTT. This would help to explain both the why mHTT affects the genes that it does, and why striatal neurons are selectively vulnerable.

A second caveat is that the collections of potential regulators against which genes were compared are biased towards experimentally accessible systems and regulators that have been of recent biological interest. A majority of the relevant ChIP datasets were performed in mice and mouse cell lines, across a variety of developmental stages, primarily in early development. It is very likely that many of the targets of these regulators will be different in mature neurons. It is also questionable to assume that regulatory interactions observed in the mouse will be conserved in human cells, though many probably are.

To address these limitations, experimentation will be needed to more definitively test the involvement of PRC2 activity in the transcriptional dysregulation caused by mHTT. It

would be informative to measure the epigenetic states (especially H3K27me3) in normal and HD striatum, or in cultured cells with and without expression of mHTT. One of the simplest hypotheses is that PRC2 activity is increased in the presence of mHTT, which leads to increased H3K27 tri-methylation at PRC2 targets in MSNs, repressing their expression. Observing an increase in H3K27me3 marks at the loci of genes down-regulated in HD in the relevant cells would provide much more convincing evidence for this hypothesis. One could then attempt to further prove and dissect the functional connection between mHTT and the changes in H3K27me3 state and expression by blocking the activity of PRC2 catalytic components and other regulators, using small-molecule inhibitors or genetic methods.

Chapter 7

Re-analysis of Cell-Type Specific Expression in Mouse HD Models (Fenster TRAP Study)

7.1 Introduction

Previous studies of transcriptional changes in mouse models of HD have relied on homogenized tissue, so potential cell-type specific transcriptional changes may have been obscured, either by differences between neuronal subtypes or expression from glial and other cell types. To address these limitations, Fenster, Heiman and colleagues conducted an experiment using the TRAP methodology [49, 50] to measure cell-type specific translational profiles of the two major cell MSN types, Drd1a and Drd2, in two transgenic mouse models of HD, R6-2 and YAC128 [37]. Translational changes (which we will also refer to as expression changes) were measured both before and after motor symptoms developed, so that potential early, pre-symptomatic changes could be compared to those emerging after development of motor phenotypes and as disease progressed.

Here, we re-examine this data to search for potential cell-type and time-dependent mechanisms of transcriptional dysregulation and to confirm some of the observations from the previous chapter in an independent dataset.

7.2 Experimental Design and Data

An outline of the experimental design is shown in Figure 7.1. The numbers of replicates are balanced for between the HD and control groups the same model and time-point, but some time points have many more replicates than others. It should be noted that the R6-2 and YAC128 models are in different background strains and have phenotypes that

develop over very different time scales, so the most meaningful comparisons are between the HD and the wild-type mice of the same model and TRAP cell type.

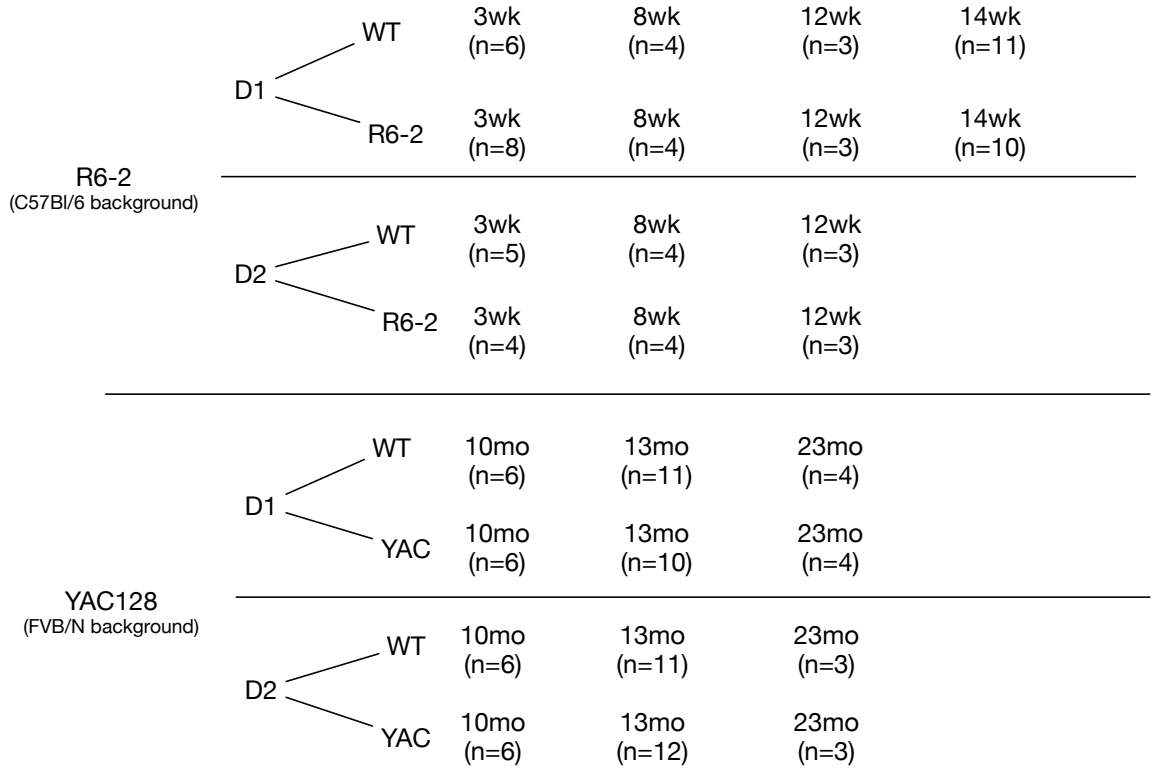


Figure 7.1: Outline of organization and number of samples per group in Fenster HD TRAP experiment.

7.3 Methods

The complete description of experimental methods, and a discussion of the motor phenotypes and the selection of the time points from which expression was assayed can be found in Fenster, 2011 [37].

7.3.1 Microarray Normalization and Quality Control

Gene expression was measured using Affymetrix Mouse 430 2.0 microarrays. Affymetrix .CEL files were processed using the Bioconductor *affy* package [41] and RMA normalization. Probe sets were mapped to gene symbols using the annotation in the Bioconductor mouse4302.db annotation package, downloaded in July 2013.

Since one objective of this experiment was to characterize expression changes across time and to compare changes between different models and cell types, we initially processed and normalized all samples together. This assumed that all samples (which are from different but closely related cell types, and two different mouse background strains) are similar enough that the majority of genes (unaffected by mHTT) are expressed at equivalent levels.

However, in some cases we observed differences (generally an increase in the number of reported significant changes) when the arrays from individual cell types and/or timepoints were processed on their own. This may reflect biological differences between cell types that affect the RMA normalization procedures, or other batch effects. In addition, probes with signals never observed above the 20th percentile in at least two samples were excluded from analysis, so slightly fewer such genes were filtered when using the combined samples.

For simplicity, and to facilitate comparisons across the entire dataset, we use the globally normalized version of the data for most of the analyses discussed in this chapter. Figures and discussion using the single-context data will be indicated.

7.3.2 Differential expression testing

We tested expression changes in several ways. In the analysis using the normalization over all samples, we first tested differential expression at individual time points using Welch's t-test. We also attempted to combine information over all time points using ANOVA, modeling expression as a function of both HD genotype and time. Given the limited number of replicates at some time points, and the observation that some probe sets have modest changes that are consistent in direction over time, the multiple-testing adjusted (Type I) ANOVA provides a reasonable alternative way to summarize expression changes, potentially capturing both global and time-specific differences. A probe-set was considered to be up- or down- regulated if either the HD genotype status main effect or the genotype-time interaction was significant, and the direction was assigned based on the sign of the largest change over the time points. In many cases, the ANOVA approach improved power and permitted more stringent control of false discovery rate across all

models and cell types.

For the analysis restricted to the individual time points and contexts, we used the the moderated t-test implemented by *limma* [116], which produces a small improvement in the apparent power and therefore more genes called significant at an FDR of 0.05.

Statistics were adjusted for multiple testing over probe sets using the Benjamini-Hochberg method [13]. Since we wished to control false discoveries within each experimental group, rather than globally, and given that power was already limited by the modest numbers of samples, no further adjustments were made for the multiplicity of time points and models.

7.3.3 Motif and Regulatory Overrepresentation Analyses

Over-representation of motifs (based on SwissRegulon) and associations with chromatin binding proteins from the ChEA database [71] were assessed as described in Chapter 6. We once again relied on the hypergeometric test since the small number of samples in many of the contexts was not conducive to methods based on permutation testing.

7.4 Results

Table 7.1 summarizes the numbers of significant expression changes (at α = 0.05) in each condition under each testing approach, both with and without Benjamini-Hochberg adjustment. Significant expression changes tend to develop in the same direction across time (Fig 2A), and a majority of changes emerge at post-symptomatic time points. Although one of our objectives was to compare changes across both models and time, this is complicated by differences in power at different time points, largely due to unavoidable experimental limitations. For example, since few YAC128 mice survive to 23 months of age, statistical power at that time point was limited, which accounts for the apparent decrease in the number of significant changes between 13 months and the terminal 23 month point.

			Number of Significantly Changed Probesets and Genes, by HD Model and TRAP cell type															
			Welch t-tests, individual time points								2-way ANOVA; genotype or genotype-time interaction							
			# of mice	Nominal p-value < 0.05				Adjusted p-value < 0.05				Nominal p-value < 0.05				Adjusted p-value < 0.05			
			(wt / HD)	Down		Up		Down		Up		Down		Up		Down		Up	
				Genes	Probes	Genes	Probes	Genes	Probes	Genes	Probes	Genes	Probes	Genes	Probes	Genes	Probes	Genes	Probes
R6-2	D1	03 wk	(6/8)	753	901	839	1021	0	0	0	0	3959	5423	3107	4022	170	225	104	124
		08 wk	(4/4)	788	972	613	756	0	0	0	0								
		12 wk	(3/3)	982	1228	618	718	0	0	0	0								
		14 wk	(11/10)	1954	2646	1583	2055	117	153	77	92								
	D2	03 wk	(5/4)	719	864	753	929	0	0	0	0	3234	4366	3036	3993	183	240	170	203
		08 wk	(4/4)	738	929	581	712	0	0	0	0								
		12 wk	(3/3)	3033	4100	3918	5357	2	2	7	7								
YAC	D1	10 mo	(6/6)	1071	1318	814	979	0	0	0	0	3086	4188	3301	4250	297	385	329	400
		13 mo	(11/10)	1689	2286	2024	2576	114	146	189	229								
		23 mo	(4/4)	1093	1421	988	1211	0	0	0	0								
	D2	10 mo	(6/6)	771	963	686	856	0	0	0	0	3045	3959	2988	3780	56	65	57	71
		13 mo	(11/12)	1460	1865	1587	2020	13	17	18	20								
		23 mo	(3/3)	663	852	590	713	0	0	0	0								

Table 7.1: Summary of numbers of genes changing across the different conditions studied, with normalization over all samples and timepoints, using either Welch's t-tests for individual contrasts or ANOVA.

Table 7.2 summarizes the number of expression changes between the HD model and corresponding wild-type control for each group using the context-specific analysis and moderated t-tests. Numbers of up- and down-regulated genes and probe sets are shown for α of 0.05 and 0.10, with and without Benjamini-Hochberg adjustments, and counts are shown for both changes of any magnitude and change of at least 2-fold. All of the the models and cell types studied exhibited changes at some time point, even after multiple-testing adjustments.

7.4.1 Validation against previous mouse HD expression data

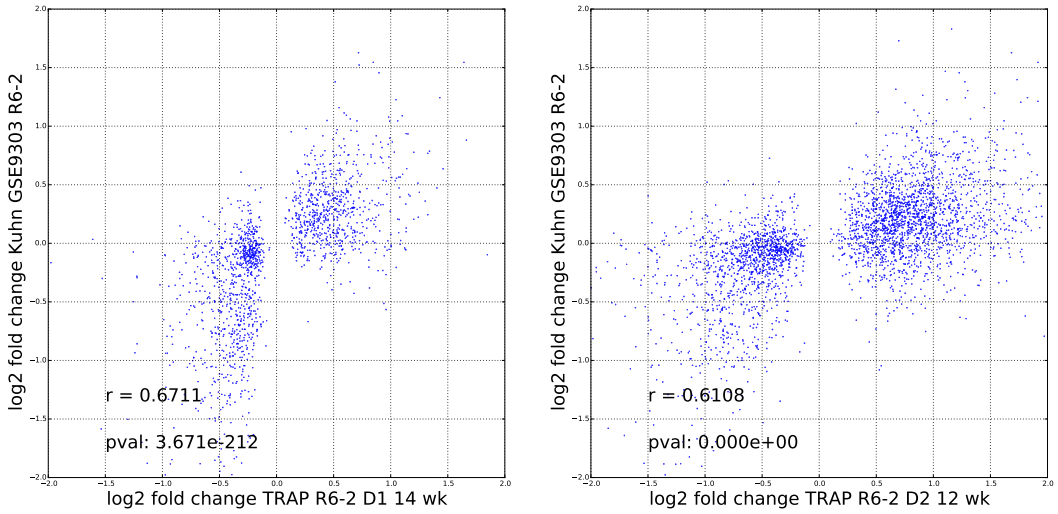
The expression changes in the TRAP experiment were compared to those from earlier studies using homogenized tissue reported by Kuhn and colleagues, which were discussed in Chapter 6. As expected, there was significant correlation between expression changes in corresponding models, especially at the later time points, for both models and cell types studied. These comparisons are summarized in the scatterplots of Figure 7.2. The R6-2 data had slightly greater correlation, which probably reflects a difference in the extent of disease progression at the time points sampled, rather than any biological difference between the models.

7.4.2 Validation against expression changes in Human HD studies

Earlier studies have noted the considerable (though imperfect) agreement between expression changes in mouse HD models and changes in human HD striatum. Figure 7.3 shows scatterplots comparing fold-changes between the mouse cells and human striatum (from Hodges [53]), for orthologous genes with nominally significant changes (α = 0.01) at selected, late time points in both the mouse and in human striatum. For such genes, there are two to three times as many genes changing in the same direction (numbers in black) in the human as there are genes changing in opposing directions (numbers in red). Both D1 and D2 cells show this concordance to the human data.

				# of Genes								# of Probes							
				down				up				down				up			
				bh		nominal		bh		nominal		bh		nominal		bh		nominal	
p-val	model	cell type	min log2 FC:	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
0.05	R6-2	D1	03 wk	0	0	809	77	0	0	698	87	0	0	932	85	0	0	838	98
			08 wk	0	0	883	164	0	0	634	117	0	0	1032	183	0	0	728	125
			12 wk	0	0	1031	335	0	0	1031	495	0	0	1248	384	0	0	1181	551
			14 wk	159	11	1870	14	131	20	1663	39	227	17	2536	21	159	26	2136	51
		D2	03 wk	0	0	789	186	0	0	745	201	0	0	918	208	0	0	854	219
			08 wk	0	0	845	244	0	0	669	210	0	0	1008	280	1	1	768	237
			12 wk	3699	179	5712	179	4402	407	5416	408	5153	211	8302	211	6270	510	7970	511
	YAC	D1	10 mo	0	0	1131	33	0	0	848	36	0	0	1348	40	0	0	977	39
			13 mo	307	8	1628	8	524	8	2214	9	394	14	2247	14	638	9	2865	10
			23 mo	3	3	1036	39	3	3	1142	141	5	5	1314	52	3	3	1385	162
		D2	10 mo	0	0	715	23	0	0	715	40	0	0	862	33	0	0	827	46
			13 mo	55	2	1786	4	75	1	1420	3	67	5	2366	8	91	1	1832	5
			23 mo	0	0	729	106	0	0	818	157	0	0	848	122	0	0	931	173
0.10	R6-2	D1	03 wk	0	0	1587	115	0	0	1473	118	0	0	1901	127	0	0	1817	132
			08 wk	0	0	1767	225	0	0	1351	169	0	0	2146	254	0	0	1634	191
			12 wk	0	0	2125	493	1	1	1855	721	0	0	2638	566	1	1	2222	823
			14 wk	285	11	2991	15	269	27	2516	39	399	17	4122	22	309	34	3322	51
		D2	03 wk	0	0	1591	283	0	0	1575	295	0	0	1917	323	0	0	1909	341
			08 wk	0	0	1622	361	0	0	1390	302	0	0	1975	408	1	1	1674	347
			12 wk	5389	179	7261	179	5266	408	6102	408	7773	211	10758	211	7706	511	9193	511
	YAC	D1	10 mo	1	1	2120	34	0	0	1571	38	2	2	2598	41	0	0	1887	42
			13 mo	457	8	2497	8	809	9	3135	9	592	14	3531	14	986	10	4172	10
			23 mo	3	3	2019	50	4	4	1994	159	7	7	2640	64	4	4	2499	187
		D2	10 mo	0	0	1408	26	0	0	1495	44	0	0	1736	36	0	0	1784	51
			13 mo	112	3	2919	4	145	1	2131	3	140	6	3994	8	170	1	2835	5
			23 mo	0	0	1536	151	0	0	1606	203	0	0	1855	172	0	0	1907	229

Table 7.2: Summary of numbers of genes and probe sets changing across conditions studied, using data normalized for each context and time point independently. Differential expression tested using *limma* moderated t-tests.



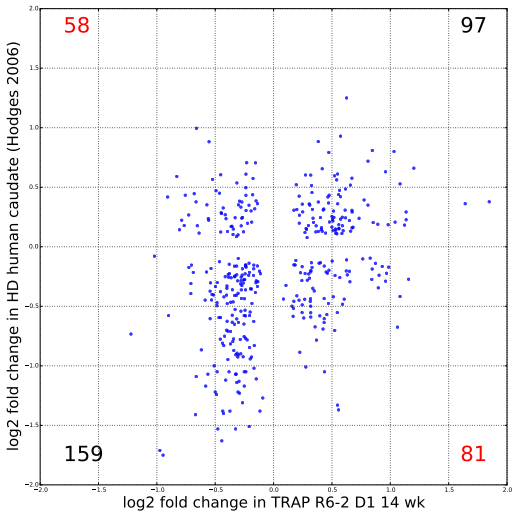
(a) R6-2 D1 vs. Kuhn R6-2

(b) R6-2 D2 vs. Kuhn R6-2

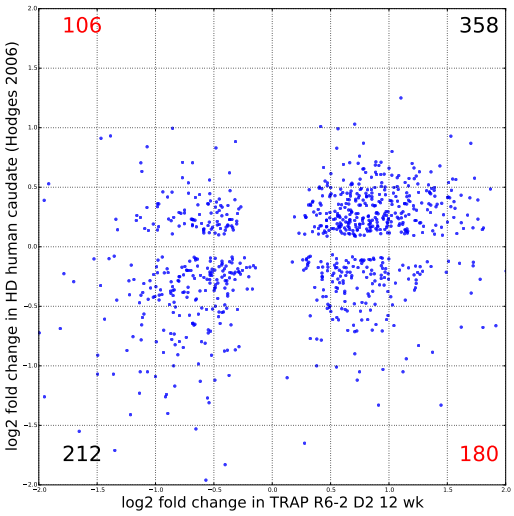
(c) YAC D1 vs. Becanovic YAC

(d) YAC D2 vs. Becanovic YAC

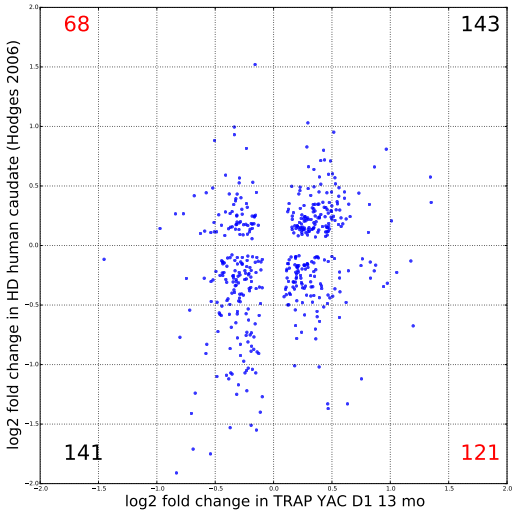
Figure 7.2: Scatterplots comparing cell-type specific TRAP (at late, representative time points) to previously published homogenized tissue data from the same mouse models.



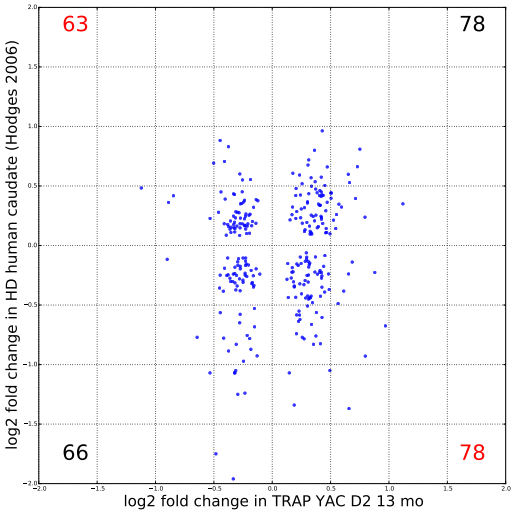
(a) R6-2 D1 (14 wk) vs. Hodges (human HD striatum)



(b) R6-2 D2 (12 wk) vs. Hodges (human HD striatum)

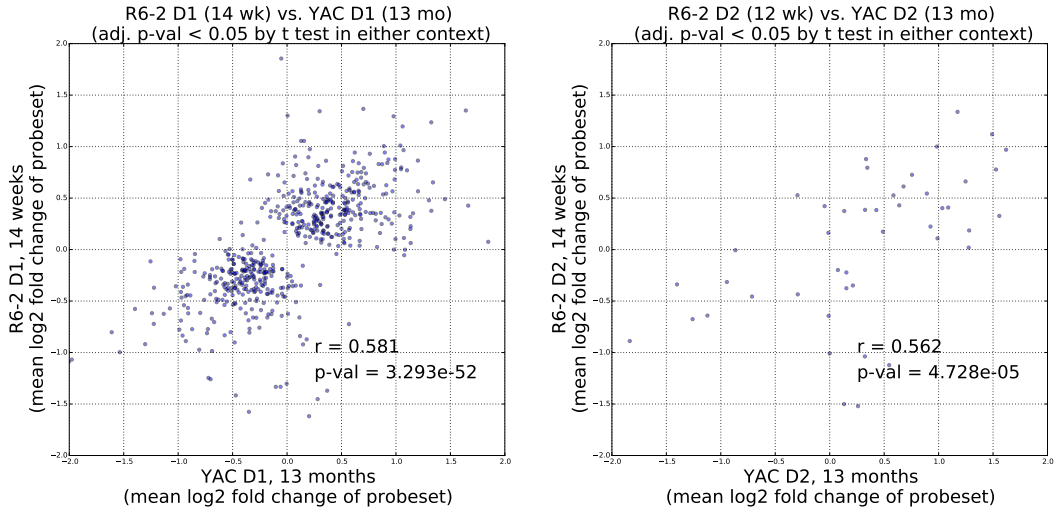


(c) YAC D1 (13 mo) vs. Hodges (human HD striatum)



(d) YAC D2 (13 mo) vs. Hodges (human HD striatum)

Figure 7.3: Concordant and discordant expression of orthologs in human HD (striatum) and mouse models. Scatterplots of log2 fold changes in from TRAP in HD vice vs. log2 fold change of orthologous gene in human HD samples, for differentially expressed genes at a nominal p-value threshold of 0.01 for both groups. The number of genes falling in each quadrant is indicated (black numbers are concordant expression changes; red indicate discordant expression changes.



(a) R6-2 D1 (14 wk) vs. YAC D1 (13 mo) (b) R6-2 D2 (12 wk) vs. YAC D2 (13 mo)

Figure 7.4: Scatterplot comparing significant changes between R6-2 and YAC models in corresponding cell-types

7.4.3 Differences between HD models

The earlier studies that compared different mouse models of HD [69] found substantial concordance between full-length (YAC128) and Exon 1 (R6-2) models of the disease. This is reinforced by analysis of the TRAP data. Figure 7.4 shows that at the late timepoints, the correlation of fold-changes between the models is highly significant (for probe-sets that are differentially expressed in at least one of the conditions).

7.4.4 Many HD dysregulated genes have MSN-specific expression

We and others had previously observed that many of the genes dysregulated in HD models are those with striatal-specific expression. To verify this and ask whether there was a difference in this selectivity between the cell types, we compared the post-symptomatic expression changes to a metric of selectivity, the wild-type expression vs. mean expression of neurons studied in Doyle et al [30]. We found that in both cell types, genes that are down-regulated in the HD condition tend to be those with relatively selective wild-type expression, and vice versa. This suggests that one of the effects of mHTT is a loss of MSN identify, and not simply a generic toxic effect, and perhaps that there are transcriptional

regulatory mechanisms essential to maintaining MSN identity that are disrupted by mHTT.

7.4.5 Expression changes in D1 vs D2 MSNs

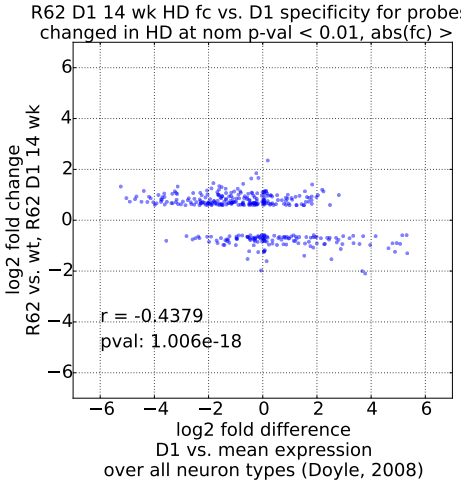
An important question which can be addressed with cell-type specific data is whether expression changes causes by mHTT differ between the distinct MSN cell types. Figure 7.6 shows scatterplots comparing changes observed in D1 vs D2 cells, for representative (late, post-symptomatic) time points in each model. Overall, at these later time points, there is substantial concordance between the changes observed by TRAP in D1 and D2 cells. At least at this global level, there is no evidence for opposing responses between the two cell types, although there are some genes that change more in one cell type vs. the other.

7.4.6 Expression changes over time and multiple contexts

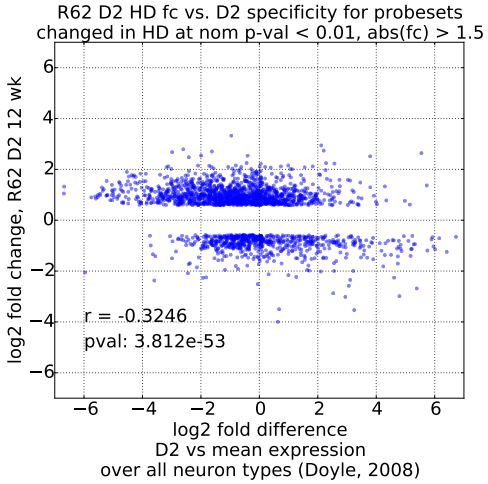
Another question motivating this experiment was to identify expression changes occurring both early and late in progression of the disease model, and assess how these may differ between the two MSN cell types. To address this systematically, we classified probe sets by their patterns of expression changes across the various experimental conditions. While the small number of samples and modest changes at some timepoints limited statistical power, we reasoned that the most biologically interesting genes would be those with changes corroborated across multiple contexts.

We examined the changes at a stringent level of nominal significance ($p < 0.001$, fold-change > 1.2 fold) as well as changes that were consistent between two groups over either cell types or time, at a more relaxed level of significance ($p < 0.05$, fold-change > 1.2 fold), and organized these by their patterns of change over the four contexts (Figure 7.7, panel A). These criteria defined 169 up-regulated genes (201 probe sets) and 166 down-regulated genes (189 probe sets) involved in diverse processes, and a number of these changes appeared to be cell-type specific.

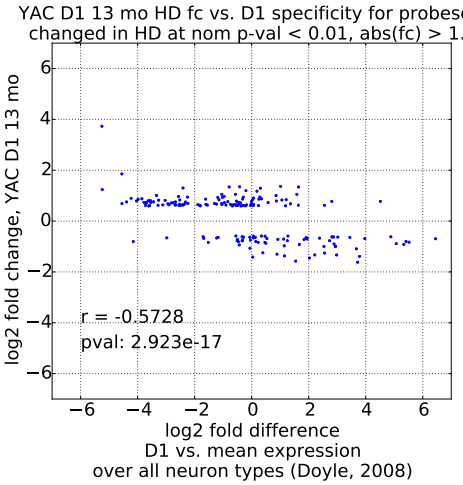
Genes with increased expression at early time-points in R6/2 include some involved in signal transduction (Pde7a, Pde1c, Nek5, Tyk2, Ikbkg, Gna12), neurotrophins (Ntf3), apoptosis signaling (Dapk1), and transcriptional regulation (eg. Zhx2, Zfp566, Zfp608,



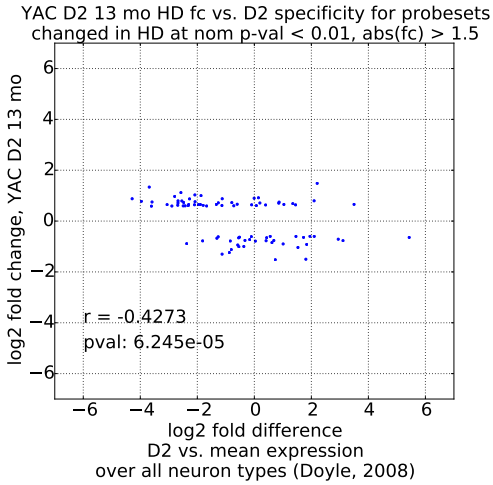
(a) R6-2 D1 (14 wk) vs. D1 selectivity



(b) R6-2 D2 (12 wk) vs. D2 selectivity

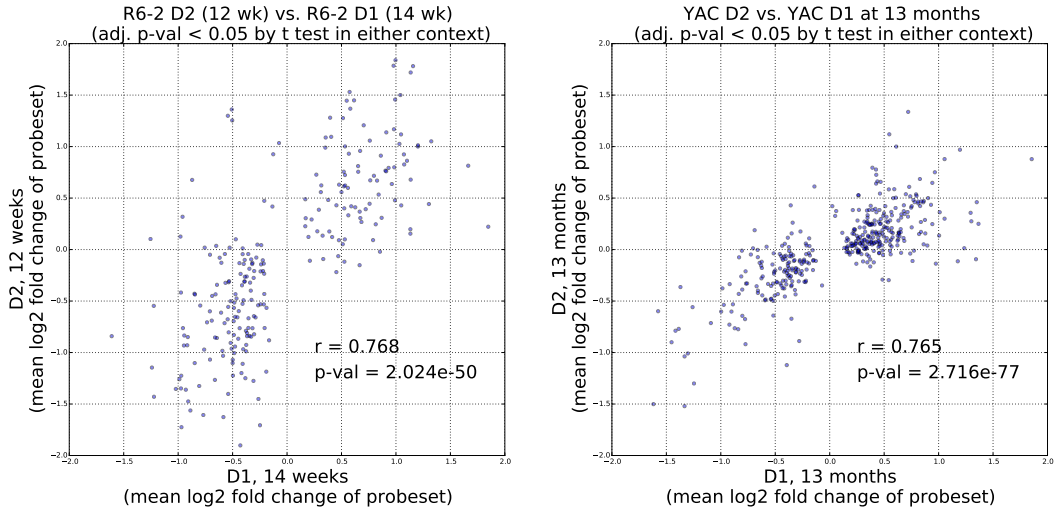


(c) YAC D1 (13 mo) vs. D1 selectivity



(d) YAC D2 (13 mo) vs. D2 selectivity

Figure 7.5: Genes dysregulated in HD models tend to have MSN-specific expression. Expression changes in HD were compared to relative expression across the neuronal cell types profiled in Doyle et al. [30]



(a) R6-2 D1 (14 wk) vs. D2 (12 wk) (b) YAC D1 (13 mo) vs. YAC D2 (13 mo)

Figure 7.6: Scatterplots comparing significant changes D1 vs. D2 cells in corresponding models and time-points

Pou6f1, Zfp949). Among the most significant selectively up-regulated genes in D2 cells (in both YAC and R6/2 models) was the Wnt receptor Ryk, which was recently reported elsewhere to be up-regulated and to promote neuronal dysfunction at early stages of HD pathogenesis [133].

Down-regulated genes at early timepoints in R6/2 models included those involved in signal transduction (eg. Pde10a), neurotransmitter receptors (Grm5), neurotransmission (Syt7), chromosome organization (Syce2), and regulation of the cytoskeleton (Tbce, Slain2).

At the 10 month point in YAC128 model, many more significant expression changes are evident, and this time-point likely represents a disease state that is further progressed compared to the 3 week pre-symptomatic point in R6/2. Among up-regulated genes in D1 cells are a number of genes involved in signal transduction (Rnd3, Nek7), apoptosis (Dlc1, Rprm), cell cycle regulation (Cdk14), transcription factors (Zfp516), and a large number of protocadherin genes (Pcdh18, Pcdh19, Pcdh20, Pcdhb16).

Comparing across models and cell types at the pre-symptomatic time points, there were few common changes. At a threshold of 1.5 fold-change, Igfbp4, Gm8154, Ddit4l, Oprk1, and Col6a4 were down-regulated in both D1 and D2 cells; no genes had common

increases.

At the late, post-symptomatic timepoints, there were several thousand genes that changed. Figure 7.7B summarizes those with the most significant changes, based on the ANOVA model, after multiple-testing adjustment. Roughly similar numbers of genes were up- and down- regulated using these statistical thresholds, although a greater number of down-regulated genes are among those with the most extreme fold-changes. Many genes changed in the same direction in both striatonigral and striatopallidal MSNs, and only a handful of genes (eg. Atp2b1, Gpm6b, Ncam1, and Ttc3) had probe sets significantly changed in opposite directions between cell types.

7.4.7 Pathway over-representation analyses

There are well-known limitations to biological inferences based on database annotations of function [112]. Nevertheless, to summarize the large number of expression changes observed at the later time points, we also assessed over-representation of gene sets associated with biological pathways from KEGG [65]. Figure 7.8 shows a graphical summary of the results. As might have been expected, many of the pathways with the most significant over-representation among dysregulated genes are involved with neuronal signaling and neurological disease, such as sets involved with "gap junction", "glutamatergic synapse", "neurotrophin signaling", the "dopaminergic synapse", and "calcium signaling".

7.4.8 Expression changes of transcriptional factors and regulators

To identify genes that might be regulating transcriptional dysregulation, we also looked specifically at changes involving transcription factors and other genes with DNA or chromatin-binding activity. Genes with these annotations and which had large (> 3-fold) and (nominally) significant changes in at least one experimental context are shown in Figure 7.9.

Among regulators with the largest or more consistent changes, Wt1, which is very highly up-regulated especially in D1 cells, is a transcription factor better known for its role in kidney development and Wilms tumor, an inherited form of kidney cancer. However, Wt1 has also been previously implicated in neurodegeneration in Alzheimer's disease [78].

Among chromatin modifying genes, Kdm6b (Jmjd3), a lysine demethylase that acts

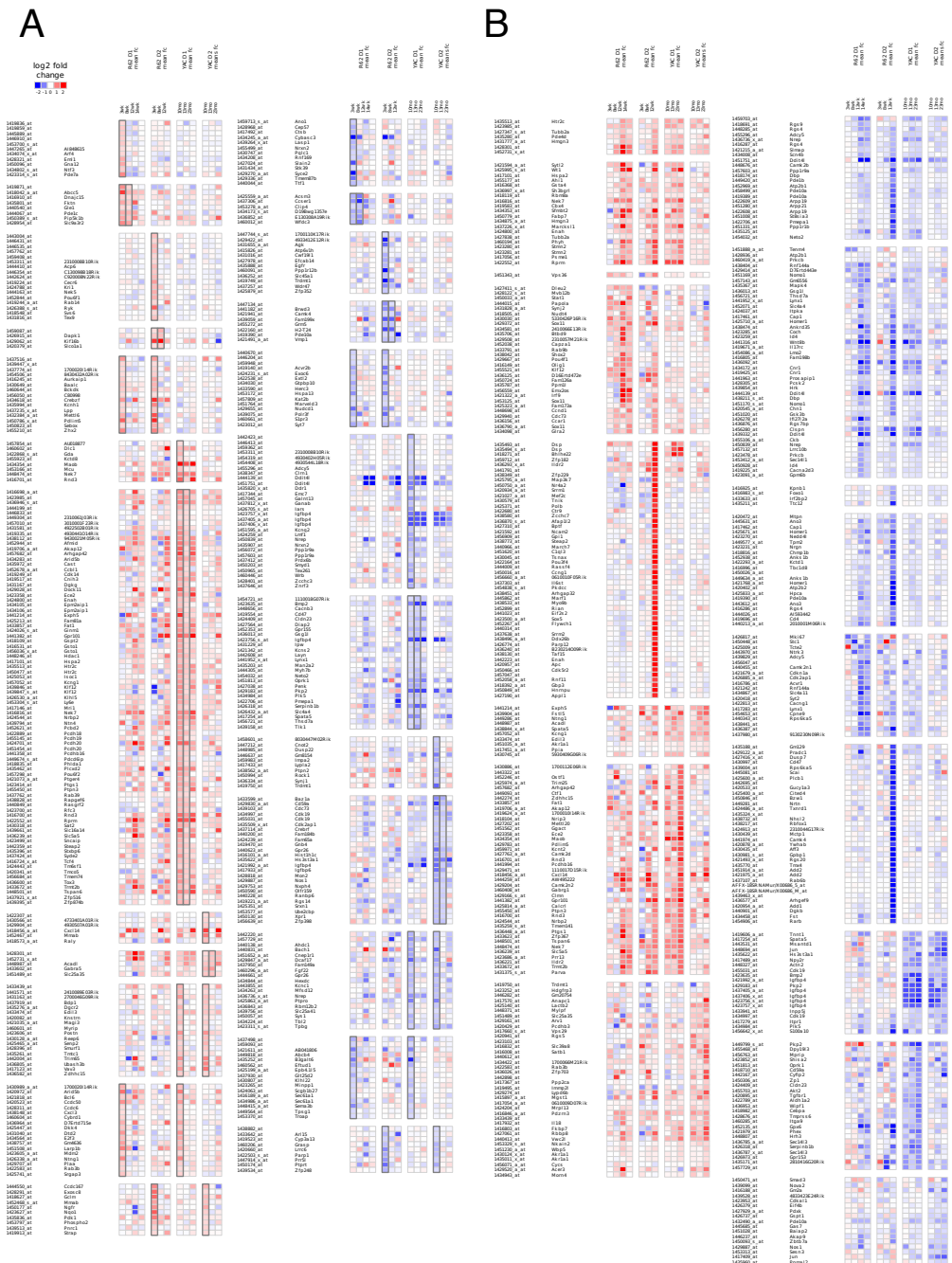


Figure 7.7: Heat map summarizing putative A) early and B) late expression changes in HD models, grouped by patterns of significant changes across models and time-points.

Figure 7.8: Summary of KEGG pathway over-representation over HD models and time-points.

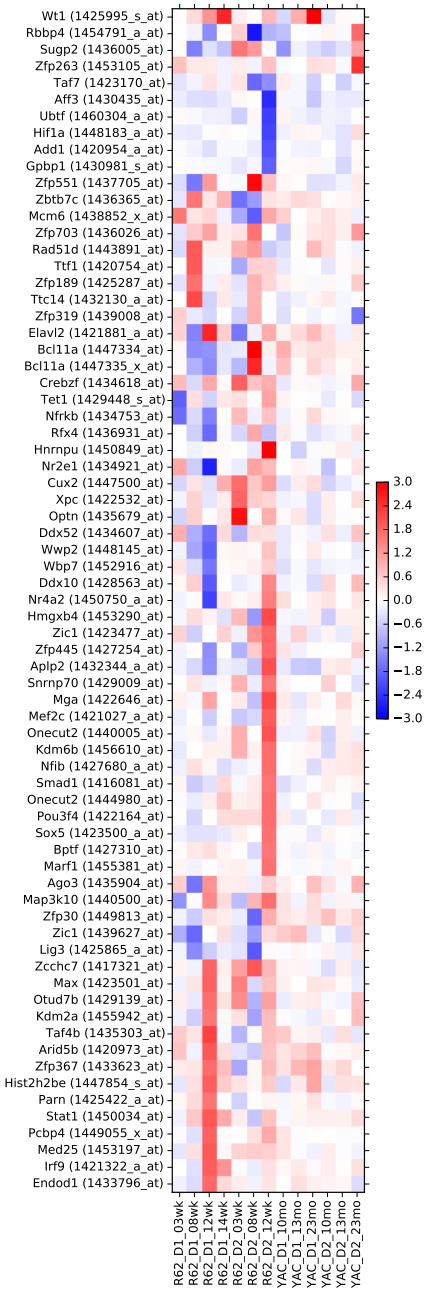


Figure 7.9: Putative changes in transcription regulation / chromatin binding related genes. Probe sets were filtered to show those with a nominal p-value < 0.001 and a fold change of at least 3-fold in at least one context, based on the independently normalized *limma* analysis.

on H3K27me2/3, appears to be up-regulated, as may be Kdm2a, a H3K36 demethylase. These might reflect homeostatic responses to increases in PRC2 activity, if indeed that is a consequence of mHTT.

7.4.9 Motif and regulatory target overrepresentation analysis

We next assessed over-representation of regulatory motifs and chromatin-binding regulators across the various models. Figure 7.10 shows a summary of the over-represented motifs, and Figure 7.11 shows over-represented regulators from the ChEA database, based on the sets of genes differentially expressed in each group at an adjusted α of 0.05.

The analysis of over-representation of the targets of chromatin-binding regulators is shown in Figure 7.11. Many of the most highly-ranked putative regulators in this analysis were the same as those found using data from homogenized tissue discussed in the previous chapter. Among the top potential regulators were components of the PRC2 complex: MTF2, SUZ12, JARID2, RNF2, and EZH2. The over-representation of targets of these proteins was most apparent among the sets of down-regulated genes. Although the test statistics tended to be more extreme in the D1 cell types, the differences in power between groups (due both to sample size and potential differences in expression of the TRAP constructs) makes it difficult to make strong claims about whether this difference is statistically significant and biologically important.

Given the large expression changes in expression of the Wt1 gene itself, it is also interesting to note that Wt1 targets are over-represented in many of the groups. While the extreme changes in Wt1 has been observed in many earlier studies ([12]), there is still relatively little known about its role in neural gene expression networks.

7.5 Discussion

7.5.1 Considerations for design of future experiments

Unbalanced numbers of replicates across experimental groups makes rigorous comparisons between groups and cell-types difficult. Given the biological variability in HD mouse models, it seems that having at least a dozen biological replicates is desirable, especially

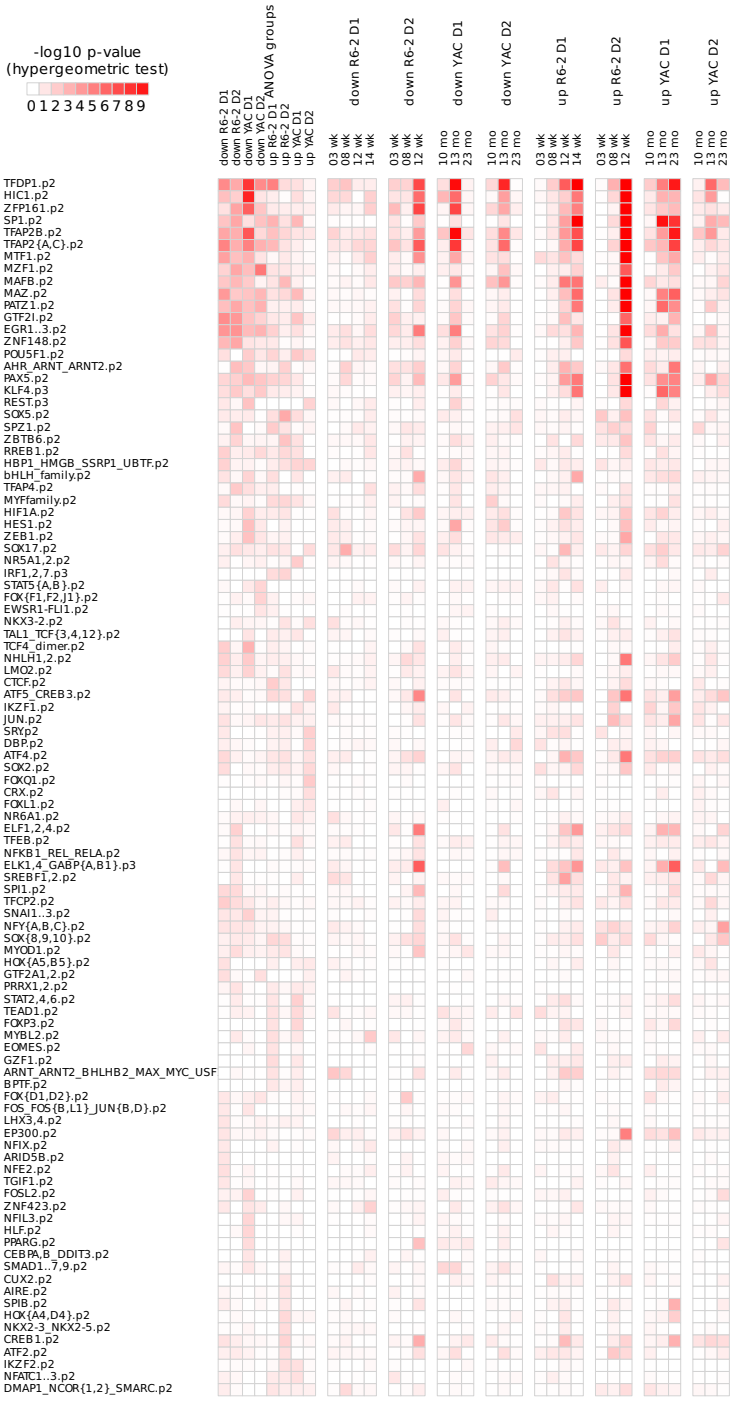


Figure 7.10: Graphical summary showing over-representation of individual *cis*- regulatory motifs in promoter regions of HD-dysregulated genes over all experimental contexts.

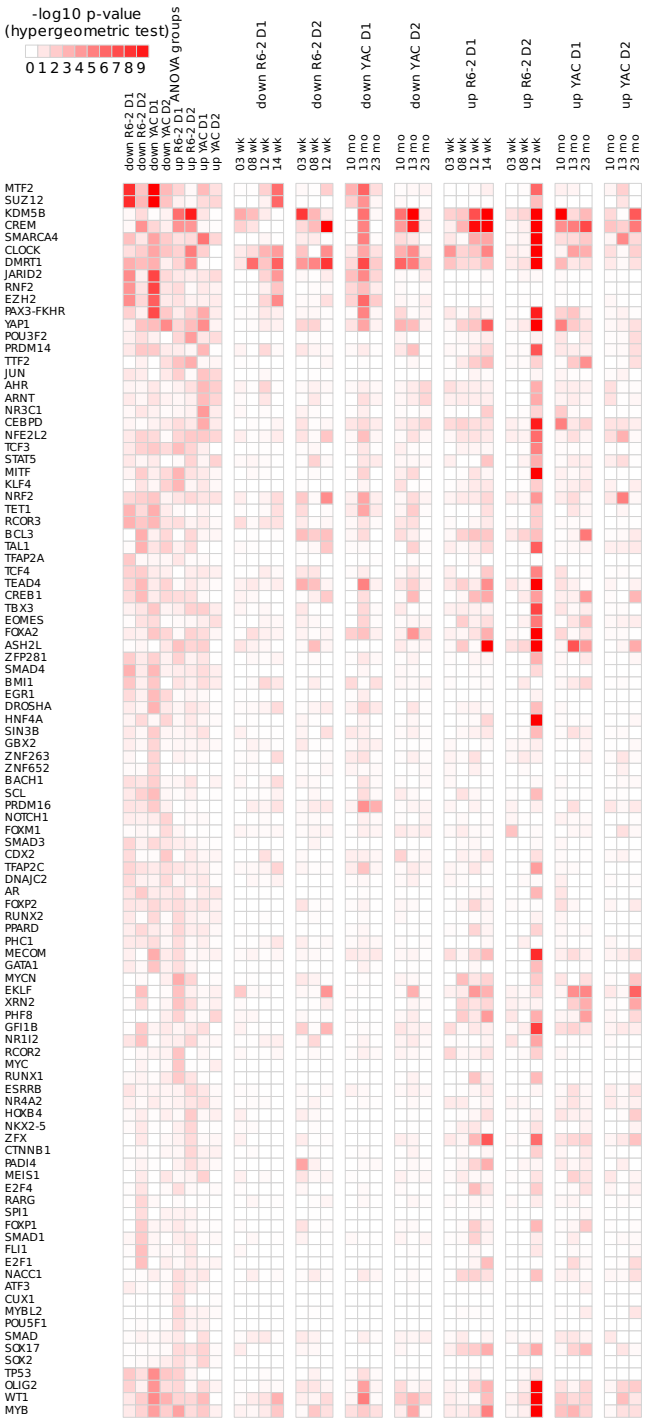


Figure 7.11: Graphical summary showing over-representation of mouse regulatory targets from the ChEA database, over all experimental contexts.

given the multiple-testing issues inherent in transcriptional profiling.

The microarray technology used to measure transcription in this study is well established and cost-effective. However, the arrays here were not designed to detect splicing variation. We observed many instances in which alternative probe sets for genes reported inconsistent changes. It will be interesting to apply long-read mRNA-seq to more thoroughly characterize splicing in these models and to validate expression changes with an independent method.

7.5.2 Cell-type specificity of HD transcriptional dysregulation

The expression changes observed across both cell types suggest that the underlying selective vulnerability of the striatum to mHTT is not D1 or D2 specific, but likely a more general characteristic of MSNs.

7.5.3 Potential Role of PRC2

The over-representation analyses in both cell types (Figure 7.11) again suggests the involvement of the PRC2 complex in transcriptional changes caused by mHTT. In the future, profiling epigenetic states and binding in the specific cell types of interest, perhaps using epitope-tagged chromatin regulators and transcription factors [16], may help to deconstruct the regulatory networks that are affected by mHTT.

Chapter 8

Analysis of Expression in the CHDI HD Allelic Series

8.1 Introduction

The age of onset of HD is dependent on CAG repeat length [33], and understanding the mechanistic basis for this genetic observation remains an open question. One approach to study CAG dependence has been the construction of a series of knock-in mice heterozygous for mHTT alleles with varying repeat lengths [139]. Recently, a consortium organized by the CHDI Foundation (`http://www.chdifoundatation.org`) conducted a systematic study to measure mRNA and miRNA expression in many tissues over this allelic series and over time.

We performed some preliminary and non-exhaustive analysis of this expansive dataset to investigate three focused questions, as well as to offer a vignette describing an application of BOMBASTIC. First, we sought to verify whether our observations about the potential role of PRC2 and H3K27me3 in transcriptional changes in HD (Chapter 6 and 7) could be replicated in this newer, bigger, and independent study. Second, using the allelic series, we could evaluate which genes and potential dysregulatory mechanisms might be dependent on mHTT repeat length. Finally, since the CHDI experiment measured expression across a wide range of tissues, including many of non-neuronal origin, we could assess how patterns of differential expression and regulators vary over tissue types and time, which could suggest additional hypotheses about the context-specificity of mHTT effects. To explore some of the many different possible comparisons among the different timepoints and tissues neeed to address these questions, we made use of the BOMBASTIC methodology and software described in Chapter 3.

8.2 Experimental Design and Data

mRNA-seq expression data, consisting of raw read counts and FPKM values for each sample, was obtained from the CHDI HDinHD.org Data Portal (`http://www.hdinhd.org`). This data had been made openly available to the community under a Creative Commons Attribution 3.0 Unported License (`https://creativecommons.org/licenses/by/3.0/us/`). For three tissues – striatum, cortex, and liver – measurements were available over the full allelic series, at 2, 6, and 10 months. For 11 additional tissues, data was available comparing the Q175 knock-in to the wild-type mice, at 6 months. For almost all comparisons, there were 8 mice tested for each genotype at each time point.

The CHDI study also measured behavioral and motor phenotypes with a battery of automated assays using sensor data and video tracking [10, 100]. For the initial analysis described in this chapter, we did not attempt to make use of this pheotypic data.

8.3 Methods

8.3.1 Differential expression analysis

DESeq 1.20 [2] was used to assess differential expression for each mHTT heterozygous knock-in vs. WT contrast, independently, for each tissue and time point. DESeq operates on transcript read counts, and performs a binomial test for differential expression after estimating dispersions for each gene, adjusting for the library size of each experiment and the expression level of the gene.

8.4 Results

Differential Expression across Tissues

We used BOMBASTIC to organize, cluster, and visualize differential expression across the tissues using statistics computed with DESeq. Genes were clustered in blocks by tissue type. We explored a variety of clustering approaches, but focus here on the quantized contrasts clustering for its simplicity and ease of interpretation. Figure 8.1 shows representative block clusterings for striatum, cortex, and liver. In the version shown, differential

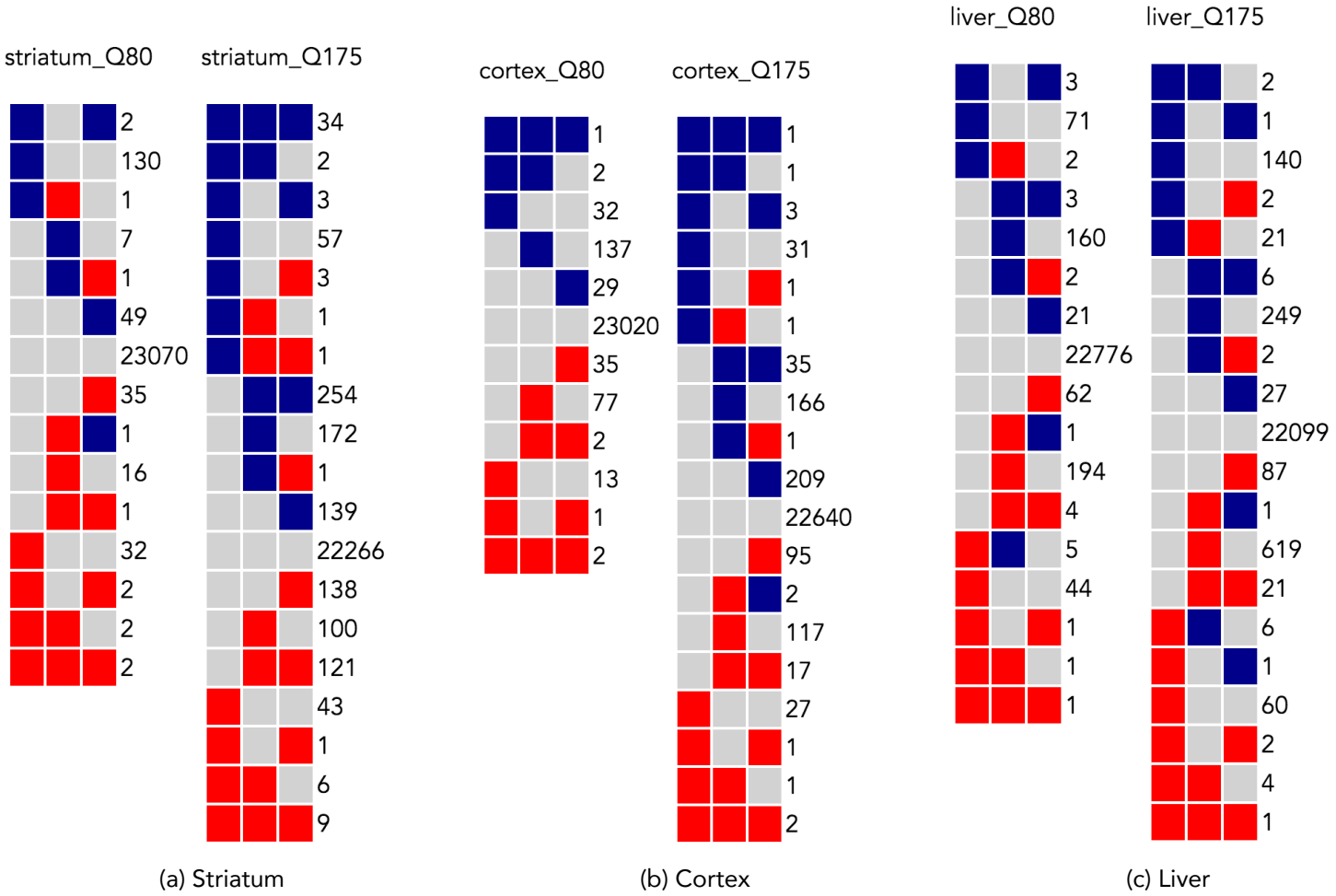
expression is classified as up, down, or unchanged, at a nominal α of 0.05 and log2 fold-change threshold of +/- 0.58. The three columns within each block represent the 2-month, 6-month, and 10-month time points, respectively.

Figure 8.2 shows an alternative set of clusterings using more granular fold change levels, as well as a more stringent statistical threshold (Benjamini-Hochberg adjusted p-value < 0.10).

Expression differences begin to develop as early as 2 months in all three tissues. Many of the genes with the biggest changes are familiar, such as Wt1. Very few differentially expressed genes change direction over time.

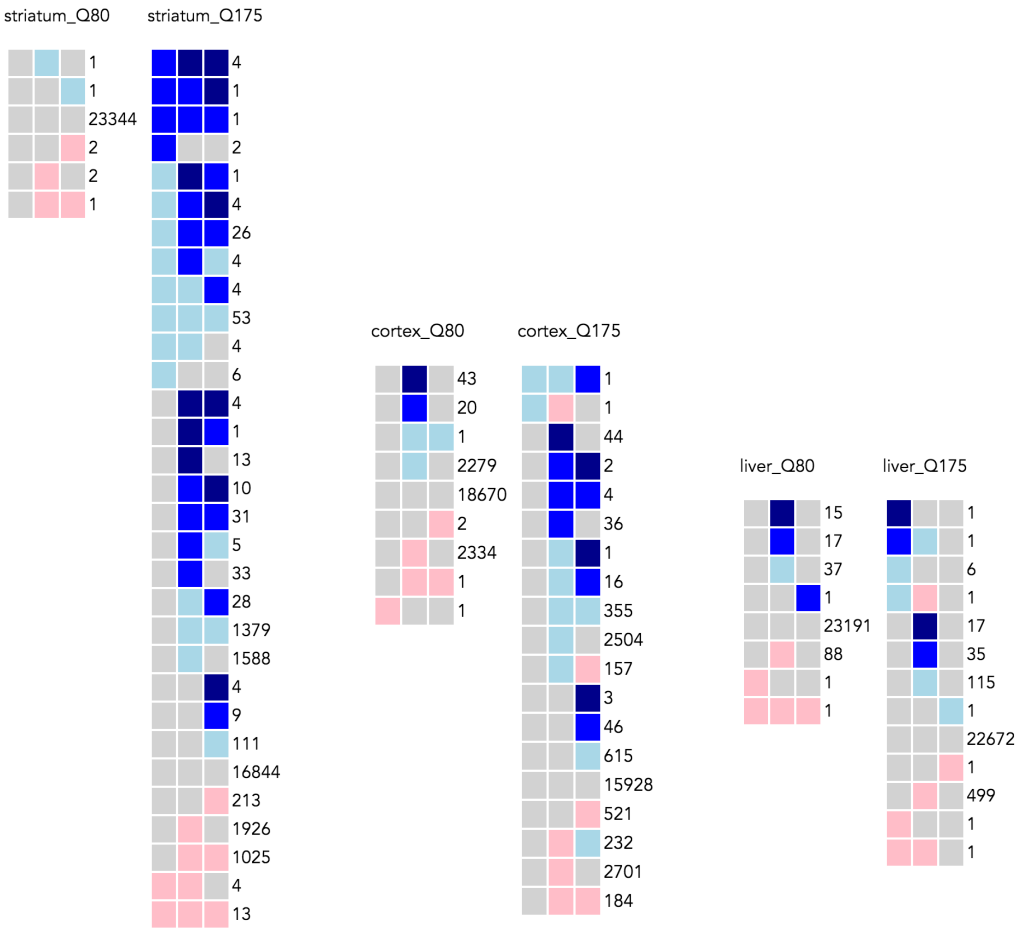
As expected, there tend to be more differentially expressed genes in the striatum than the cortex. We observed a relatively large number of genes that appear to be differentially expressed in the liver, particularly at the 6 month time point (Figure 8.1 and 8.2), a majority of which are up-regulated. That the number of genes changing in the liver might be comparable or greater than the number changing in striatum is somewhat surprising, although hepatic dysfunction and transcriptional dysregulation in liver has previously been observed [24, 55].

We also examined the tissues sampled by CHDI at 6 months. Figure 8.3 shows the distibution of expression changes in each tissue, at two alternative signficance levels. At the more stringent settings, we observe many more changes in the cerebellum than the brainstem, and very few in the corpus callosum. In peripheral tissues, there appear to be significant expression changes in skin and brown adipose tissue, and few or no changes in white adipose or muscle.



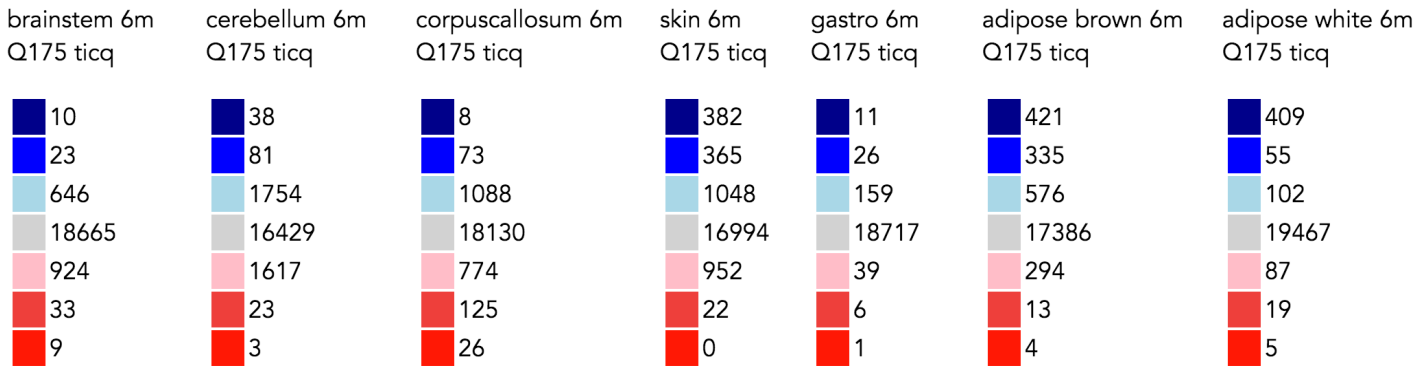
(a) Striatum (b) Cortex (c) Liver

Figure 8.1: Representative TIQCC clusterings showing distribution of genes into patterns of differential expression over time (2m, 6m, 10m) in striatum, cortex, and liver for CHDI Q80 and Q175 mice. Red indicates upregulated in mHTT vs WT; blue is downregulated, (nominal) $\alpha < 0.05$, log2 fold change quantized to one level, +/- 0.58

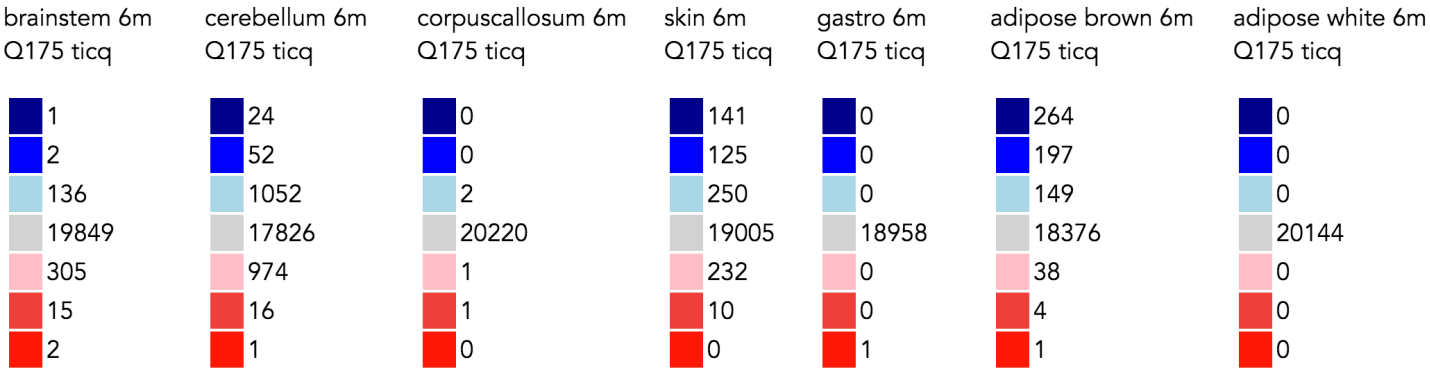


(a) Striatum (b) Cortex (c) Liver

Figure 8.2: Representative TIQCC clusterings showing distribution of genes into patterns of differential expression over time (2m, 6m, 10m) in striatum, cortex, and liver for CHDI Q80 and Q175 mice. Red indicates upregulated in mHTT vs WT; blue is downregulated, (bh adjusted) $\alpha < 0.10$, log2 fold change quantized between the levels $[-\infty, -2, -1, 1, 2, \infty]$ (from blue to red); grey indicates no significant change.



(a) nominal α = 0.05



(b) b-h adjusted α = 0.10

Figure 8.3: Representative single-contrast expression changes across the diverse tissues measured at 6m. Top panel uses a nominal p-value cutoff of 0.05; lower panel uses a Benjamini-Hochberg adjusted p-value of 0.10. Log2 fold changes are quantized between the levels $[-\infty, -2, -1, 1, 2, \infty]$ (from blue to red); grey indicates no significant change.

8.4.1 Comparisons between tissues using BOMBASTIC

We examined selected pairs of tissues and compared clusters of differentially expressed genes using BOMBASTIC. Figure 8.4 shows the the Striatum Q175 over time clustering (of Figure 8.2 (a)) vs. Cortex Q175 over time (Figure 8.2 (b)). There are many intersections between clusters that may be of potential interest and can be examined interactively, but we briefly remark upon a few examples.

One of the few genes downregulated at all time points in both striatum and cortex is Penk, pre-encephalin, whose dysregulation is a well known marker of striatal dysfunction in HD (eg. [102]). From the 10 month time-point, however, a considerable number of genes with well-known striatal changes are also dysregulated in the cortex, including Ddit4l, Plk5, Dusp18, and Rgs4. The consistency in many such expression changes across multiple neuron types suggests that at least some of the mechanisms through which mHTT affects transcription are not MSN-specific.

8.4.2 Verificiation and tissue-dependence of putative PRC2 regulation

We examined over-representation of the ChEA regulators (as described in previous chapters) across the subsets of genes with various patterns of differential expression, focusing on the Q175 groups which had the greatest number of significant expression changes. In striatal samples, there was statistically significant over-representation of Suz12 and Eed targets in the [0, -1, -1] cluster, which contained 1279 genes and the [-1, -1, -1] cluster which had 53 genes (see Fig 8.2(a)). There was also (somewhat less extreme) over-representation of Suz12 targets in the [0, 0, +1] group.

In the cortex, there was similar over-representation for Suz12 in the [0, -1, -1] group. Interestingly, in liver, the most significant enrichment for Suz12 targets was seen among up-regulated genes in the [0, +1, 0] cluster. While this preliminary result demands additional confirmation, it might suggest a common underlying mechanism leading to opposing effects due to differences in the transcriptional regulatory networks and epigenetic states of neurons and hepatocytes.

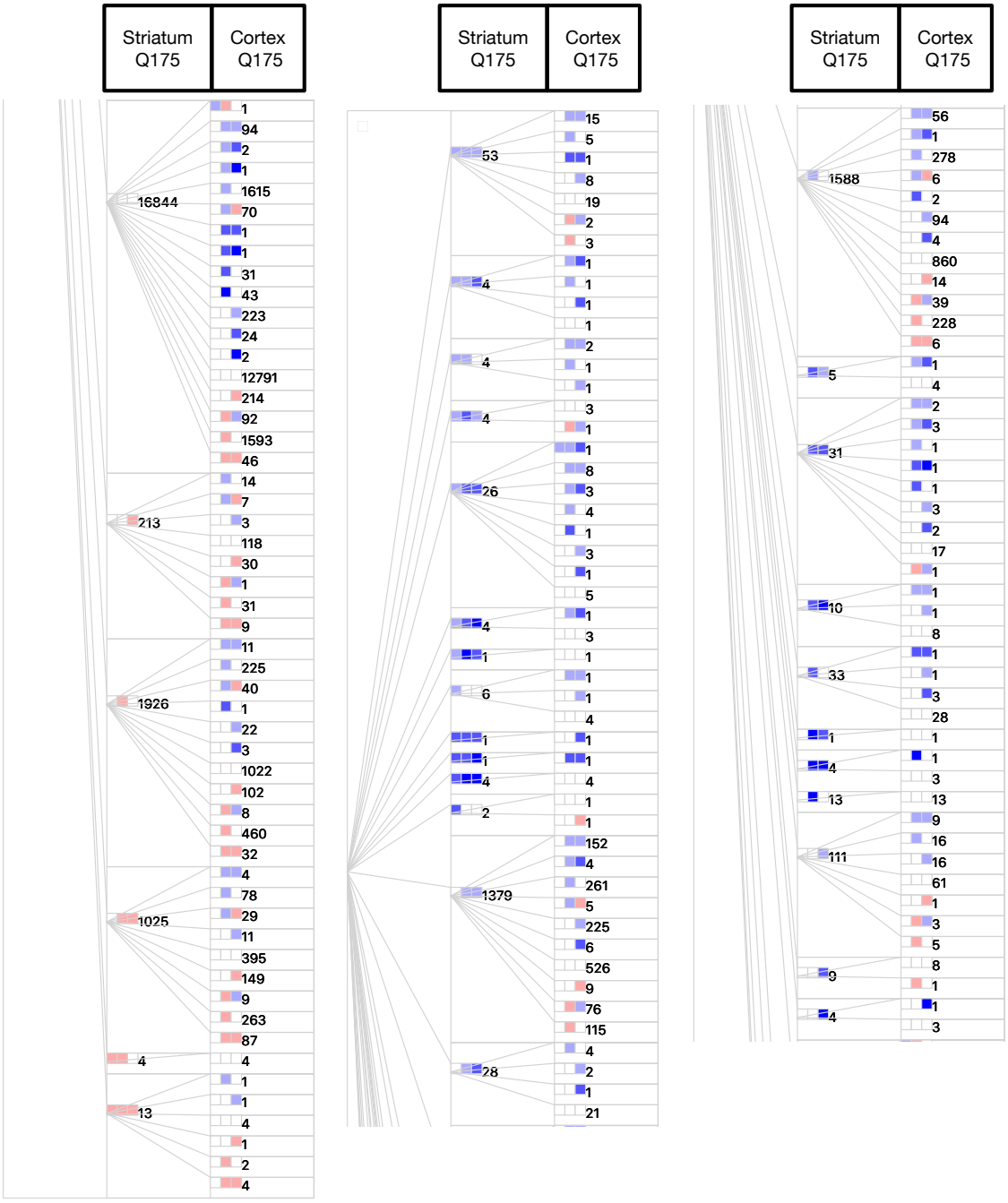


Figure 8.4: Striatum Q175 vs. Cortex Q175 clustering tree. Within each block, genes are clustered by patterns of quantized contrasts. Numbers indicate count of genes at each node having the combination of patterns indicated by each path.

8.5 Discussion

8.5.1 Future work

Better statistical modeling and tests for time course data

The statistical methodology used in this initial analysis of the CHDI dataset was chosen to be as simple as possible. Our analysis was built upon tests done using DESeq, performed independently for each contrast. It would be prudent to verify our results using alternative differential expression tests and software, such as DESeq2 [77] and *limma* [99].

Another opportunity for improvement is to use a methodology that integrates more information over the time-points and tissues studied. Given that we expect that there are common underlying mechanisms of regulation across at least some of the tissue types, one might use a simple hierarchical bayesian model including both tissue-specific and shared factors driving expression, which would allow information to be shared across the contexts to capture the intuition that consistent patterns of dysregulation over multiple tissues is evidence of *bona fide* biological changes. Having such a model is particularly important because the number of replicates performed in these and similar studies is frequently inadequate to provide sufficient power to compensate for multiple-testing adjustments required for genome-wide measurements. We have observed many instances in which small changes at early timepoints or caused by milder mHTT genotypes are not statistically significant, yet re-appear with stronger signals in other contexts. This greatly complicates comparing across contexts to track disease progression and understand early pathophysiological mechanisms.

Improvements to the BOMBASTIC software

In addition to the features proposed in Chapter 3, our experience applying BOMBASTIC to a real data analysis scenario identified two critical areas in the software needing improvement. First, the the current BOMBASTIC interface is optimized for interactive use, and lacks features for easily annotating visualizations and saving results to static figures. Second, analyses of complex datasets such as the one described in this chapter require many choices for methods and parameters, which can generate a combinatorial explo-

sion of alternative result data sets and their dependencies. For example, even before the BOMBASTIC step of intersecting clusters, all of the analyses described can be performed using different statistical thresholds, multiple-testing corrections, and levels for quantization. It is critical to be able to vary and explore such parameters, but doing so manually while properly labeling and accounting for all of the results can be challenging and tedious. This is a not a problem about which we were unaware, but the present BOMBASTIC implementation lacks some essential abstractions needed to make managing this dataflow as easy, verifiable, and efficient as it might be.

Bibliography

[1] Christopher Ahlberg, Christopher Williamson, and Ben Shneiderman. *Dynamic Queries for Information Exploration : An Implementation and Evaluation*. Tech. rep. College Park, MD: Department of Computer Science, Human Computer Interaction Laboratory, University of Maryland, 1993.

[2] Simon Anders and Wolfgang Huber. "Differential expression analysis for sequence count data." *Genome Biology* 11.10 (2010), R106.

[3] David Andrzejewski and David Buttler. "Latent topic feedback for information retrieval". *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11* (2011), p. 600.

[4] Aristotle. *The History of Animals*.

[5] Phil Arnold et al. "MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences." *Bioinformatics* 28.4 (Feb. 2012), pp. 487–94.

[6] Romina Aron Badin et al. "IRC-082451, a novel multitargeting molecule, reduces L-DOPA-induced dyskinesias in MPTP Parkinsonian primates." *PloS ONE* 8.1 (Jan. 2013), e52680.

[7] Macarena S Arrázola, Carmen Silva-Alvarez, and Nibaldo C Inestrosa. "How the Wnt signaling pathway protects from neurodegeneration: the mitochondrial scenario." *Frontiers in Cellular Neuroscience* 9.May (2015), p. 166.

[8] Incarnation Aubert et al. "Increased D1 dopamine receptor signaling in levodopa-induced dyskinesia". *Annals of Neurology* 57.1 (2005), pp. 17–26.

[9] Swneke D. Bailey et al. "ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters". *Nature Communications* 2 (2015), p. 6186.

[10] F Balci et al. "High Throughput Automated Phenotyping of Two Genetic Mouse Models of Huntington's Disease". *PLOS Currents Huntington Disease* (2013), pp. 1–31.

[11] Gillian P. Bates et al. "Huntington Disease". *Nature Reviews Disease Primers* April (2015), p. 15005.

[12] Kristina Becanovic et al. "Transcriptional changes in Huntington disease identified using genome-wide expression profiling and cross-platform analysis." *Human Molecular Genetics* 19.8 (Apr. 2010), pp. 1438–52.

[13] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". *Journal of the Royal Statistical Society. Series B (Methodological)* (1995). arXiv: 95/57289 [0035-9246].

[14] Erwan Bézard et al. "Attenuation of levodopa-induced dyskinesia by normalizing dopamine D3 receptor function." *Nature Medicine* 9.6 (2003), pp. 762–767.

[15] Marta Biagioli et al. "Htt CAG repeat expansion confers pleiotropic gains of mutant huntingtin function in chromatin regulation". *Human Molecular Genetics* (2015), pp. 1–16.

[16] Stefan Bonn et al. "Cell type-specific chromatin immunoprecipitation from multicellular complex samples using BiTS-ChIP." *Nature Protocols* 7.5 (2012), pp. 978–94.

[17] Ippolita Cantuti-Castelvetri et al. "Levodopa-induced dyskinesia is associated with increased thyrotropin releasing hormone in the dorsal striatum of hemi-parkinsonian rats." *PloS ONE* 5.11 (2010), e13861.

[18] M Angela Cenci and Christine Konradi. "Maladaptive striatal plasticity in L-DOPA-induced dyskinesia." In: *Progress in Brain Research*. Vol. 183. 10. Elsevier B.V., Jan. 2010. Chap. 11, pp. 209–33.

[19] M. Angela Cenci, C S Lee, and A Björklund. "L-DOPA-induced dyskinesia in the rat is associated with striatal overexpression of prodynorphin- and glutamic acid decarboxylase mRNA." *The European Journal of Neuroscience* 10.8 (Aug. 1998), pp. 2694–706.

[20] Jang-Ho J Cha. "Transcriptional signatures in Huntington's disease." *Progress in Neurobiology* 83.4 (Nov. 2007), pp. 228–48.

[21] P E Chabrier and M Auguet. "Pharmacological properties of BN82451: a novel multitargeting neuroprotective agent". *CNS Drug Reviews* 13.3 (2007), pp. 317–332.

[22] CHDI Foundation. *HDinHD*. 2015.

[23] Yizong Cheng and George M Church. "Biclustering of expression data." *Proceedings of the International Conference on Intelligent Systems for Molecular Biology* 8 (2000), pp. 93–103.

[24] Ming Chang Chiang, Yijuang Chern, and Chiun Gung Juo. "The dysfunction of hepatic transcriptional factors in mice with Huntington's Disease". *Biochimica et Biophysica Acta - Molecular Basis of Disease* 1812.9 (2011), pp. 1111–1120.

[25] E F Codd. "A relational model of data for large shared data banks". *Information Retrieval* 15.3 (1970), pp. 162–6.

[26] Barbara S. Connolly and Anthony E. Lang. "Pharmacological Treatment of Parkinson Disease". *Journal of the American Medical Association* 311.16 (2014), p. 1670.

[27] Wisam Dakka and Panagiotis G Ipeirotis. "Automatic Extraction of Useful Facet Hierarchies from Text Databases". In: *Proceedings of the 24th International Conference on Data Engineering, ICDE 2008, April 7-12, 2008, Cancún, México*. Ed. by Gustavo Alonso, José Blakeley, and Arbee Chen. 2008, pp. 466–475.

[28] Mickael Decressac et al. "NURR1 in Parkinson disease–from pathogenesis to therapeutic potential." *Nature Reviews Neurology* 9.11 (2013), pp. 629–36.

[29] S Déjean et al. "Clustering time-series gene expression data using smoothing spline derivatives." *EURASIP Journal on Bioinformatics & Systems Biology* 2007 (Jan. 2007), p. 70561.

[30] Joseph P Doyle et al. "Application of a translational profiling approach for the comparative analysis of CNS cell types." *Cell* 135.4 (Nov. 2008), pp. 749–762.

[31] Anthone W Dunah et al. "Sp1 and TAFII130 transcriptional activity disrupted in early Huntington's disease." *Science* 296.5576 (June 2002), pp. 2238–2243.

[32] G Dunn and Brian Sidney Everitt. *An Introduction to Numerical Taxonomy*. Cambridge: Cambridge University Press, 1982.

[33] M Duyao et al. "Trinucleotide repeat length instability and age of onset in Huntington's disease." *Nature Genetics* 4 (1993), pp. 387–392.

[34] M B Eisen et al. "Cluster analysis and display of genome-wide expression patterns." *Proceedings of the National Academy of Sciences of the United States of America* 95.25 (Dec. 1998), pp. 14863–8.

[35] Jason Ernst and Ziv Bar Joseph. "STEM: a tool for the analysis of short time series gene expression data". *BMC Bioinformatics* 7.1 (2006).

[36] Jason Ernst, Gerard J Nau, and Ziv Bar-Joseph. "Clustering short time series gene expression data." *Bioinformatics* 21 Suppl 1 (June 2005), pp. i159–68.

[37] Robert J Fenster. "CELL-TYPE SPECIFIC TRANSLATIONAL PROFILING IN HUNTINGTON'S DISEASE MOUSE MODELS". PhD thesis. Rockefeller University, 2011.

[38] Chris Fraley and Adrian E Raftery. "Model-Based Clustering, Discriminant Analysis, and Density Estimation". *Journal of the American Statistical Association* 97.458 (2002), pp. 611–631.

[39] Veronica Francardo et al. "Impact of the lesion procedure on the profiles of motor impairment and molecular responsiveness to L-DOPA in the 6-hydroxydopamine mouse model of Parkinson's disease". *Neurobiology of Disease* 42.3 (2011), pp. 327–340.

[40] Laurent Gautier et al. "Affy - Analysis of Affymetrix GeneChip data at the probe level". *Bioinformatics* 20.3 (2004), pp. 307–315.

[41] Robert C Gentleman, Vincent J Carey, Douglas M Bates, et al. "Bioconductor: Open software development for computational biology and bioinformatics". *Genome Biology* 5.10 (2004), R80.

[42] Samuel Gratzl et al. "Domino: Extracting, Comparing, and Manipulating Subsets across Multiple Tabular Datasets". *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 2023–2032.

[43] E. Grünblatt et al. "Transcriptional alterations under continuous or pulsatile dopaminergic treatment in dyskinetic rats". *Journal of Neural Transmission* 118.12 (2011), pp. 1717–1725.

[44] James F. Gusella, Marcy E. MacDonald, and Jong-Min Lee. "Genetic modifiers of Huntington's disease". *Movement Disorders* 29.11 (Sept. 2014), pp. 1359–1365.

[45] Isabelle Guyon and André Elisseeff. "An Introduction to Variable and Feature Selection". *Journal of Machine Learning Research* 3 (2003), pp. 1157–1182.

[46] Susan L. Havre et al. "Diverse information integration and visualization". In: *Visualization and Data Analysis 2006 SPIE-IS&T Electronic Imaging*. Vol. 6060. 2006, pp. 60600M–60600M–11.

[47] Adrian Heilbut, Myriam Heiman, and Eric Kolaczyk. "Declarative Interactive Visual Analytics for Clustering Structured Data". In: *Data Science Learning and Applications to Biomedical and Health Sciences*. New York, NY: New York Academy of Sciences, 2016, pp. 100–104.

[48] Adrian Heilbut et al. "BOMBASTIC: Block-Organized Model-BAsed Tree-Indexed Clustering". In: *VIZBI 2013*. Cambridge, MA, 2013.

[49] Myriam Heiman et al. "A translational profiling approach for the molecular characterization of CNS cell types." *Cell* 135.4 (Nov. 2008), pp. 738–48.

[50] Myriam Heiman et al. "Cell type-specific mRNA purification by translating ribosome affinity purification (TRAP)." *Nature Protocols* 9.6 (2014), pp. 1282–91.

[51] Myriam Heiman et al. "Molecular adaptations of striatal spiny projection neurons during levodopa-induced dyskinesia." *Proceedings of the National Academy of Sciences of the United States of America* 111.12 (Mar. 2014), pp. 4578–83.

[52] T Hey, S Tansley, and K Tolle, eds. *The Fourth Paradigm: Data-Intensive Scientific Research*. Redmond, Washington: Microsoft Research, 2009.

[53] Angela Hodges et al. "Regional and cellular gene expression changes in human Huntington's disease brain." *Human Molecular Genetics* 15.6 (Mar. 2006), pp. 965–77.

[54] J G Hodgson et al. "Human huntingtin derived from YAC transgenes compensates for loss of murine huntingtin by rescue of the embryonic lethal phenotype." *Human Molecular Genetics* 5.12 (1996), pp. 1875–1885.

[55] Rainer Hoffmann et al. "Progressive hepatic mitochondrial dysfunction in premanifest Huntington's disease." *Movement Disorders* 29.6 (2014), pp. 831–4.

[56] Andrew G Hoss et al. "MicroRNAs Located in the Hox Gene Clusters Are Implicated in Huntington's Disease Pathogenesis." *PLoS Genetics* 10.2 (Feb. 2014), e1004188.

[57] CC Huang et al. "Amyloid formation by mutant huntingtin: threshold, progressivity and recruitment of normal polyglutamine proteins". *Somatic Cell and Molecular Genetics* 24.4 (1998), pp. 217–33.

[58] Julien Jacques and Cristian Preda. *Functional data clustering : a survey*. Tech. rep. January. INRIA, 2013, Research Report 8198 –Project Team MODAL.

[59] A.K Jain, M.N. Murty, and P.J Flynn. "Data Clustering : A Review". *ACM Computing Surveys* 1999.3 (2000), pp. 264–323.

[60] Peter Jenner. "Molecular mechanisms of L-DOPA-induced dyskinesia." *Nature Reviews Neuroscience* 9.9 (Sept. 2008), pp. 665–77.

[61] Amanda Jones and Hengbin Wang. "Polycomb repressive complex 2 in embryonic stem cells: an overview." *Protein & Cell* 1.12 (2010), pp. 1056–62.

[62] Rebecka Jörnsten and Sündüz Keleş. "Mixture models with multiple levels, with application to the analysis of multifactor gene expression data." *Biostatistics (Oxford, England)* 9.3 (July 2008), pp. 540–54.

[63] Linda S Kaltenbach et al. "Huntingtin Interacting Proteins Are Genetic Modifiers of Neurodegeneration". *PLoS Genetics* 3.5 (2007), pp. 689–708.

[64] Eric R. Kandel et al. *Principles of Neural Science*. 5th ed. McGraw Hill, 2013.

[65] Minoru Kanehisa et al. "Data, information, knowledge and principle: back to metabolism in KEGG." *Nucleic Acids Research* 42.Database issue (2014), pp. D199–205.

[66] Takeya Kasukawa et al. "Quantitative expression profile of distinct functional regions in the adult mouse brain." *PloS ONE* 6.8 (Jan. 2011), e23228.

[67] Christine Konradi et al. "Transcriptome analysis in a rat model of L-DOPA-induced dyskinesia." *Neurobiology of Disease* 17.2 (Nov. 2004), pp. 219–36.

[68] Robert Kosara, Fabian Bendix, and Helwig Hauser. "Parallel sets: interactive exploration and visual analysis of categorical data." *IEEE Transactions on Visualization and Computer Graphics* 12.4 (2006), pp. 558–68.

[69] Alexandre Kuhn et al. "Mutant huntingtin's effects on striatal gene expression in mice recapitulate changes observed in human Huntington's disease brain and do not differ with mutant huntingtin length or wild-type huntingtin dosage." *Human Molecular Genetics* 16.15 (Aug. 2007), pp. 1845–61.

[70] John Labbadia and Richard I. Morimoto. "Huntington's disease: Underlying molecular mechanisms and emerging concepts". *Trends in Biochemical Sciences* 38.8 (Aug. 2013), pp. 378–385.

[71] Alexander Lachmann et al. "ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments." *Bioinformatics* 26.19 (Oct. 2010), pp. 2438–44.

[72] Jae-Ho Lee et al. "CREBZF, a novel Smad8-binding protein". *Molecular and Cellular Biochemistry* 368.1-2 (2012), pp. 147–153.

[73] S J Lee et al. "E3 ligase activity of RING finger proteins that interact with Hip-2, a human ubiquitin-conjugating enzyme." *FEBS Letters* 503.1 (2001), pp. 61–64.

[74] Alexander. Lex et al. "StratomeX: Visual Analysis of Large-Scale Heterogeneous Genomics Data for Cancer Subtype Characterization". *Computer Graphics Forum* 31.3pt3 (June 2012), pp. 1175–1184.

[75] Matthew D Li et al. "Integrated multi-cohort transcriptional meta-analysis of neurodegenerative diseases". *Acta Neuropathologica Communications* 2.1 (2014), p. 93.

[76] Shi-hua Li et al. "Interaction of Huntington Disease Protein with Transcriptional Activator Sp1". *Molecular and Cellular Biology* 22.5 (2002), pp. 1277–1287.

[77] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". *Genome Biology* 15.12 (2014), p. 550.

[78] Mark A. Lovell et al. "Wilms' tumor suppressor (WT1) is a mediator of neuronal degeneration associated with the pathogenesis of Alzheimer's disease". *Brain Research* 983.1-2 (2003), pp. 84–96.

[79] Tao Lu et al. "REST and stress resistance in ageing and Alzheimer disease". *Nature* 7493 (2014), pp. 448–54.

[80] M. Lundblad et al. "A model of L-DOPA-induced dyskinesia in 6-hydroxydopamine lesioned mice: Relation to motor and cellular parameters of nigrostriatal function". *Neurobiology of Disease* 16.1 (2004), pp. 110–123.

[81] Ping Ma et al. "A data-driven clustering method for time course gene expression data." *Nucleic Acids Research* 34.4 (Jan. 2006), pp. 1261–9.

[82] Sara C Madeira and Arlindo L Oliveira. "Biclustering algorithms for biological data analysis: a survey." *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* 1.1 (2004), pp. 24–45.

[83] L Mangiarini et al. "Exon 1 of the HD gene with an expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice." *Cell* 87.3 (Nov. 1996), pp. 493–506.

[84] S Mao, G a Neale, and R M Goorha. "T-cell oncogene rhombotin-2 interacts with retinoblastoma-binding protein 2." *Oncogene* 14 (1997), pp. 1531–1539.

[85] Raphaël Margueron and Danny Reinberg. "The Polycomb complex PRC2 and its mark in life." *Nature* 469.7330 (Jan. 2011), pp. 343–9.

[86] José Fernando Maya-Vetencourt. "Activity-dependent NPAS4 expression and the regulation of gene programs underlying plasticity in the central nervous system." *Neural Plasticity* 2013 (2013), p. 683909.

[87] Karen N McFarland et al. "MeCP2: a novel Huntingtin interactor." *Human Molecular Genetics* 23.4 (Feb. 2014), pp. 1036–44.

[88] Liliana B Menalled. "Knock-in mouse models of Huntington's disease." *NeuroRx : The Journal of the American Society for Experimental NeuroTherapeutics* 2.3 (2005), pp. 465–470.

[89] Liliana Menalled et al. *A Field Guide to Working with Mouse Models of Huntington's Disease*. Tech. rep. CHDI foundation, 2014.

[90] Tarjei S Mikkelsen et al. "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells." *Nature* 448.7153 (Aug. 2007), pp. 553–60.

[91] Lara Moumné, Sandrine Betuing, and Jocelyne Caboche. "Multiple Aspects of Gene Dysregulation in Huntington's Disease." *Frontiers in Neurology* 4.October (Jan. 2013), p. 127.

[92] Christopher W Ng et al. "Extensive changes in DNA methylation are associated with expression of mutant huntingtin." *Proceedings of the National Academy of Sciences of the United States of America* 110.6 (Feb. 2013), pp. 2354–9.

[93] Harry T Orr and Huda Y Zoghbi. "Trinucleotide repeat disorders." *Annual Review of Neuroscience* 30 (Jan. 2007), pp. 575–621.

[94] Mikhail Pachkov et al. "SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates." *Nucleic Acids Research* 41.Database issue (Jan. 2013), pp. D214–20.

[95] F Pedregosa et al. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[96] Thomas Perlmann and Asa Wallén-Mackenzie. "Nurr1, an orphan nuclear receptor with essential functions in developing dopamine cells." *Cell and Tissue Research* 318.1 (Oct. 2004), pp. 45–52.

[97] Plato. *Phaedrus*. Cambridge, MA: Harvard University Press, 265e.

[98] R Development Core Team. *R: A language and environment for statistical computing*. Tech. rep. Vienna, Austria.: R Foundation for Statistical Computing, 2008.

[99] Matthew E Ritchie et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies". *Nucleic Acids Research* 43.7 (2015), e47.

[100] Steven L. Roberds et al. "Rapid, computer vision-enabled murine screening system identifies neuropharmacological potential of two new mechanisms". *Frontiers in Neuroscience* 5.SEP (2011), pp. 1–4.

[101] Christopher A Ross. "Polyglutamine pathogenesis: Emergence of unifying mechanisms for Huntington's disease and related disorders". *Neuron* 35.5 (2002), pp. 819–822.

[102] Heike Runne et al. "Dysregulation of gene expression in primary neuron models of Huntington's disease shows that polyglutamine-related effects on the striatal transcriptome may not be dependent on brain circuitry." *The Journal of Neuroscience* 28.39 (Sept. 2008), pp. 9723–31.

[103] Alexander Russakovskii. *Mathematical Foundations of the Multidimensional Database Models*. Tech. rep. Hyperion Solutions Corporation, 1999.

[104] Giovanni Maria Sacco and Yannis Tzitzikas, eds. *Dynamic Taxonomies and Faceted Search*. Berlin: Springer, 2009.

[105] Kathleen M Salerno et al. "Sox11 modulates brain-derived neurotrophic factor expression in an exon promoter-specific manner." *Journal of Neuroscience Research* 90.5 (2012), pp. 1011–9.

[106] Emanuela Santini, Emmanuel Valjent, and Gilberto Fisone. "Parkinson's disease: levodopa-induced dyskinesia and signal transduction." *The FEBS Journal* 275.7 (Apr. 2008), pp. 1392–9.

[107] Eva Schaeffer, Andrea Pilotto, and Daniela Berg. "Pharmacological strategies for the management of levodopa-induced dyskinesia in patients with Parkinson's disease." *CNS Drugs* 28.12 (2014), pp. 1155–84.

[108] Skipper Seabold and Josef Perktold. "Statsmodels: econometric and statistical modeling with python". *Proceedings of the 9th Python in Science Conference* Scipy (2010), pp. 57–61.

[109] Ihn Sik Seong et al. "Huntingtin facilitates polycomb repressive complex 2." *Human Molecular Genetics* 19.4 (Feb. 2010), pp. 573–83.

[110] P F Shelbourne et al. "A Huntington's disease CAG expansion at the murine Hdh locus is unstable and associated with behavioural abnormalities in mice." *Human Molecular Genetics* 8.5 (1999), pp. 763–774.

[111] Dyna I. Shirasaki et al. "Network organization of the huntingtin proteomic interactome in mammalian brain." *Neuron* 75.1 (July 2012), pp. 41–57.

[112] J. Shrager. "The fiction of function". *Bioinformatics* 19.15 (Oct. 2003), pp. 1934–1936.

[113] Nicola Simola, Micaela Morelli, and Anna R. Carta. "The 6-hydroxydopamine model of Parkinson's disease". *Neurotoxicity Research* 11.3-4 (2007), pp. 151–167.

[114] Simonetta Sipione et al. "Early transcriptional profiles in huntingtin-inducible striatal cells by microarray analyses." *Human Molecular Genetics* 11.17 (Aug. 2002), pp. 1953–65.

[115] Elizabeth J. Slow et al. "Selective striatal neuronal loss in a YAC128 mouse model of Huntington disease". *Human Molecular Genetics* 12.13 (July 2003), pp. 1555–1567.

[116] Gordon K Smyth. "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." *Statistical Applications in Genetics and Molecular Biology* 3.1 (Jan. 2004), Article3.

[117] Peter H. A Sneath. "Classification of Microorganisms". In: *Essays in Microbiology*. Ed. by J R Norris and M H Richmond. London: Wiley, 1978.

[118] Robert R. Sokal and Peter H. A Sneath. *Principles of Numerical Taxonomy*. 2nd. San Francisco: Freeman and Company, 1973.

[119] Boris S Spektor et al. "Differential D1 and D2 receptor-mediated effects on immediate early gene induction in a transgenic mouse model of Huntington's disease". *Molecular Brain Research* 102 (2002), pp. 118–128.

[120] Brigitte Spinnewyn et al. "BN82451 attenuates L-dopa-induced dyskinesia in 6-OHDA-lesioned rat model of Parkinson's disease." *Neuropharmacology* 60.4 (2011), pp. 692–700.

[121] J S Steffan et al. "The Huntington's disease protein interacts with p53 and CREB-binding protein and represses transcription." *Proceedings of the National Academy of Sciences of the United States of America* 97.12 (June 2000), pp. 6763–8.

[122] C. Stolte, D. Tang, and P. Hanrahan. "Polaris: a system for query, analysis, and visualization of multidimensional relational databases". *IEEE Transactions on Visualization and Computer Graphics* 8.1 (2002), pp. 52–65.

[123] Christopher Stolte. "Query, Analysis, and Visualization of Multidimensional Databases". PhD thesis. Stanford University, 2003.

[124] Martin Strauch et al. "A Two-Step Clustering for 3-D Gene Expression Data Reveals the Main Features of the Arabidopsis Stress Response". *Journal of Integrative Bioinformatics* 4.1 (2007), p. 54.

[125] Alfred X. Sun, Gerald R. Crabtree, and Andrew S. Yoo. "MicroRNAs: Regulators of neuronal fate". *Current Opinion in Cell Biology* 25.2 (2013), pp. 215–221.

[126] Susan M Sunkin et al. "Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system." *Nucleic Acids Research* 41.Database issue (Jan. 2013), pp. D996–D1008.

[127] Jochen Supper et al. "EDISA: extracting biclusters from multiple time-series of gene expression profiles". *BMC Bioinformatics* 8 (Jan. 2007), p. 334.

[128] Francesca Telese et al. "Seq-ing Insights into the Epigenetics of Neuronal Gene Regulation". *Neuron* 77.4 (2013), pp. 606–623.

[129] The ENCODE Project Consortium et al. "An integrated encyclopedia of DNA elements in the human genome." *Nature* 489.7414 (2012), pp. 57–74.

[130] The Huntington's Disease Collaborative Research Group. "A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group." *Cell* 72.6 (1993), pp. 971–983.

[131] Elizabeth A Thomas. "Striatal Specificity of Gene Expression Dysregulation in Huntington's Disease". *Journal of Neuroscience Research* 1164.April (2006), pp. 1151–1164.

[132] Cendrine Tourette et al. "A large scale huntingtin protein interaction network implicates RHO GTPase signaling pathways in huntington disease". *Journal of Biological Chemistry* 289.10 (2014), pp. 6709–6726.

[133] Cendrine Tourette et al. "The Wnt receptor Ryk reduces neuronal and cell survival capacity by repressing FOXO activity during the early phases of mutant huntingtin pathogenicity." *PLoS Biology* 12.6 (June 2014), e1001895.

[134] Daniel Tunkelang. "Faceted search". In: *Synthesis Lectures on Information Concepts, Retrieval, and Services # 5*. Ed. by Gary Marchionini. Morgan & Claypool, 2009.

[135] L. M. Valor et al. "Genomic Landscape of Transcriptional and Epigenetic Dysregulation in Early Onset Polyglutamine Disease". *Journal of Neuroscience* 33.25 (June 2013), pp. 10471–10482.

[136] Malini Vashishtha et al. "Targeting H3K4 trimethylation in Huntington disease." *Proceedings of the National Academy of Sciences of the United States of America* 110.32 (Aug. 2013), E3027–36.

[137] Jenny E. Westin et al. "Spatiotemporal Pattern of Striatal ERK1/2 Phosphorylation in a Rat Model of L-DOPA-Induced Dyskinesia and the Role of Dopamine D1 Receptors". *Biological Psychiatry* 9.2 (2010), pp. 1–14. arXiv: NIHMS150003.

[138] Vanessa C. Wheeler et al. "Length-dependent gametic CAG repeat instability in the Huntington's disease knock-in mouse". *Human Molecular Genetics* 8.1 (1999), pp. 115–122.

[139] Vanessa C. Wheeler et al. "Long glutamine tracts cause nuclear localization of a novel form of huntingtin in medium spiny striatal neurons in HdhQ92 and HdhQ111 knock-in mice." *Human Molecular Genetics* 9.4 (2000), pp. 503–513.

[140] Hadley Wickham. "A Layered Grammar of Graphics". *Journal of Computational and Graphical Statistics* 19.1 (Jan. 2010), pp. 3–28.

[141] Leland Wilkinson. *The Grammar of Graphics*. 2nd. New York, NY: Springer, 2005.

[142] Rui Zhang et al. "Zhangfei/CREB-ZF - A Potential Regulator of the Unfolded Protein Response". *PLoS ONE* 8.10 (2013), pp. 1–14.

[143] Lizhuang Zhao and Mohammed J Zaki. "triCluster : An Effective Algorithm for Mining Coherent Clusters in 3D Microarray Data". In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*. Ed. by J Widom, F Ozcan, and R Chirkova. 2005, pp. 694–705.

[144] Jianping Zhou et al. "Visually comparing multiple partitions of data with applications to clustering". *Proc. of SPIE-IS&T Electronic Imaging Visualization & Data Analysis* 7243 (2009), 72430J–1–12.

[145] Chiara Zuccato et al. "Huntingtin interacts with REST/NRSF to modulate the transcription of NRSE-controlled neuronal genes." *Nature Genetics* 35.1 (Sept. 2003), pp. 76–83.

Vita

Education

2010 – present	**Ph.D candidate, Bioinformatics**	
	Boston University	
	(visiting student at Broad Institute)	
	Advisors: Dr. Eric Kolaczyk (BU, Mathematics and Statistics)	
	Dr. Myriam Heiman (MIT, Brain and Cognitive Science)	
2008 – 2010	**M.S., Bioinformatics**	
	Boston University	
1996 – 2001	**B.Sc. (Hon) Computer Science (Major)**	
	and Neuroscience (Specialist)	
	University College, University of Toronto	

Professional Experience

Jan 2010– Aug 2011	**Bioinformatics Specialist**	Center for Human Genetic Research
	Gusella Laboratory	Massachusetts General Hospital
Jan 2005– Aug 2008	**Associate Scientist, Computational Biology**	CombinatoRx Inc.
		Cambridge, MA
Sept 2002–Dec 2004	**Research Assistant**	Department of Biochemistry
		University of Toronto
Jan 2000–Aug 2002	**Bioinformatics Software Developer**	MDS Proteomics Inc.
		Toronto, ON

Teaching Experience

Fall 2014	Co-Instructor	Bioinformatics
	Co-taught BF527, Bioinformatics Applications	Boston University
Spring 2014	Adjunct Instructor	Dept of Mathematics and Statistics
	Co-taught MA213, Introduction to Statistics	Boston University
Spring 2014	Adjunct Instructor	Lesley University
	CMATH2142, Introduction to Statistics	
Fall 2013	Teaching Fellow	Boston University
	BI211, Introduction to Physiology	

Spring 2011	Teaching Fellow	Boston University
	CS103, Introduction to Internet Technologies and Web Programming	
Fall 2010	NSF GK12 Fellow, AP Biology	John D. O'Bryant High School
		Boston, MA
Fall 2009	Teaching Fellow	Boston University
	CS111, Introduction to Computer Science	
Spring 2009	Teaching Fellow	Boston University
	BI315, Introdution to Physiology	

Publications

Heiman M, **Heilbut A**, Francardo V, Kulicke R, Fenster RJ, Kolaczyk ED, Mesirov JP, Surmeier DJ, Cenci MA, Greengard P

Molecular adaptations of striatal spiny projection neurons during levodopa-induced dyskinesia

Proceedings of the National Academy of Sciences 2014 111 (12) 2578-4583.

Talkowski ME, Rosenfeld JA, Blumenthal I, Pillalamarri V, Chiang C, **Heilbut A**, Ernst C, Hanscom C, Rossin E, Lindgren AM, Pereira S, Ruderfer D, Kirby A, Ripke S, Harris DJ, Lee JH, Ha K, Kim HG, Solomon BD, Gropman AL, Lucente D, Sims K, Ohsumi TK, Borowsky ML, Loranger S, Quade B, Lage K, Miles J, Wu BL, Shen Y, Neale B, Shaffer LG, Daly MJ, Morton CC, Gusella JF

Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries

Cell 2012 Apr 27;149(3):525-37.

Chiang C, Jacobsen JC, Ernst C, Hanscom C, **Heilbut A**, Blumenthal I, Mills RE, Kirby A, Lindgren AM, Rudiger SR, McLaughlan CJ, Bawden CS, Reid SJ, Faull RL, Snell RG, Hall IM, Shen Y, Ohsumi TK, Borowsky ML, Daly MJ, Lee C, Morton CC, Macdonald ME, Gusella JF, Talkowski ME

Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration.

Nature Genetics. Mar 4 2012

Talkowski ME, Ernst C, **Heilbut A**, Chiang C, Hanscom C, Lindgren A, Kirby A, Liu S, Muddukrishna B, Ohsumi TK, Shen Y, Borowsky M, Daly MJ, Morton CC, Gusella JF

Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research

American Journal of Human Genetics 8;88(4):469-81 April 2011.

Suzuki Y, Onge RP, Mani R, King O, **Heilbut A**, Labunskyy Y, Chen W, Pham L, Zhang LV, Tong AH, Nislow C, Giaever G, Gladyshew VN, Vidal M, Schow P, Lehar J, Roth FP

Knocking out multigene redundancies via cycles of sexual assortment and fluorescence selection

Nature Methods. Jan 9 2011

Lehar J, Krueger AS, Avery W, **Heilbut AM**, Johansen LM, Price ER, Rickles RJ, Short GF 3rd, Staunton JE, Jin X, Lee MS, Zimmermann GR, Borisy AA

Synergistic drug combinations tend to improve therapeutically relevant selectivity

Nature Biotechnology 27(7):659-66 2009

Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, McBroom L, Robinson MD, O'Connor L, Li M, Taylor R, Dharsee M, Ho Y, **Heilbut A**, et al.

Large-scale mapping of human protein-protein interactions by mass spectrometry

Molecular Systems Biology 3:89 2007

Lehar J, Zimmermann GR, Krueger AS, Molnar RA, Ledell JT, **Heilbut AM**, Short GF III, Giusti LC, Nolan GP, Magid OA, Lee MS, Borisy AA, Stockwell BR, Keith CT.

Chemical combination effects predict connectivity in biological systems

Molecular Systems Biology 3:80 2007

Bader GD, **Heilbut A**, Andrews B, Tyers M, Hughes T, Boone C.

Functional Genomics and Proteomics: Charting a Multidimensional Map of the Yeast Cell

Trends in Cell Biology. 2003 July.

Ho Y, Gruhler A, **Heilbut A**, et al.

Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.

Nature. 2002 Jan 10;415(6868):180-3.

Cowan KN, **Heilbut A**, Humpl T, Lam C, Ito S, Rabinovitch M.

Complete reversal of fatal pulmonary hypertension in rats by a serine elastase inhibitor

Nature Medicine 2000 Jun;6(6):698-702.