Whole-organism integrative expressome for *C. elegans* enables *in silico* study of developmental regulation

BY

LUKE A. D. HUTCHISON

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2011

Signature of Author

Luke A. D. Hutchison
20 May 2011

Certified By

Professor Isaac S. Kohane, M.D. Ph.D.
Pediatrics and Health Sciences & Technology, Harvard Medical School

Certified By

Professor Bonnie A. Berger, Ph.D.
Science, MIT

Accepted By

Professor Leslie A. Kolodziejski
Chair, Committee on Graduate Students

WHOLE-ORGANISM INTEGRATIVE EXPRESSOME FOR *C. ELEGANS* ENABLES *IN SILICO* STUDY OF DEVELOPMENTAL REGULATION

LUKE A.D. HUTCHISON

June 2011

"Nature hides her secret because of her essential loftiness,
but not by means of ruse."

– Albert Einstein

Whole-organism integrative expressome for *C. elegans* enables *in silico* study of developmental regulation

by

Luke A. D. Hutchison

Submitted to the Department of Electrical Engineering
and Computer Science on May 20, 2011,
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in Computer Science

ABSTRACT

The *C. elegans* nematode has been extensively studied as a model organism since the 1970s, and is the only organism for which the complete cell division tree and the genome are both available. These two datasets were integrated with a number of other datasets available at WormBase.org, such as the anatomy ontology, gene expression profiles extracted from 8000 peer-reviewed papers, and metadata about each gene, to produce the first ever whole-organism, cell-resolution map of gene expression across the entire developmental timeline of the organism, with the goal to find genomic features that regulate cell division and tissue differentiation. Contingency testing was performed to find correlations between thousands of gene attributes (e.g. the presence or absence of a specific 8-mer in the 3′ UTR, the CG-content of the sequence upstream of the transcriptional start site, etc.) and thousands of cell attributes (e.g. whether cells that express specific genes die through apoptosis, whether cells become neurons or not, whether cells move in the anterior or posterior direction after division). The resulting database of contingency test scores allow us to quickly ask a large number of biologically-interesting questions, like, "Does the length of introns of expressed genes increase across the developmental timeline?"; "Across what period of development and in which cell types is this specific gene most active?"; "Do regulatory motifs exist that switch on or off genes in whole subtrees of the cell pedigree?"; "Which genes are most strongly implicated in apoptosis?", etc. This whole-organism expressome enables direct and powerful *in silico* analysis of development.

Thesis Co-Supervisors:

Isaac S. Kohane, Professor, Pediatrics and Health Sciences & Technology, Harvard Medical School

Bonnie A. Berger, Professor, Applied Mathematics and Computer Science, MIT

OVERVIEW

The *C. elegans* nematode has been extensively studied as a model organism since the 1970s. *C. elegans* was also the first organism to have its genome fully sequenced, and it is the only organism for which the complete tree of cell divisions is known, from the zygote to the fully-developed adult worm. By integrating these two datasets with a number of other datasets available at WormBase.org, it is possible to start looking for a mapping from the *C. elegans* genome to its cell division tree, i.e. to identify genomic regulators of cell fate and cell phenotype.

Two different versions of the cell fate tree for *C. elegans* were linked and merged to maximize the metadata available for each cell, then the cell fate tree was cross-linked with the anatomy ontology, or hierarchical map of containment and relatedness of the worm's anatomical features. Reachability analysis was performed on the anatomy ontology to obtain a list of organs and tissue types that each cell is part of. A dataset of reported expression levels of thousands of genes in different tissue types and organs, as extracted from the gene expression results in 8000 peer-reviewed papers, was cross-linked with the anatomy ontology, and gene expression reported at tissue or organ level was propagated through the anatomy ontology to the individual cells that comprise those anatomical features. A gene metadata database was also integrated to provide metadata about the genes active in each cell. This combination of the two linked cell fate trees, the anatomy ontology, the gene expression database and the gene metadata database yields the first whole-organism, cell-resolution map of gene expression across the entire developmental timeline of the organism.

Given this integrated database of gene expression, contingency testing was performed to find correlations between thousands of different potential gene attributes (e.g. the presence or absence of a specific 8-mer in the 3′ UTR, the GC-content of the sequence upstream of the transcriptional start site, etc.) and thousands of different potential cell attributes (e.g. whether cells that express specific genes die through apoptosis, whether they become neurons or not, whether they merge into syncitia, whether they move in the anterior or posterior direction after division). The resulting database of contingency test scores allow us to quickly ask a large number of biologically-interesting questions, like "Does the length of the introns of expressed genes increase across the developmental timeline?"; "Across what period of development and in which cell types is this specific gene most active?"; "Do regulatory motifs exist that switch on or off genes in whole subtrees of the cell pedigree?"; "Which genes are most strongly implicated in apoptosis?"; "Which genes cause cells to stop dividing and become leaf nodes in the cell pedigree?", etc. In querying for genes correlated with apoptosis in cells or daughter cells, for example, the database lists a large number of genes that have not previously been implicated in apoptosis. This whole-organism expressome enables direct and powerful *in silico* analysis of development on an unprecedented scale.

CONTENTS

I INTRODUCTION 1

1 INTRODUCTION: FINDING BIOLOGY'S BLUEPRINT FOR MACRO-SCALE STRUCTURE 3
- 1.1 The fractal structural pattern language of biology 3
- 1.2 Test organism: *C. elegans* 3

2 THE *c. elegans* NEMATODE AND WORMBASE.ORG 5
- 2.1 The *C. elegans* nematode as a model organism for studying development 5
- 2.2 *C. elegans* genome statistics 6
- 2.3 WormBase.org 9

II DATA PREPROCESSING 11

3 OVERVIEW: THE DATA ANALYSIS PIPELINE 13

4 INPUT DATASETS 17
- 4.1 Datasets retrieved from WormBase.org 17
  - 4.1.1 The Anatomy Ontology 17
  - 4.1.2 The Two Electronic *C. elegans* Cell Pedigrees 17
  - 4.1.3 The Anatomy Association Database 19
  - 4.1.4 The *C. elegans* Genome 21
  - 4.1.5 The Gene Metadata Database; "gene features" 21

5 DATA PREPROCESSING (I): LINKING/MERGING CELL PEDIGREES 23
- 5.1 Overview of the two available *C. elegans* cell pedigrees 23
  - 5.1.1 The AceDB "Old Cell Pedigree" 23
  - 5.1.2 Problems with the AceDB Cell lineage 27
  - 5.1.3 The anatomy ontology pedigree 31
  - 5.1.4 Problems with the anatomy ontology pedigree 36
- 5.2 Pedigree merging 37
  - 5.2.1 The pedigree merging task 37
  - 5.2.2 Manual vs. semi-automated vs. automated pedigree merging 41
- 5.3 Final linked/merged pedigree 42
  - 5.3.1 Practical considerations 42
  - 5.3.2 Data format for the linked pedigree 45
  - 5.3.3 Examples of linked pedigree nodes 47

5.3.4 Pre-parsed metadata for linked pedigree nodes 49

5.3.5 Usefulness and simplicity of linked format 50

5.4 Annotating the linked pedigree with all known cell metadata to produce *cell attributes* 51

5.5 Adding reachable terms in anatomy ontology to each cell's attributes 51

5.6 Special case handling for unusual lineages 54

5.6.1 Definition: "Canonical Path" 54

5.6.2 Z1/Z4 handling and symmetry inversion 56

6 DATA PREPROCESSING (II): MAPPING GENE EXPRESSION ONTO THE CELL PEDIGREE 59

6.1 Overview 59

6.2 Lineage-specific vs. tissue-specific gene associations 59

6.3 The gene-annotated pedigree 60

6.4 Propagating gene metadata through genes onto the gene-annotated pedigree 61

6.4.1 Definition of *gene features* 61

7 THE *c. elegans* EXPRESSOME 63

7.1 The linked, gene-annotated *C. elegans* pedigree: an unprecedented resource 63

7.2 Varying data quality of WormBase.org data 63

7.3 On leveraging The Unreasonable Effectiveness of Data 64

7.4 Usefulness of the final linked, annotated cell pedigree 64

III SEARCHING FOR GENOMIC INFLUENCERS OF CELL PHENOTYPE 65

8 INTRODUCTION: SEARCHING FOR GENOMIC INFLUENCERS OF CELL PHENOTYPE 67

9 CELL ATTRIBUTES AND GENE PREDICATES 69

9.1 Cell attributes 69

9.1.1 Definition of cell attribute 69

9.1.2 List of cell attributes 69

9.1.3 Example cell attributes 72

9.2 Gene predicates 75

9.2.1 Definition of gene predicate 75

9.2.2 Otsu thresholding for binarization of continuous domains into predicate form 75

9.2.3 List of gene predicates studied 76

9.3 List of genes studied 82

10 CONTINGENCY TESTING 87

10.1 Overview 87

10.2 Contingency test methods 87
10.3 Bayesian contingency testing 88
10.3.1 Contingency Table 89
10.3.2 Hypothesis 89
10.3.3 Beta-Binomial model 90
10.4 The z-score of a contingency test 90
10.5 New parametric estimator for Bayesian contingency test 91
11 PERFORMING CONTINGENCY TESTS BETWEEN CELL ATTRIBUTES AND GENE PREDICATES 95
11.1 Method 95
11.2 Example usages of results 101
11.3 Pairwise comparison of z-score profile to compute a "phenotype influence network" 101
11.4 Multiresolution analysis of clusters in phenotype influence networks 102
11.5 Biological Relevance of Results 110
12 TOOL FOR VISUALIZATION OF GENE EXPRESSION PATTERNS ACROSS THE CELL PEDIGREE 113
13 CONCLUSION 115

BIBLIOGRAPHY 117

LIST OF FIGURES

Figure 1 C. elegans cell nuclei, fluorescently labeled. 6

Figure 2 The C. *elegans* cell fate tree 7

Figure 3 The C. *elegans* cell fate tree (detail) 8

Figure 4 The main data processing pipeline. 15

Figure 5 An example of a specific gene expression pattern. 20

Figure 6 Basic topological conventions in the AceDB Old Cell Pedigree. 28

Figure 7 Gender labeling in AceDB can get complicated when one gender differentiates into a cell with a different name and/or function. 28

Figure 8 Connectivity in AceDB can get complicated if the underlying biology is complicated. 29

Figure 9 An Equivalence_fate link in AceDB represents two possible interchangeable cell fates. 29

Figure 10 Equivalence fates in AceDB can get complicated, with multiple equivalence fates chained together through separate branches of the pedigree in some cases. 30

Figure 11 Normal notation in AceDB for lineage-specific cell division connects four daughter cells to parent cell and marks gender with gender-appropriate self-links. 31

Figure 12 There are many cases of strange connectivity in AceDB, in gender-specific cases. 31

Figure 13 Gender confusion in AceDB. 32

Figure 14 Another case of gender confusion in AceDB. 32

Figure 15 Gender labeling can be superfluous in AceDB. 33

Figure 16 Complexity in the data model of AceDB. 34

Figure 17 The anatomy ontology pedigree currently consists of many disjoint trees, making it hard to write whole-tree analysis algorithms. 38

Figure 18 The anatomy ontology pedigree is not fully connected and therefore cannot be traversed programmatically in its entirety. 39

Figure 19 In the anatomy ontology pedigree, gender can incorrectly switch partway along a line. 40

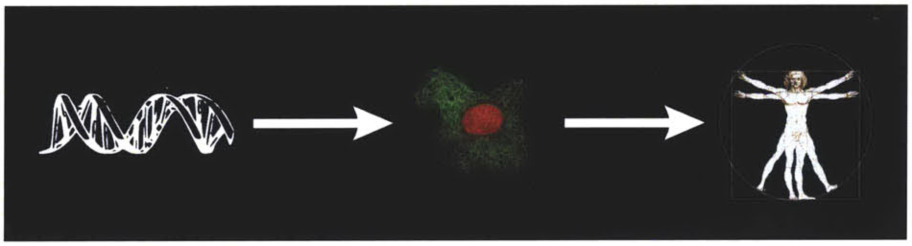
Figure 20 Pedigree merging (simple example, with only a small section of the tree shown). 43

Figure 21 Male and hermaphrodite Z1 and Z4 lineage early division patterns. 57
Figure 22 Projection of genes through tissues and organs in anatomy ontology onto individual cells 60
Figure 23 Otsu thresholding finds the threshold t which simultaneously (and equivalently) maximizes between-cluster variance (blue) and minimizes the sum of within-cluster variance (red). 76
Figure 24 (a) Contingency testing between all cell attributes and all gene predicates; (b) comparison of pairs of entire rows and columns of contingency test scores to produce a pairwise distance matrix for all pairs of cell attributes that are correlated with similar gene predicate contingency score profiles, and all pairs of gene predicates with similar cell attribute contingency score profiles. 96
Figure 25 Lineage-specific z-score results. 98
Figure 26 Tissue-specific z-score results. 99
Figure 27 The largest negative z-score results from tissue-specific gene associations. 100
Figure 28 Pairwise "phenotype influence network" for intronic motifs. 103
Figure 29 The 3′ UTR motif "phenotype influence network" with a low z-score edge-add threshold. 105
Figure 30 A pairwise "phenotype influence network" for genes 111
Figure 31 Screenshot of tree visualization tool 114

Part I

INTRODUCTION

1 INTRODUCTION: FINDING BIOLOGY'S BLUEPRINT FOR MACRO-SCALE STRUCTURE



1.1 THE FRACTAL STRUCTURAL PATTERN LANGUAGE OF BIOLOGY

The goal of this thesis was originally to test for evidence in the genome of some sort of structural pattern language that is employed by biology for encoding macro-scale structure, for example sort of a *cell addressing* mechanism, by which individual cells and/or cell pedigree path prefixes or suffixes were used to regulate gene expression or the phenotypic traits of cells.

To create a problem statement that was tractable and achievable in the scope of a PhD thesis, a more specific and concrete goal was established of developing useful concrete tools that would give insights into the genomic regulators of cell phenotype rather than attempt the larger question of searching for macro-scale structure descriptors in the genome.

1.2 TEST ORGANISM: *C. elegans*

C. elegans is the perfect organism in which to search for genomic regulators of cell phenotype. It is one of the best-studied organisms in all of biology, and in particular it is the only complex organism for which the complete genome *and* the complete cell fate pedigree tree are available (Chapter 2.1). In theory if it is possible to find some sort of functional mapping from the genome to the cell fate tree that does not simply amount to simulating the entire development process, then we will have reverse-engineered the parser for some sort of cell phenotype regulation language embedded in the DNA. This thesis employs a number of different techniques from statistics in an attempt to look for correlations between the genome of *C. elegans* and its cell division pedigree.

2 THE *C. ELEGANS* NEMATODE AND WORMBASE.ORG

2.1 THE *c. elegans* NEMATODE AS A MODEL ORGANISM FOR STUDYING DEVELOPMENT

The *C. elegans* nematode worm [1] is one of the most studied model organisms in biology, and has already been instrumental in increasing our understanding of cell cycle and the development of complex organisms [12]. There are huge numbers of data sources in different forms dealing with different aspects of the organism and at different levels of system functioning. The data available on *C. elegans* is being curated in one main central location, WormBase.org.

C. elegans became popular primarily through the efforts of Sydney Brenner to promote *C. elegans* as a model organism [6], beginning in the 1970s [5][31].

C. elegans has several properties that lend it extremely well to the study of the developmental biology of complex organisms:

1. The worm's body is transparent, meaning its internal functioning can be studied in vivo without dissection. (Figure 1.)

2. *C. elegans'* transparency allowed Sydney Brenner and John Sulston to painstakingly observe and trace every single cell division during development, from the fertilized egg to the full adult worm, which feat (along with their work on genes implicated in apoptosis) won them the Nobel Prize in 2002 along with Bob Horvitz for his work on understanding apoptosis in *C. elegans* [20]. As a result of their efforts, we now have the complete cell fate tree for *C. elegans*: the complete tree of cell divisions from the fertilized egg to the full adult worm. (Figure 2, 3.) In fact *C. elegans* is the only complex organism for which we currently have the complete cell fate tree, and the ramifications to the study of development are huge. (*C. elegans* is also the only complex organism for which the entire neural wiring diagram has been reverse-engineered.)

3. The cell lineage of *C. elegans* is almost completely invariant [10]: There are almost always exactly 556 cells in the newly-hatched worm and exactly 969 cells in the adult hermaphrodite worm (1031 cells in the adult male) after exactly 131 die through apoptosis Cel [1]. Similarly, knocking out single cells with a laser shows that cells rarely compensate for each other: cells rely more on cell-intrinsic signals and not inductive signals. The invariant lineage is useful

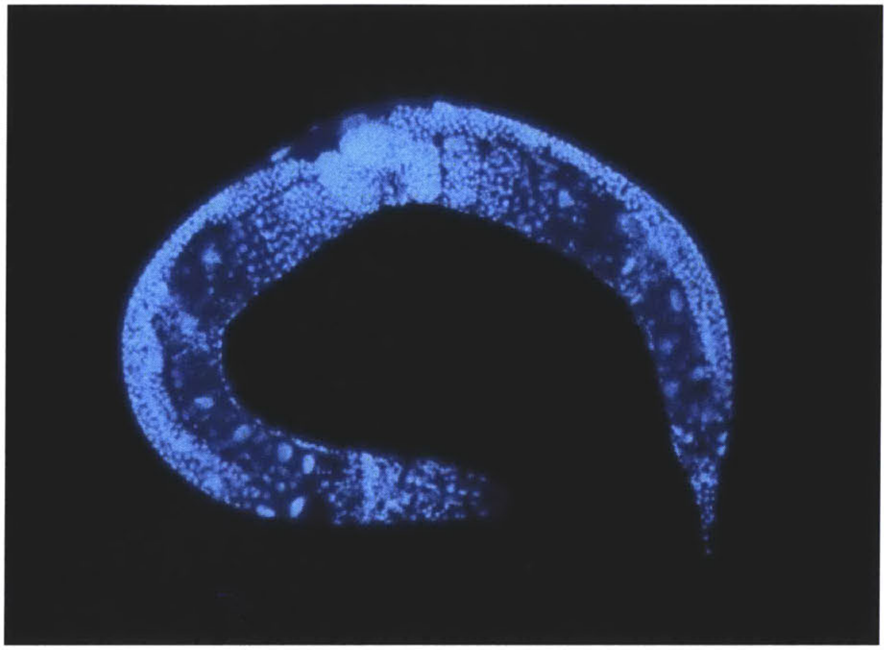


Figure 1: *C. elegans* cell nuclei, fluorescently labeled. *C. elegans* is an ideal organism for the study of development, because it is possible to view the internal structure of the worm as it grows. *[Credit: NIH]*

for studying cell-intrinsic control of cell fate (as opposed to cases where cell fate is affected by external signals). This is important when the signals are sought in the genome, because differentiation signals that come from outside the cell are likely to be too many layers of complexity removed from the genomic sequence to yield to standard inference techniques.

4. The complete genome for *C. elegans* has been sequenced and is available [21].

5. There is a huge amount of other data of different forms available on WormBase.org [7], including the curated annotations from 8000+ papers detailing which genes are active in which parts of the worm's anatomy at various stages of development.

2.2 *c. elegans* GENOME STATISTICS

C. elegans has a genome approximately 100Mb in size, containing at least 17,500 but probably over 20,000 genes [11][30]. (WormBase contains around 20,500 gene coding sequences, or around 25,000 when counting alternative splice forms; these coding sequences cover about 10% of the genome.)

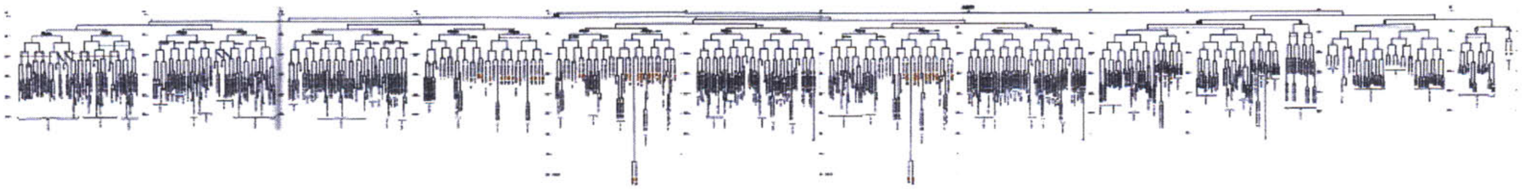


Figure 2: The *C. elegans* cell fate tree

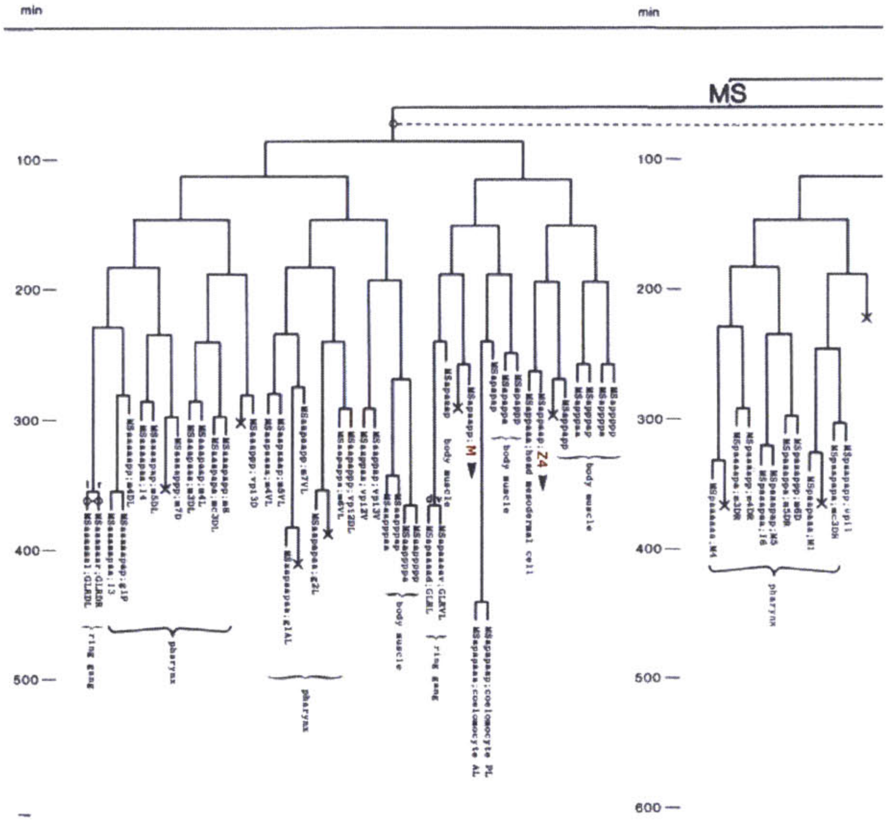


Figure 3: The *C. elegans* cell fate tree (detail)

2.3 WORMBASE.ORG

WormBase[7][2] is one of the organizations participating in the *Generic Model Organism Database (GMOD)* project. WormBase comprises at least the following main datasets[1]:

- The annotated genomes of *Caenorhabditis elegans, Caenorhabditis briggsae, Caenorhabditis remanei, Caenorhabditis brenneri, Caenorhabditis species 3, Pristionchus pacificus, Haemonchus contortus, Meloidogyne hapla, Meloidogyne incognita,* and *Brugia malayi;*
- Hand-curated annotations describing the function of ~20,500 *C. elegans* protein-coding genes and ~16,000 *C. elegans* non-coding genes;
- Gene families;
- Orthologies;
- Genomic transcription factor binding sites
- Comprehensive information on mutant alleles and their phenotypes;
- Whole-genome RNAi (RNA interference) screens;
- Genetic maps, markers and polymorphisms;
- The *C. elegans* physical map;
- Gene expression profiles (stage, tissue and cell) from microarrays, SAGE analysis and GFP promoter fusions;
- The complete cell lineage of the worm;
- The wiring diagram of the worm nervous system;
- Protein-protein interaction Interactome data;
- Genetic regulatory relationships;
- Details of intra- and inter-specific sequence homologies (with links to other model organism databases).
- In addition, WormBase contains an up-to-date searchable bibliography of *C. elegans* research and is linked to the WormBook project.

WormBase is extremely useful for a *C. elegans* integrative genomics project like the one presented in this thesis, in that it provides a central clearinghouse for numerous datasets about the worm that can be combined in limitless novel ways such that the information in one dataset can provide leverage for better understanding data in another dataset.

[1] Source of this summary: `http://en.wikipedia.org/wiki/Wormbase`

Part II

DATA PREPROCESSING

3

OVERVIEW: THE DATA ANALYSIS PIPELINE

As described previously, the "blue skies" goal of this research was to try to find some sort of systematic language for encoding structure in the genome, in other words to try to find *a mapping between the genome and the physical structure of an organism*. For *C. elegans*, we can ask a (probably) simpler and better-defined question by looking at the structure of the cell fate tree rather than the structure of the organism formed of the cells in the tree that exist at any given point; i.e. we can look for *a mapping from the genome to the cell fate tree*. This is convenient in *C. elegans* because:

Ultimate goal: to find a mapping between the genome and the physical structure of an organism.

Better-defined goal in C. elegans*: find a mapping from the genome to the cell fate tree.*

1. The complete cell fate tree is available – which is unique among organisms at this point;
2. The lineage of *C. elegans* is quite invariant, and the fate of individual cells is mostly not affected by their immediate environs: cell fate is mostly determined by cell lineage alone, i.e. by internal signaling that ultimately arises from information in the genome and not through interactions with other cells, which means that it is likely there are fewer orders of complexity between the information in the genome and cell fate in *C. elegans*.

It is likely that looking for a mapping from genome sequence to the pattern of cell divisions and tissue differentiation is a much simpler problem than looking for a mapping from genome sequence to overall organism structure, because the structure is further formed from interactions between cells that arise from the cell division process, so macro-scale structure is more orders of complexity separated from genomic encoding than the cell division tree.

The mapping from the genome to the cell fate tree goes through at least one hidden layer of complexity – probably many – so it is predictable that we could not easily find the mapping directly using some sort of blind machine learning approach, we need some sort of additional information about the hidden layers. Fortunately for *C. elegans*, we also have a lot of information about gene activity. Furthermore, since genes and their promoters can be located in the genome, and since gene activity is recorded as associated with specific regions of the anatomy, gene activity concretely ties together the domain (the genome sequence sequence) and range (the cell fate tree) of the mapping we are seeking to understand. Gene activity is therefore a great first feature to examine when trying to understand the hidden layers.

The activity of genes concretely ties together genome sequence and the cell fate tree, which will assist in finding the mapping between them.

There was previously no single database that gives gene expression for *C. elegans* for the entire organism at per-cell resolution at all stages of development – only at mixed resolution (the anatomy association database: Section 4.1.3), with gene expression recorded in some cases for individual cells but a lot more recorded for tissues, organs and other body structures. The first thing that needed to be achieved was to find which cells comprise which anatomical parts of the worm, and then to propagate the gene expression data from the per-tissue level to the per-cell level – otherwise it would be hard to correlate gene expression with the cell-based pedigree. How that is accomplished is covered in the next chapter. Also covered is work done to merge two different versions of the *C. elegans* cell pedigree together so that the structure of the pedigree is accurate while maximizing the richness of the metadata available for each cell, in order to generate a more complete picture of the phenotypic profile of each cell. These phenotypic traits are then used compared to the properties of the genes active in each cell, and a contingency test is performed between the phenotypic traits of cells and the properties of the genes expressed in those cells, in order to try to find factors that regulate gene expression that are also highly correlated with specific cell attributes.

This effectively constitutes the "low-hanging fruit" of the bigger question described above – the mapping from genome to cell pedigree. However the results thus obtained are still immensely useful, because they give lists of candidate genes (as well as other genomic features not yet described, such as sequence motifs and other genomic features like SNPs and miRNA binding sites) that are strongly implicated as being somehow involved with cellular functioning that is typically several orders of complexity above the level of gene-gene interactions that are typically studied, specifically at the level of whole-cell behavior.

Figure 4 depicts the overall data analysis pipeline. The pipeline is explained in more detail in subsequent chapters.

A number of extremely useful artifacts are produced in this data analysis pipeline. In particular, by projecting gene expression onto the individual cells in the organism across all stages of development, we create a gene-annotated pedigree with huge potential for use in a variety of different ways in future research of organism development.

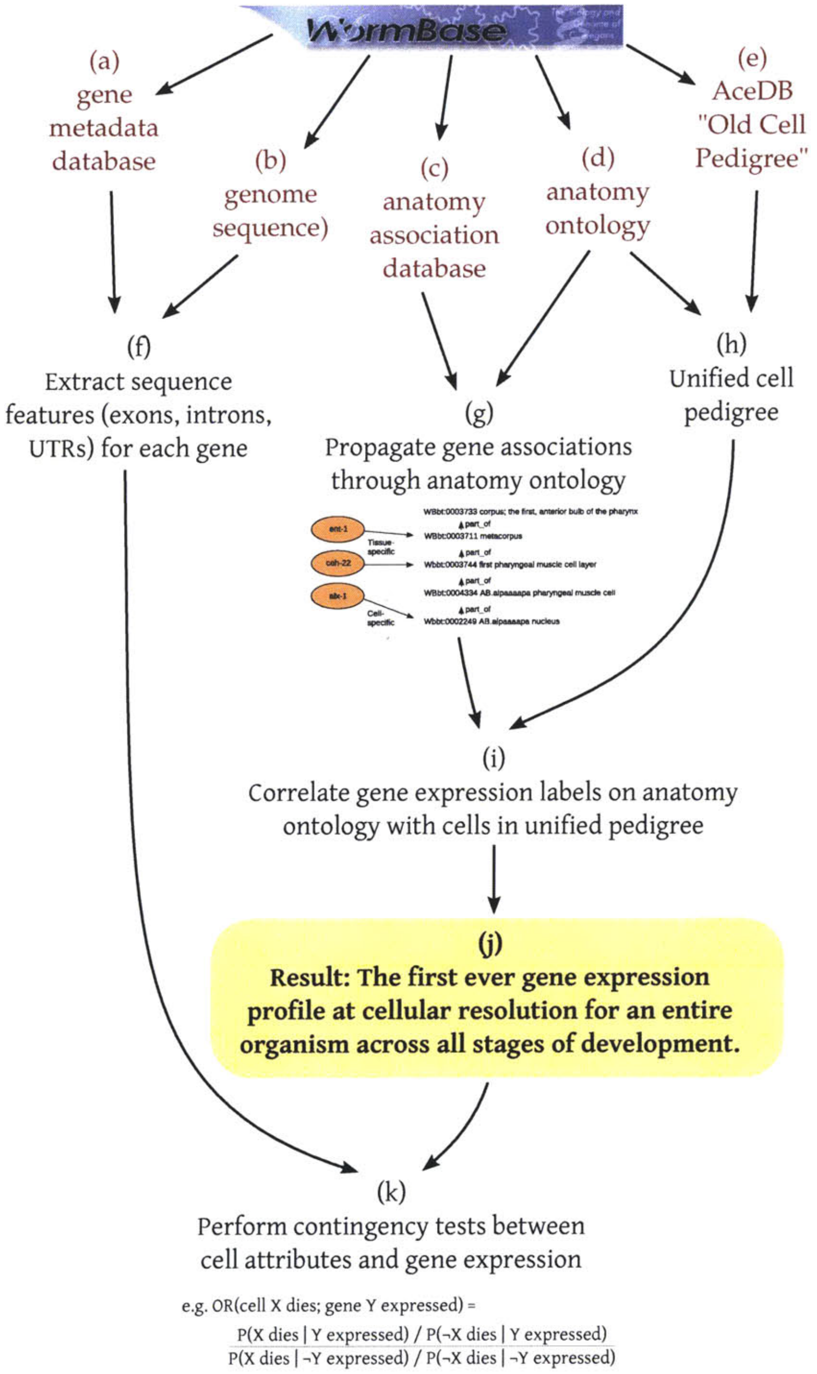

WormBase
(a)
gene
metadata
database
(b)
genome
sequence)
(c)
anatomy
association
database
(d)
anatomy
ontology
(e)
AceDB
"Old Cell
Pedigree"
(f)
Extract sequence
features (exons, introns,
UTRs) for each gene
(g)
Propagate gene associations
through anatomy ontology
WBbt:0003733 corpus; the first, anterior bulb of the pharynx
part_of
ent-1
WBbt:0003711 metacorpus
Tissue-
specific
part_of
ceh-22
Wbbt:0003744 first pharyngeal muscle cell layer
part_of
WBbt:0004334 AB.alpaaaapa pharyngeal muscle cell
part_of
Cell-
specific
Wbbt:0002249 AB.alpaaaapa nucleus
(h)
Unified cell
pedigree
(i)
Correlate gene expression labels on anatomy
ontology with cells in unified pedigree
(j)
Result: The first ever gene expression
profile at cellular resolution for an entire
organism across all stages of development.
(k)
Perform contingency tests between
cell attributes and gene expression
e.g. OR(cell X dies; gene Y expressed) =
P(X dies | Y expressed) / P(¬X dies | Y expressed)
P(X dies | ¬Y expressed) / P(¬X dies | ¬Y expressed)

Figure 4: The main data processing pipeline. (a)-(e) are separate databases from WormBase.org. Information on each node in the pipeline is contained in the following chapters / sections: (a) 4.1.5; (b) 4.1.4; (c) 4.1.3; (d) 4.1.1; (e) 4.1.2; (f) 4.1.5; (g) 6; (h) 5.1; (i) 7.1; (j) 7.1; (k) 6.4, Chapter 10.1, Chapter 10.1.

4

INPUT DATASETS

4.1 DATASETS RETRIEVED FROM WORMBASE.ORG

The following datasets were downloaded from WormBase.org and integrated:

- The anatomy ontology – including an ontology of anatomical terms, the bottom terms of which constitute cells and nuclei.
- The AceDB "Old Cell Pedigree", which preceded the anatomy ontology and has rich cell metadata.
- The anatomy association database, which associates gene expression to anatomical terms.
- The *C. elegans* genome.
- The gene metadata database.

4.1.1 *The Anatomy Ontology*

The *anatomy ontology* [25] contains all commonly-used anatomical terms (tissues, organs, structural elements, cells, nuclei), and their relationships and containment hierarchies, using is-a / has-a relationships etc.

The "bottom terms" of the anatomy ontology are cells and nuclei. However there is not much metadata available for either cells or nuclei that describe the phenotypic traits of individual cells, so other cell metadata is needed.

The anatomy ontology is a very large and complicated directed acyclic graph (DAG), and is too large to lay out in a figure here. However some glimpses into its structure and content are given in Section 5.5.

4.1.2 *The Two Electronic* C. elegans *Cell Pedigrees*

4.1.2.1 *The Old Cell Pedigree (AceDB)*

AceDB is an older *C. elegans* database used to obtain what is known as the "old Cell pedigree". This is the original WormBase.org cell fate tree; it has been superseded by anatomy ontology (described below) because it doesn't contain cells' relationships to anatomical features, and because the format cannot properly capture the DAG-like nature of the actual cell pedigree (including muscle cell fusion events etc.). There is rich metadata available for each cell, e.g. the developmental timing of cell divisions, the presence of apoptosis events, cell name aliases, and even the 3D position of cells during development. However the data is quite messy, especially in the pedigree connectivity information, so his datasource needed a lot of normalization work to fully extract the pedigree info in a useful way.

```
Cell : "ABprppaapp"
Main_name         "ABprppaapp"
Other_name        "AB.prppaapp"
Program  "division"
Embryo_division_time      "305"
Parent   "ABprppaap"
Daughter          "AIAR"
Daughter          "DB7"
Lineage_name      "AB.prppaapp"
Lineage_name      "ABprppaapp"
Cell_group        "AB lineage"
Cell_group        "all enclosing embryo cells"
Life_stage        "embryo"
Life_stage        "enclosing embryo"
Life_stage        "gastrulating embryo"
Life_stage        "late cleavage stage embryo"
Life_stage        "proliferating embryo"
Anatomy_term      "WBbt:0006684"
Tree_Node         "ABprppaapp"
Reconstruction    "N2-EMB-1" Birth 203.000000
Reconstruction    "N2-EMB-1" Timepoint 215.000000 XYZ 35.700001
    9.900000 35.200001
Reconstruction    "N2-EMB-1" Timepoint 230.000000 XYZ 37.299999
    10.600000 35.200001
Reconstruction    "N2-EMB-1" Timepoint 245.000000 XYZ 37.500000
    12.500000 35.200001
```

4.1.2.2 *The Anatomy Ontology Pedigree*

The anatomy ontology contains a cell nucleus pedigree as part of the ontology ("descendant-of", "descendant-in-male", "descendant-in-herm"), but the actual Cell term hierarchy has not yet been finished by WormBase contributors. There are numerous other problems with this datasource because this particular ontology is still a work in progress, and because numerous terms need cleaning up (e.g. the nucleus pedigree sometimes consists of disjoint subtrees; naming systems are not always consistent; some of the information in the file is stored in notes rather than normalized symbolic form; etc.)

```
[Term]
id: WBbt:0004327
name: pm2VR
alt_id: WBbt:0003650
def: "Pharyngeal muscle cell (nucleus)" [ISBN:0-87969-307-X]
synonym: "lineage name\: AB.arapaaaap" []
synonym: "m2VR" []
is_a: WBbt:0004017
relationship: develops_from WBbt:0002258
relationship: part_of WBbt:0003633

[Term]
id: WBbt:0002520
name: P6.papl nucleus
def: "nucleus of pedigree P6.papl" [WB:rynl]
relationship: DESCINHERM WBbt:0002519

[Term]
id: WBbt:0004017
name: Cell
alt_id: WBbt:0003730
def: "a cellular object that consists of subcellular components\,
    expresses genes or functions." [WB:rynl]
synonym: "Cell type" []
is_a: WBbt:0000100

[Term]
id: WBbt:0006988
name: P6.pp
def: "cell\, posteior daughter of P6.p" [WB:rynl ""]
is_a: WBbt:0004017
```

Note that in the anatomy ontology, the cells are not connected into a lineage, the nucleus terms are, whereas in the Old Cell Pedigree the cells are connected. There are many such "impedance mismatches" between the two pedigrees that made linking information from these two sources together difficult (see Section 5.1).

4.1.3 *The Anatomy Association Database*

The *anatomy association database* contains a mapping between gene names and anatomical terms; produced by WormBase contributors by curating results from more than 8000 different scientific papers that have published gene expression results for *C. elegans*. The resolution of each gene association varies from cell-specific ("ced-1 → ABpappa") to tissue-specific ("clic-1 → endothelium"), and varies in certainty (Uncertain, Partial Cer-

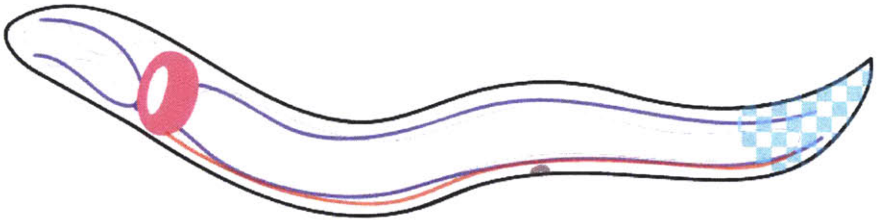


Figure 5: An example of a specific gene expression pattern referenced by Expr#### in the anatomy association file.

tain or unspecified). Each association is also given an expression pattern ID, and visualizations have been generated for the anatomical distribution of each expression pattern (e.g. Figure 5.)

```
WB      WBGene00001752  gst-4           WBbt:0003675    WBPaper
    00005888 Expr_pattern    Expr2612
WB      WBGene00001752  gst-4           WBbt:0003681    WBPaper
    00005888 Expr_pattern    Expr2612
WB      WBGene00000253  bli-3           WBbt:0005733    WBPaper
    00004841 Expr_pattern    Expr2613
WB      WBGene00001981  hnd-1   Uncertain       WBbt:0006444
    WBPaper00006002 Expr_pattern    Expr2615
WB      WBGene00006742  unc-2   Certain WBbt:0006749    WBPaper
    00006005 Expr_pattern    Expr2616
WB      WBGene00006742  unc-2           WBbt:0004758    WBPaper
    00006005 Expr_pattern    Expr2616
WB      WBGene00001053  dop-2   Certain WBbt:0004007    WBPaper
    00006021 Expr_pattern    Expr2618
WB      WBGene00000036  ace-2   Uncertain       WBbt:0003829
    WBPaper00006040 Expr_pattern    Expr2633
WB      WBGene00000036  ace-2   Certain WBbt:0004074    WBPaper
    00006040 Expr_pattern    Expr2633
WB      WBGene00000037  ace-3           WBbt:0004096    WBPaper
    00006040 Expr_pattern    Expr2634
WB      WBGene00000037  ace-3           Wbbt:0004947    WBPaper
    00006040 Expr_pattern    Expr2634
WB      WBGene00000037  ace-3           WBbt:0004949    WBPaper
    00006040 Expr_pattern    Expr2634
WB      WBGene00000516  cki-1   Partial WBbt:0003681    WBPaper
    00006027 Expr_pattern    Expr2640
WB      WBGene00001485  fre-1   Certain WBbt:0005451    WBPaper
    00006128 Expr_pattern    Expr2642
```

4.1.4 *The* C. elegans *Genome*

The *C. elegans* genome is approximately 100Mbase in size, and is available in FASTA .fa format, which lists chromosome number followed by sequence.

```
>I
gcctaagcctaagcctaagcctaagcctaagcctaagcctaagcctaagc
ctaagcctaagcctaagcctaagcctaagcctaagcctaagcctaagcct
aagcctaagcctaagcctaagcctaagcctaagcctaagcctaagcctaa
gcctaagcctaagcctaagcctaagcctaagcctaagcctaagcctaagc
ctaagcctaagcctaagcctaagcctaagcctaagcctaagcctaagcct
```

4.1.5 *The Gene Metadata Database; "gene features"*

The *GFF3 gene info file* is a huge database containing metadata on all known genomic features for *C. elegans*, for example gene intron/exon/TSS position, SNPs, RNA sequences, etc. The database landmark coordinates must match the reference genome version.

```
I       Coding_transcript       gene    1016407 1017182 .       +
        .                               ID=Gene:WBGene00005002;Name=
    WBGene00005002;Alias=C54G6.5,spp-17;Dbxref=CGC:spp-17
```

From this data, *"gene features"* are extracted for each gene: 5′ and 3′ UTR sequence, introns, exons and 1kb of sequence upstream of the first 5′UTR. Many other genome landmarks are extracted (SNPs, short tandem repeats, miRNA binding sites, etc. etc.) and the closest gene feature to each is found (see Section 9.2.3).

5

DATA PREPROCESSING (I): LINKING/MERGING CELL PEDIGREES

5.1 OVERVIEW OF THE TWO AVAILABLE *C. elegans* CELL PEDIGREES

As mentioned previously, the cell pedigree for *C. elegans* at WormBase.org exists in two different forms:

1. The "Old Cell Pedigree" extracted from AceDB
2. The nucleus lineage terms in the anatomy ontology, connected via "DESCENDANTOF", "DESC_IN_MALE" and "DESC_IN_HERM" links.

These data sources are both very informative for *C. elegans* research where a researcher is manually examining a lineage chart to see approximately where a cell might fit into the pedigree, but there are many problems with using these data sources for whole-organism statistical search for developmental regulators, because neither data source has complete enough information (gene expression in the anatomy association database is symbolically linked to the the anatomy ontology, but the AceDB pedigree has much richer per-cell metadata). To create a gene-linked pedigree with the richest per-cell metadata possible (so that gene expression can be linked to cell metadata), we need to link or merge these two pedigrees.

Gene expression is linked to the anatomy ontology cell lineage pedigree, but the AceDB "Old Cell Pedigree" has much richer cell metadata, so we need to link the two pedigrees.

5.1.1 *The AceDB "Old Cell Pedigree"*

The Old Cell Pedigree has rich metadata about each cell. For example, the record for the cells AB.alaaaarr and AB.alaaaarra are as follows:

```
Cell : "ABalaaaarr"
Main_name       "ABalaaaarr"
Other_name      "AB.alaaaarr"
Program  "division"
Embryo_division_time     "315"
Remark   "AB.alaaaarl and AB.alaaaarr: identical lineage"
Parent   "ABalaaaar"
Daughter        "RMER"
Daughter        "ABalaaaarra"
Lineage_name    "AB.alaaaarr"
Lineage_name    "ABalaaaarr"
Cell_group      "AB lineage"
Cell_group        "all enclosing embryo cells"
Life_stage        "embryo"
Life_stage        "enclosing embryo"
Life_stage        "gastrulating embryo"
Life_stage        "late cleavage stage embryo"
Life_stage        "proliferating embryo"
Anatomy_term      "WBbt:0006500"
Tree_Node         "ABalaaaarr"
Reconstruction    "N2-EMB-1" Birth 207.000000
Reconstruction    "N2-EMB-1" Timepoint 215.000000 XYZ 9.300000
    18.400000 27.200001
Reconstruction    "N2-EMB-1" Timepoint 230.000000 XYZ 8.000000
    19.700001 28.799999
Reconstruction    "N2-EMB-1" Timepoint 245.000000 XYZ 9.800000
    21.100000 27.200001

Cell : "ABalaaaarra"
Main_name         "ABalaaaarra"
Other_name        "AB.alaaaarra"
Brief_id          "programmed cell death"
Program  "death"
Embryo_division_time     "405"
Remark   "dies"
Parent   "ABalaaaarr"
Lineage_name      "AB.alaaaarra"
Lineage_name      "ABalaaaarra"
Cell_group        "cells_that_die"
Cell_group        "embryonic_death"
Cell_group        "AB lineage"
Cell_group        "all enclosing embryo cells"
Life_stage        "bean embryo"
Life_stage        "comma embryo"
Life_stage        "elongating embryo"
Life_stage        "embryo"
Life_stage        "enclosing embryo"
Life_stage        "late cleavage stage embryo"
Life_stage        "proliferating embryo"
Anatomy_term      "WBbt:0006224"
Tree_Node         "ABalaaaarra"
```

5.1.1.1 *Fields present in Cell records in AceDB*

Anatomy_term the Term in the anatomy ontology that corresponds with this AceDB Cell. This is inconsistently present and is inconsistent in which class of Terms it links too (Cells, syncitia, etc.)

Brief_id Seems to be used to give freeform information about unusual traits of the cell, e.g. programmed cell death in one or both genders.

Cell The name of the cell.

Cell_group The group of cells that this cell is part of (syncitia, cells that die through apoptosis, cells of certain stages of development, etc.). There is additional metadata about each cell group in AceDB but for the purposes of the research herein, the cell group name was simply used as a semantic label to represent a grouping of cells with some common attribute or attributes, i.e. the metadata for the cell group itself was not extracted from AceDB.

Cell_type "founder", "blast", "neuron" etc.

Daughter Links to daughter Cell records. There should only be zero, two or four links: two links for the common case of cell division and four links for the case where both genders divide and produce different daughter cells (two are marked with *Male_fate* and two are marked with *Herm_fate*). However the format is not constrained to this number, so you will see cases with one or three daughters (as explained below).

Embryo_division_time The time at which the cell divides in the embryo, according to one set of measurements.

Equivalence_fate There should be two of these links if they are present, and no *Daughter* fields. (However there are incorrect cases where there are both *Daughter* fields and *Equivalence_fate* fields.) Designates two possible fates that may be assumed by the current cell. There should be another cell that also has the same two *Equivalence_fate* fields (although this is not always the case either). This is not a cell division event, but simply represents an alternative lineage decision point where two lineages may take on alternate roles.

Equivalence_origin The backlink for *Equivalence_fate*.

Herm_fate The syncitial fate of a hermaphrodite cell, or as a self-link from a Cell node back to itself, indicates the Cell is present only in the hermaphrodite.

Herm_origin The backlink for *Herm_fate*.

Life_stage An indicator as to what stage(s) of life the cell is present in, "2-cell embryo", "blastula embryo", "proliferating embryo", "gastrulating embryo", etc.

Lineage_name One or more names for the cell using lineage prefix form. Z is the zygote, AB is the early-stage AB lineage, and most cells are described as descending from one of those two developmental stages, using one letter for each cell division according to which direction the cell typically moves after division: "a" for anterior, "p" for posterior, "l" for left, "r" for right, "d" for dorsal, "v" for ventral.

Link_diagram filenames of neural wiring diagrams in the case of neurons

Main_name The name the cell is primarily referred to by in AceDB

Male_fate The syncitial fate of a male cell, or as a self-link from a Cell node back to itself, indicates the Cell is present only in the male.

Male_origin The backlink for *Male_fate.*

Neurodata The connectivity information for the neural wiring of the worm (only present for neurons).

Other_name One or more name aliases for the cell. (Nerve cells and some other cells have anatomy-based aliases)

Parent The backlink for *Daughter* links.

Program "division", "differentiation", "death in hermaphrodite", "division in male" etc.

Reconstruction A 3D reconstruction of the lineage at different time points obtained by automatic lineaging, i.e. can be used to track the motion of the nucleus of the cell over time to determine its neighboring cells etc.

Reference Reference to one or more scientific papers from which the data was derived.

Remark Free-form text comments, e.g. "Receives a few synapses in the ring, has a posteriorly directed process that runs sublaterally"; "Somatic gonad precursor cell"; "Ventral cord interneuron, like AVD but outputs restricted to anterior cord"; "Neuron, ciliated ending in head, no supporting cells, associated with ILso"

Summary Free-form text explanation for some weird effects that are not explained by the database topology itself, e.g. "either makes proctodeum OR is engulfed by F.(r/l)d"

Syncitial_cell In the case of syncitia, gives the name of the Cell node that the cell merges into, e.g. "hyp7 embryonic". Not always present.

Tree_Node Yet another cell name, more amenable for displaying the nodes in a tree.

5.1.2 *Problems with the AceDB Cell lineage*

Even though the metadata associated with each cell in the AceDB pedigree is rich and broad in scope, the data schema itself does not cleanly map onto the biology of C. *elegans*. Also the topology of the graph for gender-specific lines is quite convoluted, and topological conventions have been inconsistently applied across the pedigree.

Figure 6 shows basic conventions for connecting cells in the AceDB Old Cell Pedigree. Descendancy is shown by *Daughter* links and a link of type *Male_fate* or *Herm_fate* from a node back to itself indicates a cell is present in only one gender. If *Male_fate* or *Herm_fate* connects a cell to another cell entry in the database, typically the meaning is that a cell merges into a syncitium like hyp7. There are separate Cell entries in AceDB for each cell in the syncitium, which represents the fact that the cells merge but the nuclei in the syncitium are still separate. The **_origin* terms are back-links for **_fate* terms.

Note that even in this simple example, a lot of inconsistency can be seen: in the case of R3BL, R3stL etc. (cells linked directly into the pedigree via *Daughter* links), *Male_fate* is an incoming edge (and actually there are two such incoming edges – a self link and a link from the actual Cell node that should probably be in the pedigree, V6L.papppapa etc. – R3BL is just another name for this cell). In the case of V6L.papppp, the *Male_fate* link is an outgoing edge to a syncitium.

Note that gender-specific labeling can get complicated when a cell is present in both genders but has distinct roles in each (Figure 7), or when the underlying biology is complicated (Figure 8).

Another convention that is quite common in the AceDB pedigree is the concept of an *Equivalence_fate* (Figure 9). Biologically, two cells can switch fates depending on the order in which certain developmental events occur, e.g. the order in which they come in contact with a certain other cell. This is represented in the pedigree as *Equivalence_fate* links from each of two cells to each of the two alternative fates (with *Equivalence_origin* backlinks).

This particular convention is also not used consistently: for example, in some cases only one of the two equivalent cells is linked (SPVR is linked to B.alaarda and B.araada in the figure, but links are missing from its paired equivalent cell).

Equivalence fates can get complicated in some cases, with multiple equivalence fates chained together through separate branches of the pedigree (Figure 10). It is possible that equivalence fates can be more simply represented (see the discussion on "canonical paths" in Section 5.6.1).

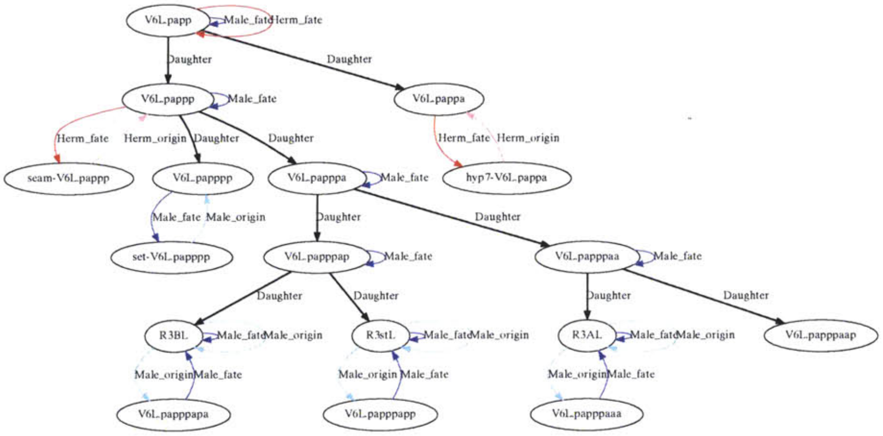


Figure 6: Basic topological conventions in the AceDB Old Cell Pedigree: *Daughter* links connect cells to two daughters; cells present in only one gender (e.g. V6L.papppap) are flagged with *Male_fate* or *Herm_fate* self-links; cells that merge into a syncitium have *Male_fate* or *Herm_fate* links.

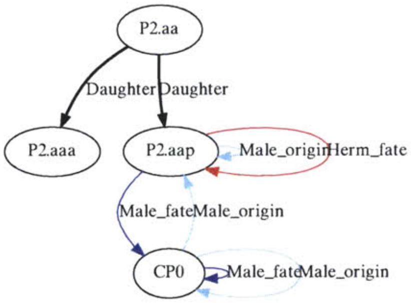


Figure 7: Gender labeling in AceDB can get complicated when one gender differentiates into a cell with a different name and/or function.

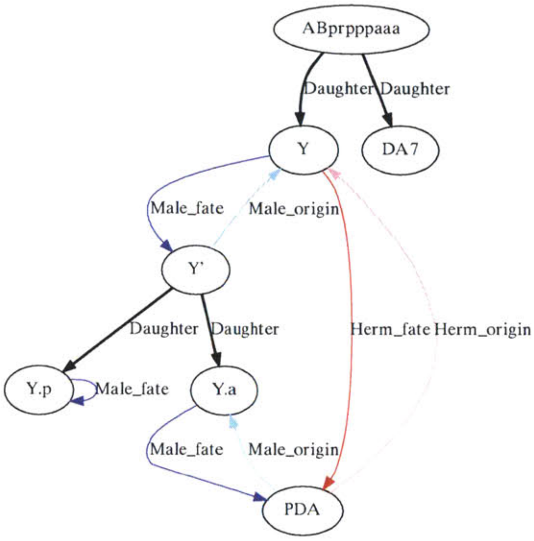


Figure 8: Connectivity in AceDB can get complicated if the underlying biology is complicated, for example in this case a cell divides one more time in male than in herm before producing the cell PDA.

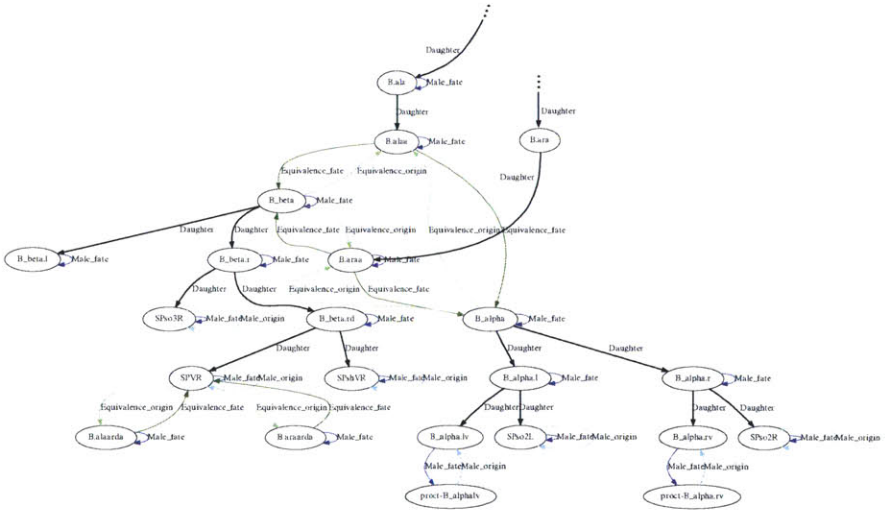


Figure 9: An Equivalence_fate link in AceDB represents two possible interchangeable cell fates.

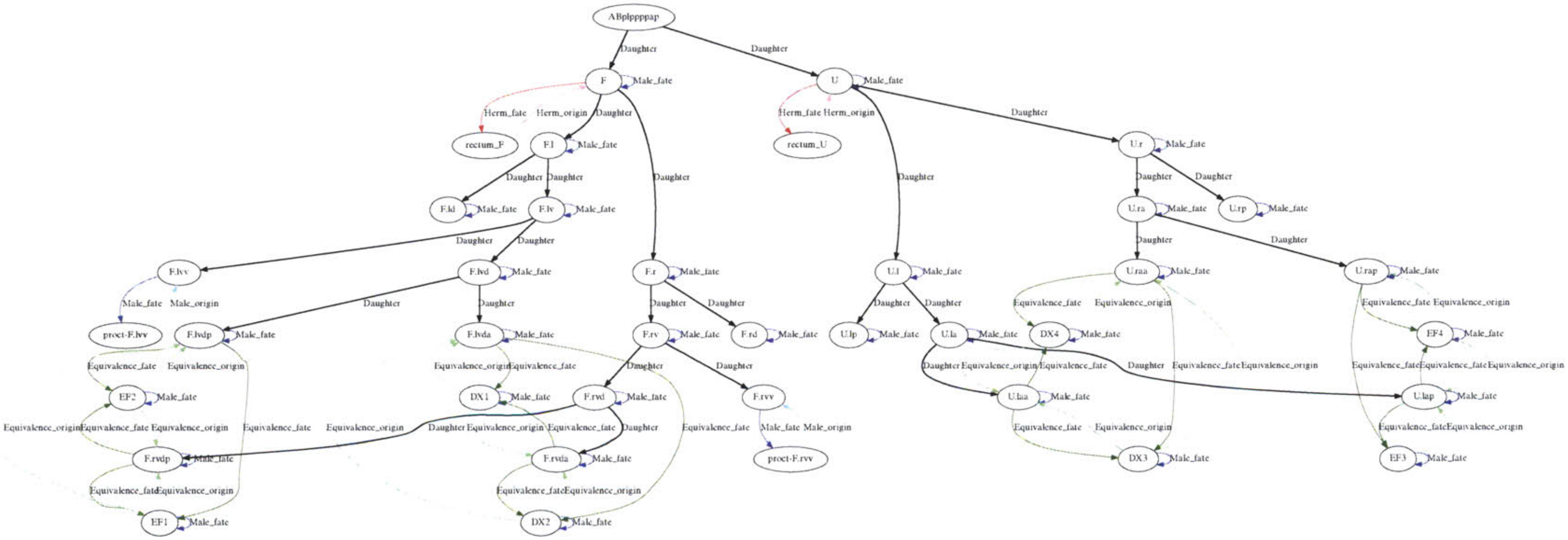


Figure 10: Equivalence fates in AceDB can get complicated, with multiple equivalence fates chained together through separate branches of the pedigree in some cases.

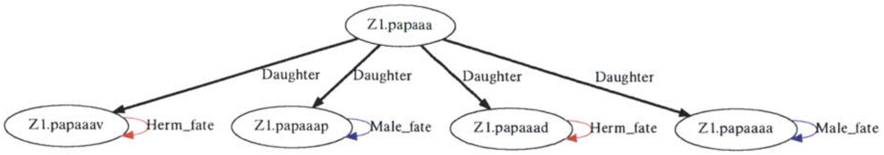


Figure 11: Normal notation in AceDB for lineage-specific cell division connects four daughter cells to parent cell and marks gender with gender-appropriate self-links.

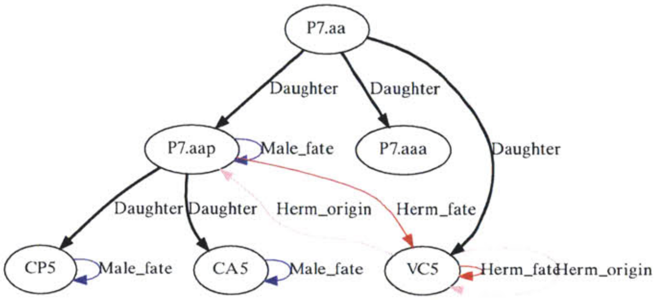


Figure 12: There are many cases of strange connectivity in AceDB, in gender-specific cases, that do not seem to conform to any specific standard. Here P7.aa has three daughter cells (the third is the herm-specific version of P7.aap, VC5).

Numerous other problems are present in the AceDB pedigree. A small sample of the problems includes the following:

- When gender-specific differentiation is present, rather than two Daughter links, a Cell node is supposed to have four Daughter links, two for each gender (Figure 11). However there are many cases of strange connectivity in gender-specific cases that do not follow this rule, e.g. one or three daughters (Figure 12).
- There are many cases of missing or contradictory gender-specific labeling, e.g. Figure 13, 14.
- Gender-specific labeling can be superfluous, in that some Cell nodes have self-links for both genders (Figure 15).

The almost *ad hoc* usage of different linking conventions in AceDB combined with the underlying complexity of the biology can lead to some very convoluted linking situations (Figure 16).

5.1.3 *The anatomy ontology pedigree*

The anatomy ontology is a newer project of the WormBase Consortium to build a complete anatomical graph for the *C. elegans* organism [25]. It is still incomplete but already consists of thousands of anatomical terms linked together in a directed acyclic graph (DAG).

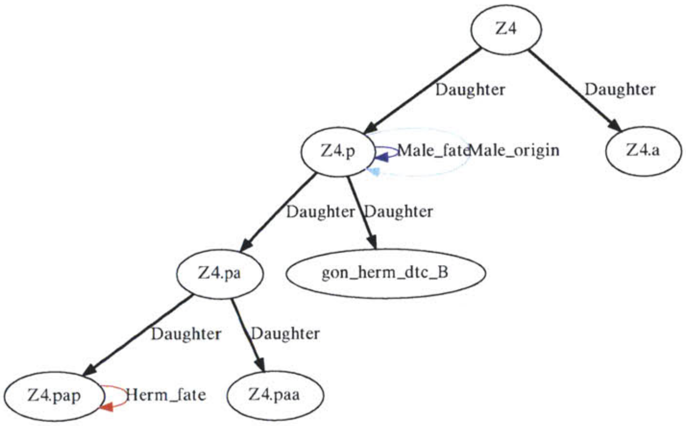


Figure 13: Gender confusion in AceDB: Z4.p is marked as a male-specific cell but has a descendant cell Z4.pap that is herm-specific (and there is no male-specific version of Z4.pap in the graph).

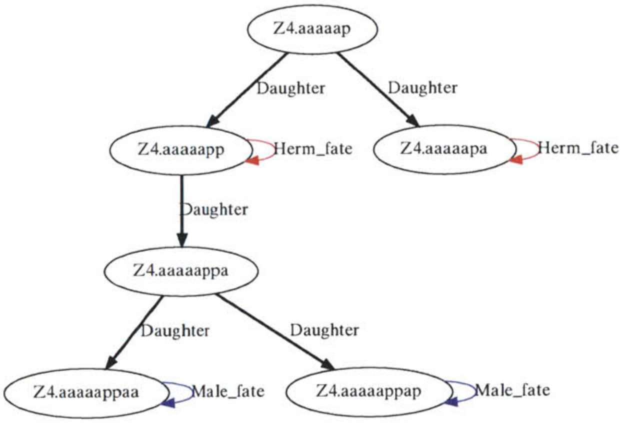


Figure 14: Another case of gender confusion in AceDB: Z4.aaaaapp is marked as herm-specific but has two descendants that are male-specific.

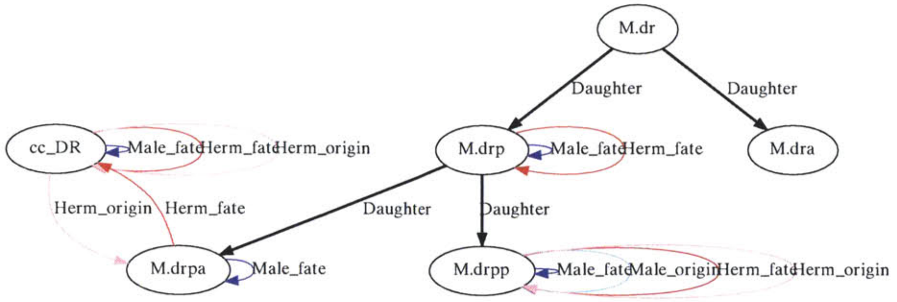


Figure 15: Gender labeling can be superfluous in AceDB (in the case of M.drpp, which is gender-specifically labeled for both genders using Male_fate and Herm_fate self-links), and/or be missing gender-specific links (cc_DR is present in male according to its Male_fate self-link, but the link should be from M.drpa to cc_DR if the Herm_fate edge is correct), and/or contain multiple contradictory information (cc_DR has both itself and M.drpa as its Herm_origin), and/or be missing information (M.drp is missing backlinks to itself – there should be a corresponding backlink Male_origin for every Male_fate link, and the same for herm), etc.

5.1.3.1 *The nucleus pedigree contained in the anatomy ontology*

As well as containing tissues, organs and body parts, the ontology also includes cell and nucleus terms. nuclei are linked together to form a pedigree. Terms are of the following form:

```
[Term]
id: WBbt:0004017
name: Cell
alt_id: WBbt:0003730
def: "a cellular object that consists of subcellular components\,
    expresses genes or functions." [WB:rynl]
synonym: "Cell type" []
is_a: WBbt:0000100

[Term]
id: WBbt:0007030
name: post-embryonic cell
def: "a cell that comes to being in a worm after hatching." [WB:rynl
    ""]
is_a: WBbt:0004017

[Term]
id: WBbt:0007048
name: B.alapa
def: "post-embryonic cell of pedigree B.alapa" [WB:rynl]
is_a: WBbt:0007030
```

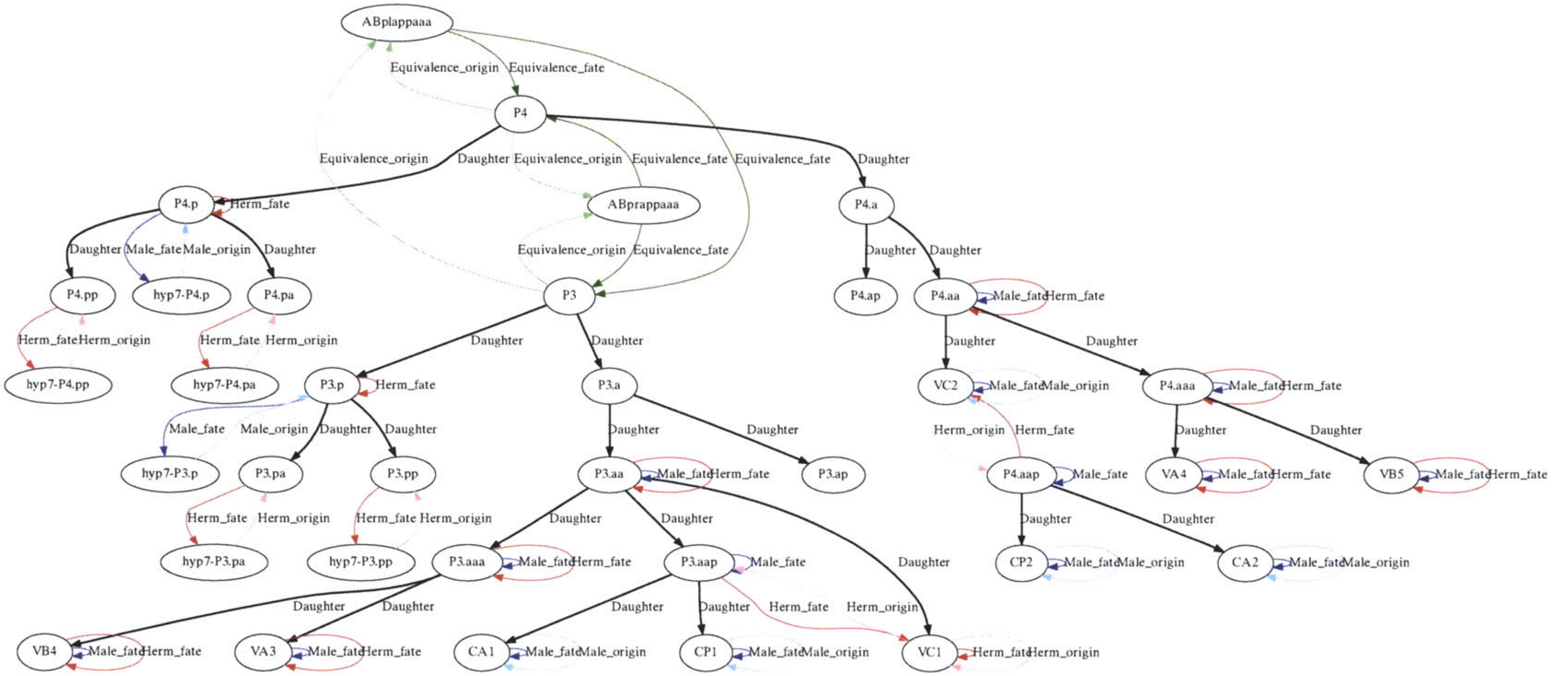


Figure 16: Complexity in the data model of AceDB. In this case within a small region of the pedigree we have an Equivalence_fate situation, several cases of gender-specific syncitia or gender-specific cell line termination (P4.p in male, P3.p in male, P3.aap in herm, VC2 in male), a node with three Daughter links (P3.aa, which is actually P3.aap in the hermaphrodite), a node that is entirely disconnected from the rest of the Daughter-linked pedigree (P4.aap) but that relies on a Herm_origin backlink to weakly connect it back in to its cell alias VC2.

```
[Term]
id: WBbt:0001720
name: B.alapa nucleus
def: "nucleus of pedigree B.alapa" [WB:rynl]
relationship: DESCINMALE WBbt:0001719

[Term]
id: WBbt:0002101
name: MS.aaaaapaa nucleus
def: "nucleus of pedigree MS.aaaaapaa" [WB:rynl]
relationship: DESCENDENTOF WBbt:0002308

[Term]
id: WBbt:0003634
name: M2 neuron
def: "neuron type\, a set of two pharyngeal motor neuron." [WB:rynl
    ""]
comment: LINKOUT\:\:WORMATLAS\:\:<http\://www.wormatlas.org/neurons.
    htm/M2.htm>
synonym: "M2" []
is_a: WBbt:0003677
is_a: WBbt:0005409
is_a: WBbt:0006840
relationship: part_of WBbt:0003732
```

Generally for the construction of the pedigree, the only types in the "relationship" field that we need to follow are DESCENDENTOF, DESCINMALE, DESCINHERM and develops_from. The other relationship types however are useful for pairing gene expression with cells, by propagating gene expression through the anatomy ontology onto the cell level (see Chapter 6).

The anatomy ontology is very much a work in progress: it is an asymptotically hard problem to build a correct ontology of all anatomical parts for a complete organism. A lot of useful information is already contained in the *C. elegans* anatomy ontology, but the cell lineage in the anatomy ontology is still somewhat incomplete, has a different topology than the AceDB pedigree in many cases, and has a data model that is not rich enough to support the complex metadata that AceDB supports.

Furthermore, the anatomy ontology pedigree is nucleus-based rather than cell-based, and represents gender-specific lines, syncitia and equivalence fates in a different way, so it doesn't match the structure of the AceDB pedigree in many places.

5.1.3.2 *Fields present in the anatomy ontology*

The following fields are present in Terms in the anatomy ontology:

[Term] The start of a Term record

alt_id Alternate ID(s) that can refer to the same Term in other WormBase databases

comment Typically contains a link to some resource describing the Term.

def A freeform text definition of the Term, e.g. "nucleus of pedigree AB.alapaaaaa". This field often has to be pattern-matched to locate cell Terms that correspond with a nucleus Term of the same name, e.g. "embryonic cell of pedigree AB.alapaaaaa".

exact_synonym Term IDs for Terms that are exactly equivalent.

id The unique Term ID.

is_a The main connector between a term and its ontological predecessor. (e.g. ABprappaaa is_a Cell).

is_obsolete "true" for old terms

name Freeform text name, e.g. "ABprappaaa" (for cells), "ABprappaaa nucleus" (for nucleus lineage terms), and "ventral cord motor neurons" etc. for non cell/nucleus-related terms.

relationship One of "DESCENDENTOF", "DESCINMALE", "DESCINHERM", "develops_from", "is_a", "part_of".

synonym Other names for the given Term, e.g. the cell AB has synonyms "AB blastomere" and "Po.a". These synonyms are important in cross-linking anatomy ontology Terms to AceDB Cell nodes.

5.1.4 *Problems with the anatomy ontology pedigree*

The anatomy ontology pedigree has a number of problems:

- The anatomy ontology currently consists of many disjoint trees, sometimes linked by freeform text, such as a a term named "B_alpha or B_beta fated" that infers it is connected to the two equivalent fates B_alpha and B_beta. Having the pedigree split into multiple trees that require special parsing to reconnect makes it hard to write whole-tree analysis algorithms (Figure 17).

- The nucleus terms in the ontology are usually not linked to the cell term they are a part of, so linking nuclei to cells also often has to be done based on parsing and matching of text in freeform text fields or comments, or based on similarity of ontology term names (Figure 18). In some but not all cases, this corresponds to "Equivalence_fate" links in the AceDB pedigree.

- The gender-specific lineage representation in the anatomy ontology is different from the one in the AceDB pedigree: in AceDB, if a cell and its descendants develops differently in the two genders, then two different descendant lines are created. In the anatomy ontology, nuclei are followed, not cells, so even if the cells manifest different phenotypes in the two genders, only one nucleus node is created. Gender-specific nodes represent past-end-of-line in the other gender.

- There are a number of errors in gender linkage in the tree, including accidental switches of gender on gender-specific lines (Figure 19).

5.2 PEDIGREE MERGING

5.2.1 *The pedigree merging task*

To link gene expression (which is tied to the anatomy ontology pedigree via normalized text tokens in the gene association database) to cell phenotype attributes (most of which are only available in the AceDB "Old Cell Pedigree"), we need to link these two pedigrees together.

This turned out to me a mammoth task: there were many fields in multiple records that had to be cross-checked for each decision to unify one anatomy ontology *Term* record with one AceDB *Cell* record:

- When comparing one Term to one Cell, parent and daughter records of each also needed to be compared to make sure they matched and were connected together in the same way.

- Past the point of gender differentiation, one Term had to be compared to two Cell records in each case, one for each gender.

- In the case of Male_fate, Male_origin, Herm_fate, Herm_origin, Equivalent_fate and Equivalent_origin links, two or more records had to be compared for each Cell record to make sure they referred to the same physical cell that corresponded with the given Term.

- Each nucleus Term had to be additionally unified with a matching anatomy ontology Term. The part_of link, if present, had to be checked to see if it corresponded with a Term that represents the cell the nucleus is part of, or added if not present.

- In the case of both anatomy ontology Terms and AceDB Cells, syncitia had to be identified and the corresponding Terms and Cells that represent the syncitia also had to be unified into a single record for the cell, so that all metadata applicable to the cell could be brought into one place. The way syncitia are connected in Cells and Terms is very different, and in many cases the information is not cleanly linked.

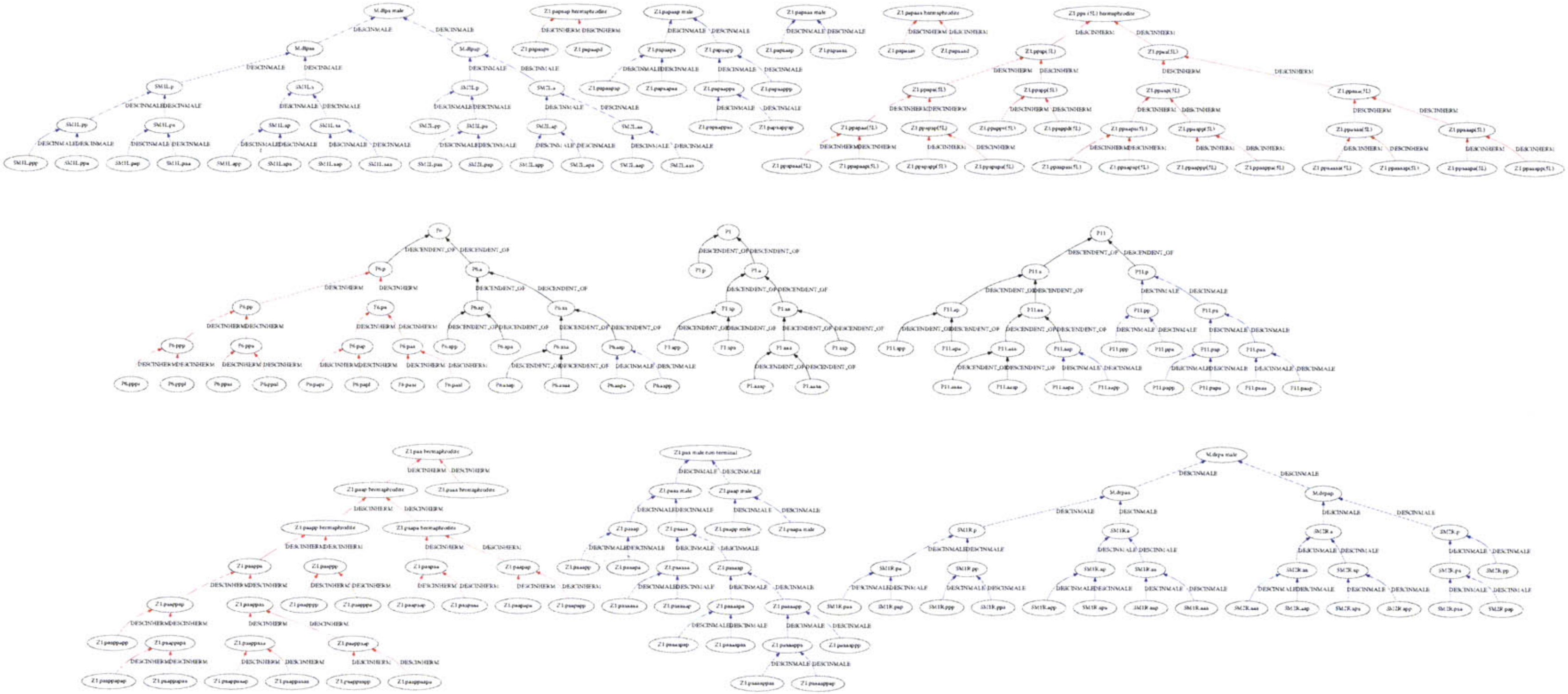


Figure 17: The anatomy ontology pedigree currently consists of many disjoint trees, making it hard to write whole-tree analysis algorithms.

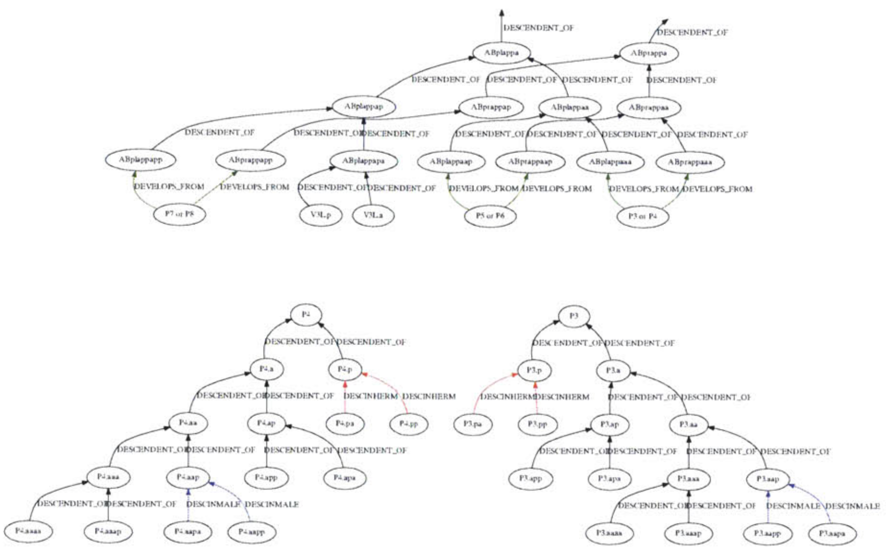


Figure 18: The anatomy ontology pedigree is not fully connected and therefore cannot be traversed programmatically in its entirety. Equivalent fates are sometimes represented by DESCENDS_FROM links to nodes, and then freeform text names are used to loosely link these nodes with disconnected subtrees. However this makes it hard to write whole-tree analysis algorithms, because the links can't be traversed directly.

- "Equivalence fate" situations are also somewhat differently encoded in the two pedigrees.

- Both pedigrees had multiple nodes representing the same cells in many situations, because there are multiple different naming schemes for many cells. Sometimes these Terms/Cells were not linked into the pedigree at all but contained useful information about the cell or potentially were linked to by entries in the gene association database, so they needed to be linked in to maximize information about each cell. Much of this linking required fuzzy string matching on name fields, notes fields, summary fields etc.

- Cell nomenclature between the two pedigrees frequently does not coincide, and the complete list of synonyms given for a certain Cell or Term often does not coincide. This makes it hard to do approximate matching based on name. Also within a given system of nomenclature, notation may be inconsistent, e.g. AB.alaaaalal is also referred to as ABalaaaalal (without the dot). And in many cases, a certain cell is given a name (e.g. F) and then naming re-starts with that cell name as a prefix, e.g. F.lvda – but it is not always the case that all pedigree nodes in both pedigrees re-base their nomenclature for all descendant cells from that reference point.

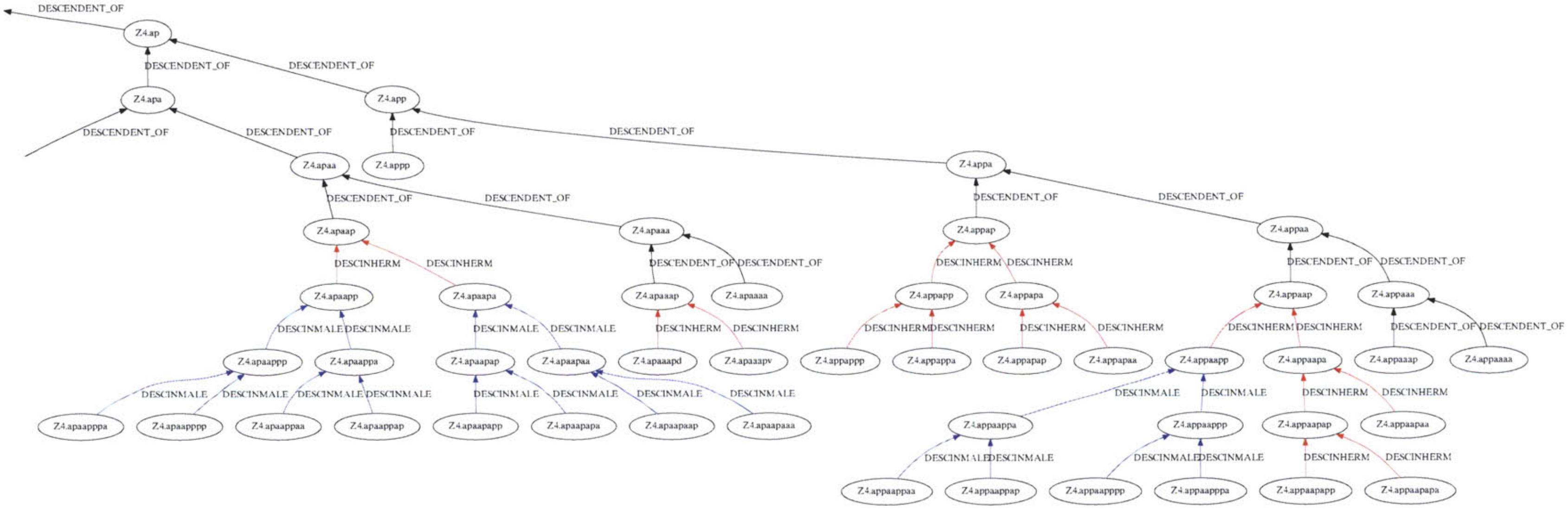


Figure 19: In the anatomy ontology pedigree, gender can incorrectly switch partway along a line, in this case between herm (red) and male (blue).

To summarize, to produce a single unified record for each cell in the pedigree, a large number of incidental records (parents/daughters) and had to be examined, along with information that needed to be directly unified with the target record (syncitia, equivalence fates, alternative names for cells), and some fuzzy searches on various fields had to be performed to find candidate matches for alternative records that could also represent any given cell.

5.2.2 *Manual vs. semi-automated vs. automated pedigree merging*

It turned out that even though there were only around 3,500 records in either of the databases to merge into a single unified pedigree, so many subtle rules emerged when going through the actual vagaries of the data that this task proved too hard to do systematically by hand – and at the same time the rules were so complex that it was often more effort to code them up than simply return to doing it the brute force way:

- By hand: it was too hard to ensure that exactly the same rules were applied in every similar situation (to guarantee the data were treated consistently), making sure to cross-check all linked records, while having to perform the searches for record IDs in a text editor by hand for each link that needed checking. The human brain shuts down very quickly when it's immersed in a large, convoluted, interlinked data graph that is not presented in graphical form.

- Graphically: graph layout algorithms were developed to try to present the data in a way that fields could quickly be cross-checked and records unified (Figure 20). But showing all the relevant fields on the graph and laying out three complex, interlinked graphs (the anatomy ontology pedigree, the AceDB pedigree, and the merged pedigree) on one 2D plane while building the right UI for merging nodes turned out to be a hard enough task that it seemed easier to going back to doing it by hand again, since it only had to be done once.[1]

- Computationally: many of the merging and cross-checking tasks were eventually performed computationally, but the more automatic linkage that was performed, the more it became apparent that the data was full of subtle quirks that required special-case handling. Nevertheless, any merging work done by hand was cross-checked with some sort of automated test after it was completed.

The literally hundreds of cross-checked field relationships and thousands of special cases that were handled will be omitted here because they

[1] This is the point at which the task started to feel like "yak shaving": http://en.wiktionary.org/wiki/yak_shaving

are best described by the automated and semi-automated merging code (present in the thesis source code), as well as the annotations made in the final merged pedigree that describe some of the major data problems that were encountered that could not be solved by smarter fuzzy matching algorithms.

5.3 FINAL LINKED/MERGED PEDIGREE

5.3.1 *Practical considerations*

5.3.1.1 *Goal: produce a practical, easy-to-parse, clean unified pedigree*

The goal in producing the final linked pedigree was to produce something relatively simple to parse and run whole-tree analyses that was completely-connected, unified as much information about each cell as possible into a single record, was "mostly correct" within the limits of accuracy of the aggregate data about each cell and other linked or similarly-named records available in each of the two pedigrees, and to link the record as accurately as possible to all cell-level or nucleus-level Terms in the anatomy ontology, so that gene expression could be correlated with each cell in the pedigree.

5.3.1.2 *Simplifying sex dimorphism*

As much as possible the complexity of the cell fate tree was represented using some representations that were simpler than those used in either AceDB or the anatomy ontology, but which still captured the biologically relevant information about each cell. For example, the decision was made to unify male-specific and hermaphrodite-specific information into a single cell record (effectively undoing AceDB's policy of splitting them into two separate lines), because it was clear that the two gender-specific lines share a common overlapping subtree (usually one gender's subtree is subsumed within the other gender's subtree), but gender-specific anatomy ontology Terms and AceDB Cells were kept separate within the linked pedigree node. Similarly, information about syncitial events was added to each participant cell's linked pedigree node, in a gender-specific way when only one gender joined a syncitium. However the complexity of "equivalence fates" was still captured directly in the linked pedigree, because the ability for two cells to exchange fate subtrees is a biologically relevant event that actually changes the topology of the tree (and therefore should affect the traversal order of any recursive analysis algorithm).

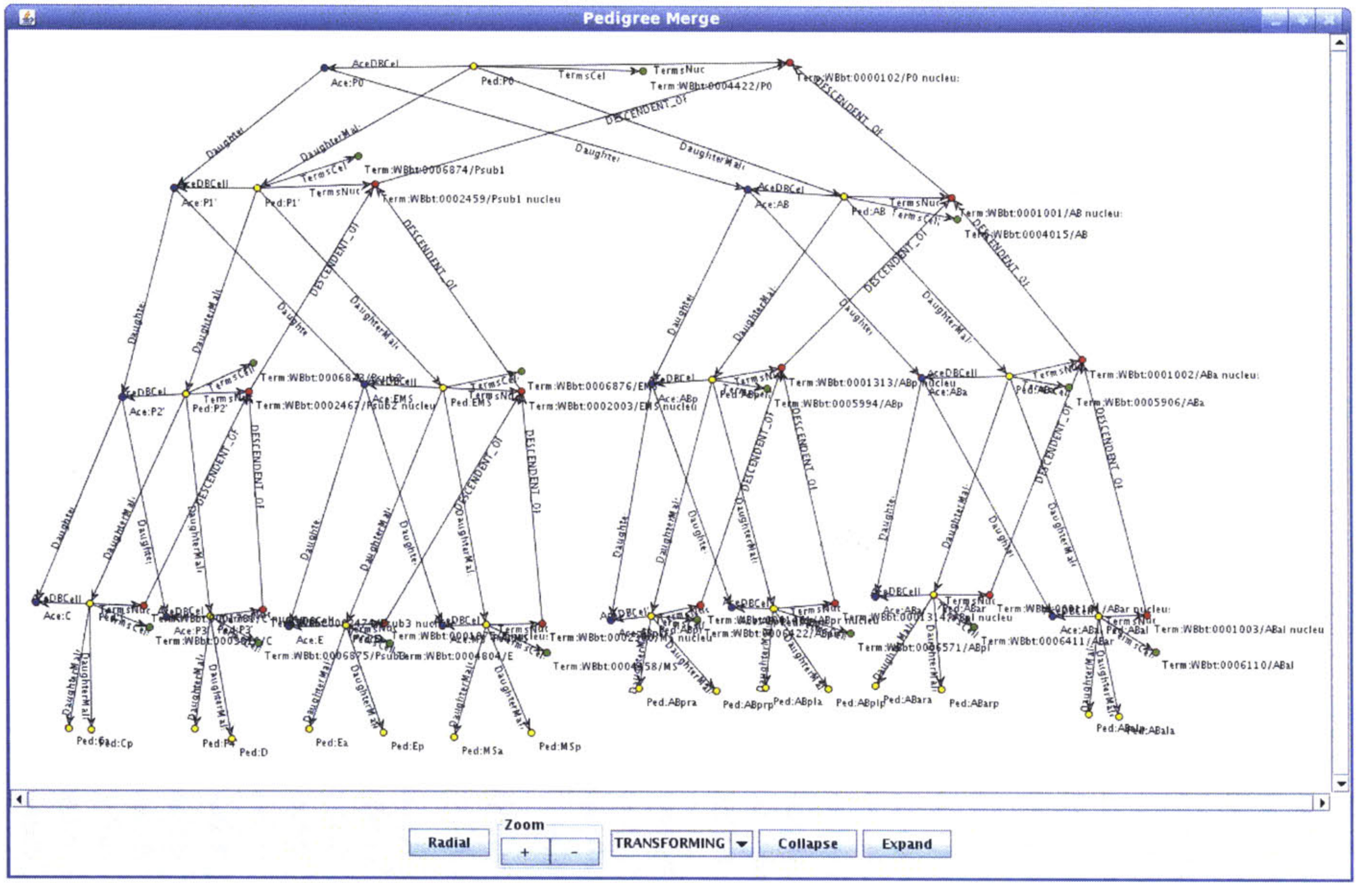


Figure 20: Pedigree merging (simple example, with only a small section of the tree shown). Blue = AceDB Old Cell Pedigree nodes; red = anatomy ontology nucleus terms; green = anatomy ontology cell terms; yellow = merged pedigree nodes. In this simple case, the graph layout problem is straightforward, but for complex situations involving many records that represented the same cell, the situation quickly became very difficult to lay out effectively (especially when fuzzy name-matches produced multiple spurious hits, and when the connectivity of the two pedigrees was not the same). Also the pedigree graph is extremely wide, making it hard to see nodes in context.

5.3.1.3 *Linked pedigree consistency guarantees*

A number of guarantees are provided by the linked pedigree that should make working with the data therein much simpler (due to the elimination of the need for certain consistency checks). Additionally, using the code produced in this thesis to parse the linked pedigree, these consistency checks are performed on every read and every write of the data. Consistency checks include the following, among many others:

1. A linked pedigree node will unify as many anatomy ontology Terms and as many AceDB Cell entries that correspond to the same physical cell as possible into one record. (There may still be some Terms or Cells that should be linked in, but most of them should be included in the final linked pedigree at this point.)

2. As a practical consideration, parsing data formats that allow the use of forward references is extra work, because it requires either two passes through the file, or the construction of "loose end" placeholder records (based on forward reference ids that have not been defined yet) that are tied together after the file is read. Since the linked pedigree is a directed acyclic graph or DAG[2], it can be sorted using a *topological sort* into an order such that every "Parent" and "EquivOrigin" backlink is to a record that has already been defined.

3. Back-links ("Parent", "EquivOrigin") and forward-links ("Daughters", "EquivFate") are always paired, i.e. where one exists, the other will always exist in the linked record, and will link correctly back in the other direction.

4. All reasonable guarantees about the gender of cell lines should be covered, e.g.

   a) When the cell line terminates for one gender through apoptosis, ceasing to divide, or becoming part of a syncitium, then beyond that point, all cells will be marked as the other gender, and there will be no switches of gender or return to bi-gendered cells in any descendant cell.

   b) If the Gender field of a linked pedigree node is not "Both" (as defined below), i.e. if the linked pedigree node is only present in one gender, then all Term and Cell fields will be null for the other gender – i.e. there won't be any links to AceDB Cells or anatomy ontology Terms for genders that don't exist at the current point.

[2] The pedigree itself is a tree, even though many cells join to form syncitia, because each syncitial event is encoded separately with each participant cell. However the presence of "equivalence fates" make the tree into a DAG, because for each equivalence fate, two nodes are both connected to each of the two equivalence fate nodes, creating multiple convergent paths through the data structure.

5. A linked pedigree node will not have both Daughter nodes and EquivFates, or a Parent node and EquivOrigins (again defined below). (This is not the case in AceDB).

6. A linked pedigree node will always have exactly zero or two children. (This is not the case in the unlinked pedigrees.)

5.3.2 *Data format for the linked pedigree*

The fields included in the final linked pedigree are defined as follows.

Note that fields marked with an asterisk (*) can consist of two separate sets of optionally tab-delimited values, one for each gender, delimited by "<tab> | <tab>" (with values for male followed with values for the hermaphrodite). When such a gender separator does not exist, the list of values for the given field is assumed to be the same for both genders. It is also possible to list "null" for one or both genders, and in particular when "Gender Male" or "Gender Herm" is specified, all Term and Cell values for the other gender need to be specified as being null (because the pedigree is beyond end-of-line for the other gender), otherwise an error is thrown during parsing. There are many examples where the list of Terms or Cells for a given liked pedigree node are not the same for both genders, even if the cell is present in both genders, for example the cell ABprpaapapp has the lineage AB.prpaapapp in both genders, but is specifically named CEMVR in male, so has different male-specific terms corresponding to CEMVR in addition to a Term and Cell entry that corresponds simply with the lineage AB.prpaapapp. (In some cases, the generic term was obviously intended to correspond only to the other gender from that of the specific term, but it appears this convention was only very inconsistently applied in both source pedigrees.)

Name The unique name of the linked pedigree node. Prefixed with "Ped:" to disambiguate it from the name of an AceDB Cell record or an anatomy ontology Term record.

Gender The gender in which the cell is present: "Male" or "Herm". (The default is "Both" if this field is not present.) If "Male" or "Herm" is specified, then by definition this cell is beyond end-of-line in the other gender.

DeathIn The gender in which the cell dies through apoptosis: "Male", "Herm" or "Both". (The default is "Neither" if this field is not present.)

SynIn If present, indicates that this cell merges into a syncitium in "Male", "Herm", or "Both" genders. (The default is "Neither" if this field is not present.)

Lineages A tab-separated list of any and all lineage names and tree paths that this cell is known by in either the anatomy ontology or AceDB. "Z." is the prefix of the zygote – i.e. Z corresponds with Po. There may be more than one lineages if the cell is also listed with a path relative to another cell, e.g. D or E.

LineagePath A single fully-qualified path from the zygote Z to the cell. This is generally unique among all linked pedigree nodes, except in the case of pairs of cells that are both equivalence fates of some other cell (and, therefore, the descendants of such pairs of equivalent cells that have overlapping path suffixes). Note that after equivalence fates, both paths to the current cell will be listed in square brackets, e.g.
[Z.aprppppapaaraa/Z.aprppppapaalaa]rda.
When passing through multiple equivalence fates to get to the same node, this process is repeated, e.g. yielding
[[Z.aprppppapaarpp/Z.aprppppapaalpp]ard/
[Z.aprppppapaarpp/Z.aprppppapaalpp]ald].

CanonicalPath The single canonical path for the cell (see Section 5.6.1). This may be shared by more than one pedigree node. Effectively paths contain tildes ("~") to indicate symmetry (usually left/right symmetry, i.e. there is an entire subtree that differs only in "l" or "r" at that position) as well as the ambiguity of equivalence paths. A "?" is also used in one place to indicate what seems to be an error in the anatomy ontology tree where a generation was missed out.

Direction The last letter of "LineagePath", interpreted as the direction the cell moves after division: "a" for anterior, "p" for posterior, "l" for left, "r" for right, "d" for dorsal, "v" for ventral. (It is hoped this may be a developmentally interesting cell phenotype attribute.)

AceDBCell* The AceDB Cell record or records that were merged into the current linked pedigree node.

SynTerms* If this cell merges into a syncitium in one or both genders, this field gives the syncitium Terms in the anatomy ontology that correspond with the final syncitium. Generally there are separate Terms for the entire syncitium as well as one term for the cell's own syncitial fate.

SynNames* The names of any syncitia that the merged AceDB Cells or anatomy ontology Terms were claimed to merge into (since these syncitial names were not always normalized and did not always correspond perfectly with an anatomy ontology Term).

TermsCell* The anatomy ontology Term or Terms that represent this cell, generally found by fuzzy matching of the name of the nucleus term that was merged into the pedigree at this position. (As indicated by the asterisk, there may be different Terms for each gender, mainly in the case of syncitia.)

TermsNuc* The anatomy ontology Term or Terms that represent the nucleus of this cell. Generally the nuclear pedigree follows the final linked pedigree.

EquivFate The "equivalence fate" of the cell, in the case where a cell can take on more than one fate. Lists the two equivalence fates and then, as the third value, other linked pedigree node that shares the same two equivalence fates.

EquivOrigin Gives the two backlinks to the nodes that have this node as an equivalence fate, and then the third value gives the other linked pedigree node that is an equal equivalence fate for those two nodes.

Parent The parent linked pedigree node, prefixed by "Ped:".

Daughters The daughter linked pedigree nodes, prefixed by "Ped:", i.e. backlinks to the Parent link.

DataProbs A tab-separated (often non-exhaustive) list of data problems encountered in merging the anatomy ontology records with the AceDB Cell records to form this linked pedigree node.

5.3.3 *Examples of linked pedigree nodes*

The final linked pedigree contains records of the following form (a range of different records are shown):

```
Name    Ped:m3VL
SynIn   Both
Lineages        AB.alpappppp
LineagePath     Z.aalpappppp
CanonicalPath   Z.aalpappppp
Direction       p
AceDBCell       M3      m3VL
SynTerms        WBbt:0003629/pm3VL-pm3VR        WBbt:0003740/pm3
           WBbt:0005595/m3[OBSOLETE]
SynNames        m3
TermsNuc        WBbt:0002265/AB.alpappppp nucleus       WBbt:0004322/
    pm3VL[ABalpappppp]
Parent  Ped:AB.alpapppp
```

```
DataProbs       (Original TermsCell Term had to get moved to TermsNuc
     because it has a DEVELOPS_FROM relationship with the TermsNuc
    term or is marked "(nucleus)" or similar)   Missing TermsCell
    term

Name    Ped:HSNR
DeathIn Male
Lineages        AB.prapppappa
LineagePath     Z.aprapppappa
CanonicalPath   Z.ap~apppappa
Direction       a
AceDBCell       HSNR
TermsCell       null    |       WBbt:0004757/HSNR[ABprapppappa]
TermsNuc        WBbt:0002089/AB.prapppappa nucleus
Parent  Ped:AB.prapppapp
DataProbs       Ped:HSNR is not marked as gender-specific, but
    AceDBCell contains a single Herm_fate: [empty fate string]
    Ped:HSNR is not marked as gender-specific, but AceDBCell contains
     gender-specific Tree_Node: herm_anat_nerv_280, HSNR

Name    Ped:TR.apapa
Gender  Male
SynIn   Male
Lineages        TR.apapa
LineagePath     Z.aprapppppapapa
CanonicalPath   Z.ap~apppppap~pa
Direction       a
AceDBCell       TR.apapa        hyp7-TR.apapa   |       null
SynTerms        WBbt:0004167/hyp7[TRapapa]      WBbt:0005146/hyp7
    post-embryonic male[OBSOLETE] |       null
SynNames        hyp7 post-embryonic male        |       null
TermsCell       WBbt:0004167/hyp7[TRapapa]      |       null
TermsNuc        WBbt:0007906/TR.apapa nucleus   |       null
Parent  Ped:TR.apap     |       null

Name    Ped:AB.prpaapapp
DeathIn Herm
Lineages        AB.prpaapapp
LineagePath     Z.aprpaapapp
CanonicalPath   Z.ap~paapapp
Direction       p
AceDBCell       ABprpaapapp     CEMVR   |       ABprpaapapp
TermsCell       WBbt:0004939/CEMVR[ABprpaapapp] WBbt:0006292/
    ABprpaapapp        |       WBbt:0006292/ABprpaapapp
TermsNuc        WBbt:0001561/ABprpaapapp nucleus
Parent  Ped:AB.prpaapap

Name    Ped:B.alaa
Gender  Male
Lineages        B.alaa
```

```
LineagePath     Z.aprppppapaalaa
CanonicalPath   Z.ap~ppppapaa~aa
Direction       a
AceDBCell       B.alaa  |       null
TermsCell       WBbt:0007046/B.alaa     |       null
TermsNuc        WBbt:0001718/B.alaa nucleus     WBbt:0007862/B_alpha
    or B_beta fated nucleus     |       null
EquivFate       Ped:B_beta      Ped:B_alpha     Ped:B.araa
Parent  Ped:B.ala       |       null

Name    Ped:B_alpha
Gender  Male
Lineages        B.alaa  B.araa
LineagePath     [Z.aprppppapaaraa/Z.aprppppapaalaa]
CanonicalPath   Z.ap~ppppapaa~aa
Direction       a
AceDBCell       B_alpha |       null
TermsNuc        WBbt:0007862/B_alpha or B_beta fated nucleus    WBbt
    :0007863/B_alpha nucleus    |       null
EquivOrigin     Ped:B.araa      Ped:B.alaa      Ped:B_beta
Daughters       Ped:B_alpha.l   Ped:B_alpha.r   |       null
DataProbs       Missing TermsCell term
```

5.3.4 *Pre-parsed metadata for linked pedigree nodes*

The linked pedigree gives references to the AceDB Cell records and anatomy ontology Term records that constitute the same physical cell. However parsing the linked pedigree for any purpose other than looking at the overall shape of the pedigree and the few cell phenotypic traits that are included directly in the linked pedigree (e.g. whether or not the cell dies through apoptosis) would require additionally parsing both other pedigree formats. As a result, all useful metadata from both other sources has been pre-parsed after the pedigrees were linked, and is available in separate files (see Chapter 6).

A number of other useful files are produced, such as an inverted index of which anatomy ontology terms are cited in which linked pedigree lineages:

```
WBbt:0002692/TL.appaaaaa nucleus    Z.aplappppappaaaaa
WBbt:0007367/TL.appaaaaa            Z.aplappppappaaaaa
WBbt:0002874/TL.appaaaap nucleus    Z.aplappppappaaaap
WBbt:0007368/TL.appaaaap            Z.aplappppappaaaap
WBbt:0002875/TL.appaaap nucleus     Z.aplappppappaaap
WBbt:0007369/TL.appaaap             Z.aplappppappaaap
WBbt:0003990/R8BL                   Z.aplappppappaaapa
```

5.3.5 *Usefulness and simplicity of linked format*

As a result of the linking efforts, the entire resulting pedigree tree can be easily parsed, by reading the records in order and (due to the topological sort order of the records) connected without the need for supporting forward references by simply using Parent and EquivOrigin links to previously-listed records.

The linked pedigree can also easily be traversed recursively, by simply visiting all Daughter nodes and EquivFate nodes from the current node. (This was not possible in either the anatomy ontology or the AceDB pedigree, because the pedigree graph was not completely connected in either case, and the linkage topology was extremely convoluted in AceDB especially.)

```
P0
|AB
||Aba
|||Abar
||||Abarp
|||||Abarpa
||||||Abarpaa
|||||||Abarpaap
||||||||Abarpaapp
|||||||||Abarpaappp
|||||||||Abarpaappa
||||||||Abarpaapa
|||||||||Abarpaapap
|||||||||Abarpaapaa
|||||||Abarpaaa
[...]
||||||||||||||Z4.aapaa
|||||||||||||||Z4.aapaad (in herm)
|||||||||||||||Z4.aapaaa (in male)
||||||||||||||||Z4.aapaaap
||||||||||||||||Z4.aapaaaa
|||||||||||||||Z4.aapaav (in herm)
|||||||||||||||Z4.aapaap (in male)
||||||||||||||||Z4.aapaapp
|||||||||||||||||Z4.aapaappa
[...]
```

Furthermore, the relationship of a cell to its parents and daughter cells, to syncitia, and to apoptosis events, is well-defined and easy to parse.

5.4 ANNOTATING THE LINKED PEDIGREE WITH ALL KNOWN CELL METADATA TO PRODUCE *cell attributes*

The final step in producing a useful pedigree is to pull in all useful cell metadata from the AceDB Cell entries (see Section 5.1.1) and the anatomy ontology Term entries (see Section 5.1.3), and combine them into a list of metadata for each linked pedigree node, so that a separate parser doesn't have to be written to make use of the data in the linked pedigree.

To simplify the analysis of cell metadata, each metadata field is turned into a Boolean attribute: if a specific metadata field (e.g. "cell is a neuron") is present, then the linked pedigree node for the cell is labeled with a cell attribute tag that implies it is of that type. The set of such tags for a cell is known as its *cell attributes*. Note that the presence of a given cell attribute indicates that the cell is believed or observed to have the given attribute, but the absence of a tag indicates *either* that the cell doesn't have the attribute *or*, in some cases, that there is insufficient data about the cell to infer that it has the attribute. In many cases with WormBase data, the data is simply too sparse for absence of evidence to be used as a proxy for evidence of absence.

The presence of a cell attribute indicates that the cell is believed to have the given attribute; the absence indicates that there is insufficient evidence to believe the cell has the attribute.
In many cases, WormBase data is too sparse for absence of evidence to be used as evidence of absence.

The types of cell attributes extracted for each cell are fully detailed in Section 9.1.2.

Once the list of attributes for each cell is generated, the linked pedigree data can easily be used without having to write additional parsers for the AceDB and anatomy ontology formats, instead only requiring a very simple parser to pull in the list of attributes for each cell from the generated file.

An example of the cell attributes extracted for a randomly-chosen cell is given in Section 9.1.3.

5.5 ADDING REACHABLE TERMS IN ANATOMY ONTOLOGY TO EACH CELL'S ATTRIBUTES

During this depth-first search process, the list of all possible paths through the anatomy ontology (following "is_a" and "part_of" links etc.), starting with each given cell, is output to a separate file, again so that all anatomic hierarchies can be easily read for a given cell without writing separate parsing and DAG traversal code for the anatomy ontology.

Some examples of complete paths through the anatomy ontology DAG are shown below. The first-level bullet point is the name of the starting cell, and the indented list below each cell is a list of all paths that can be traced through the DAG from the cell (excluding the cell itself), only reversed in order from the most general or largest term on the left to the term that is most specific to the cell on the right. (The list of paths through the anatomy ontology should give some idea of the overall structure of the DAG without showing the actual graph, which is much too large and complicated to display usefully here).

- ABaraapaaa
  - Cell/embryonic cell
- e3D
  - Anatomy/body region/digestive tract/pharynx/pharyngeal cell/pharyngeal epithelial cell/e3
  - Anatomy/body region/digestive tract/pharynx/pharyngeal segment/corpus/procorpus/e3
  - Anatomy/organ/pharynx/pharyngeal cell/pharyngeal epithelial cell/e3
  - Anatomy/organ/pharynx/pharyngeal segment/corpus/procorpus/e3
  - Cell
  - Cell/epithelial cell/interfacial epithelial cell/pharyngeal epithelial cell/e3
  - Function/Organ system/alimentary system/pharynx/pharyngeal cell/pharyngeal epithelial cell/e3
  - Function/Organ system/alimentary system/pharynx/pharyngeal segment/corpus/procorpus/e3
  - Function/Organ system/epithelial system/interfacial epithelial cell/pharyngeal epithelial cell/e3
- Z1.apa
  - Anatomy/body region/reproductive tract/gonad/hermaphrodite gonad/anterior gonad arm/anterior gonadal sheath cell/gon herm dish A
  - Anatomy/body region/reproductive tract/gonad/hermaphrodite gonad/hermaphrodite somatic gonadal cell/gonadal sheath cell/anterior gonadal sheath cell/gon herm dish A
  - Anatomy/body region/reproductive tract/gonad/somatic gonad/hermaphrodite somatic gonadal cell/gonadal sheath cell/anterior gonadal sheath cell/gon herm dish A
  - Anatomy/organ/gonad/hermaphrodite gonad/anterior gonad arm/anterior gonadal sheath cell/gon herm dish A
  - Anatomy/organ/gonad/hermaphrodite gonad/hermaphrodite somatic gonadal cell/gonadal sheath cell/anterior gonadal sheath cell/gon herm dish A
  - Anatomy/organ/gonad/somatic gonad/hermaphrodite somatic gonadal cell/gonadal sheath cell/anterior gonadal sheath cell/gon herm dish A
  - Cell/hermaphrodite somatic gonadal cell/gonadal sheath cell/anterior gonadal sheath cell/gon herm dish A
  - Cell/muscle cell/body muscle cell/smooth muscle/gonadal sheath cell/anterior gonadal sheath cell/gon herm dish A
  - Cell/muscle cell/non-striated muscle/smooth muscle/gonadal sheath cell/anterior gonadal sheath cell/gon herm dish A
  - Function/Organ system/muscular system/muscle cell/body muscle cell/smooth muscle/gonadal sheath cell/anterior gonadal sheath cell/gon herm dish A

  - Function/Organ system/muscular system/muscle cell/non-striated muscle/smooth muscle/gonadal sheath cell/anterior gonadal sheath cell/gon herm dish A
  - Function/Organ system/reproductive system/gonad/hermaphrodite gonad/anterior gonad arm/anterior gonadal sheath cell/gon herm dish A
  - Function/Organ system/reproductive system/gonad/hermaphrodite gonad/hermaphrodite somatic gonadal cell/gonadal sheath cell/anterior gonadal sheath cell/gon herm dish A
  - Function/Organ system/reproductive system/gonad/somatic gonad/hermaphrodite somatic gonadal cell/gonadal sheath cell/anterior gonadal sheath cell/gon herm dish A
  - Function/Organ system/reproductive system/muscle of the reproductive system/gonadal sheath cell/anterior gonadal sheath cell/gon herm dish A
  - Function/Sex specific entity/hermaphrodite-specific/hermaphrodite gonad/anterior gonad arm/anterior gonadal sheath cell/gon herm dish A
  - Function/Sex specific entity/hermaphrodite-specific/hermaphrodite gonad/hermaphrodite somatic gonadal cell/gonadal sheath cell/anterior gonadal sheath cell/gon herm dish A

- I4
  - Anatomy/body region/digestive tract/pharynx/pharyngeal cell/pharyngeal neuron/pharyngeal interneuron
  - Anatomy/body region/digestive tract/pharynx/pharyngeal nervous system/pharyngeal neuron/pharyngeal interneuron
  - Anatomy/body region/digestive tract/pharynx/pharyngeal segment/terminal bulb
  - Anatomy/organ/pharynx/pharyngeal cell/pharyngeal neuron/pharyngeal interneuron
  - Anatomy/organ/pharynx/pharyngeal nervous system/pharyngeal neuron/pharyngeal interneuron
  - Anatomy/organ/pharynx/pharyngeal segment/terminal bulb
  - Cell
  - Cell/neuron/interneuron
  - Cell/neuron/pharyngeal neuron/pharyngeal interneuron
  - Function/Organ system/alimentary system/pharynx/pharyngeal cell/pharyngeal neuron/pharyngeal interneuron
  - Function/Organ system/alimentary system/pharynx/pharyngeal nervous system/pharyngeal neuron/pharyngeal interneuron
  - Function/Organ system/alimentary system/pharynx/pharyngeal segment/terminal bulb
  - Function/Organ system/nervous system/pharyngeal nervous system/pharyngeal neuron/pharyngeal interneuron

Each individual node that is reachable in the anatomy ontology from a given cell is added as an "anat:"-prefixed cell attribute (see an example in Section 9.1.3). This is to allow inferences to be made about cells based not only on the properties of the individual cells themselves, but also on the types of tissues and morphological structures that each cell is part of.

5.6 SPECIAL CASE HANDLING FOR UNUSUAL LINEAGES

5.6.1 *Definition: "Canonical Path"*

An experimental alternative to the lineage path was generated for each cell in the linked pedigree, and is referred to as the "canonical path". In many places the lineage is either ambiguous (e.g. in the case of equivalence fates, which are generally symmetric to each other too), or simply symmetric (in the case of entire subtrees that are duplicated, one on the left and one on the right). The canonical path is an attempt to increase the density of the available gene expression data, by collecting gene expression from both equivalent branches. If gene expression is correlated with the lineage path and the canonical path separately, there are fewer canonical paths and so more gene expression patterns map onto each path.

It turns out that symmetries and equivalence fates almost all occurred at cell division depths 3 and 12 in branches, and there was a great degree of symmetry between subtrees beyond those points. After careful analysis and comparison of pedigree subtrees, it made sense to replace the "l"/"r" at depth 3 and the "a"/"p" at depth 12 in subtrees that exhibited equivalence fate with a wildcard "~". Equivalence fates were all in the Z.ap (AB.p) subtree. Cell division between depth 4 and 11, in all Z.ap~ lineages that divide as far as depth 12, appear to primarily be dividing for purposes of scaling up the structure of the body to the target size. Lineages that don't divide as far as depth 12 and lineages that divide beyond depth 12 appear to be more involved with building specific anatomical features. For these reasons even though there were a few lineages for which an equivalence fate was not recorded at depth 12, since equivalence fate *was* recorded for the majority of lineages that match the above description, these lineages were also given a "~" wildcard at depth 12 in their canonical path.

Some examples of canonical paths are shown below, with linked pedigree node ID on the left, lineage path in the middle, and canonical path on the right.

```
Ped:AB.pl          Z.apl                          Z.ap~
Ped:AB.pla         Z.apla                         Z.ap~a
Ped:AB.plaa        Z.aplaa                        Z.ap~aa
Ped:AB.plaaa       Z.aplaaa                       Z.ap~aaa
Ped:AB.plaaaa      Z.aplaaaa                      Z.ap~aaaa
Ped:AB.plaaaaa     Z.aplaaaaa                     Z.ap~aaaaa
Ped:AB.plaaaaaa    Z.aplaaaaaa                    Z.ap~aaaaaa
Ped:URADL          Z.aplaaaaaaa                   Z.ap~aaaaaaa
Ped:AB.plaaaaaap   Z.aplaaaaaap                   Z.ap~aaaaaap
Ped:AB.plaaaaap    Z.aplaaaaap                    Z.ap~aaaaap
...                ...                            ...
```

```
Ped:V5L.pa          Z.aplapapaappa            Z.ap~apapaapp~
Ped:V5L.paa         Z.aplapapaappaa           Z.ap~apapaapp~a
Ped:PDEL            Z.aplapapaappaaa          Z.ap~apapaapp~aa
Ped:V5L.paap        Z.aplapapaappaap          Z.ap~apapaapp~ap
Ped:PVDL            Z.aplapapaappaapa         Z.ap~apapaapp~apa
Ped:V5L.paapp       Z.aplapapaappaapp         Z.ap~apapaapp~app
...                 ...                       ...
Ped:V5L.pppppa      Z.aplapapaappppppa        Z.ap~apapaapp~pppa
Ped:V5L.pppppaa     Z.aplapapaappppppaa       Z.ap~apapaapp~pppaa
```

The canonical path can dramatically simplify how the path through the linked pedigree to any given node can be specified (with the caveat, of course, that this simplicity is only useful if the canonical path actually maps onto a biologically useful concept, which is to be investigated as future work). This is especially true in cases where there are multiple equivalence fates between the zygote and a given node:

```
Ped:P1              [Z.aplapaapp/Z.aprapaapp]     Z.ap~apaapp
Ped:P1.a            [Z.aplapaapp/Z.aprapaapp]a    Z.ap~apaappa
Ped:P1.aa           [Z.aplapaapp/Z.aprapaapp]aa   Z.ap~apaappaa
Ped:P1.aaa          [Z.aplapaapp/Z.aprapaapp]aaa  Z.ap~apaappaa~
Ped:P1.aaaa         [Z.aplapaapp/Z.aprapaapp]aaaa Z.ap~apaappaa~a
Ped:AVFR            [[Z.aplapaapp/Z.aprapaapp]aaaa/Z.aprapaapaaaa]
                                                  Z.ap~apaappaa~
Ped:AVFL            [[Z.aplapaapp/Z.aprapaapp]aaaa/Z.aprapaapaaaa]
                                                  Z.ap~apaappaa~
Ped:proctodeum      [[Z.aprppppapaarpp/Z.aprppppapaalpp]ard
                     /[Z.aprppppapaarpp/Z.aprppppapaalpp]ald]
                                                  Z.ap~ppppapaa~ppa~d
```

In one place in the pedigree, the pedigree was quite symmetric to other cousin branches but appeared offset by one generation. By symmetry, this could indicate an error in the pedigree (to be verified). There was sparse enough data in the two source pedigree versions that it was hard to verify if this was an error or was biologically significant. As a result, a "?" was inserted into the canonical path at this position to indicate the possibility of a missing cell in the lineage:

```
Ped:PVPL         Z.aplppppaaa                       Z.ap~ppppaaa
Ped:rect_VL      Z.aplppppaap                       Z.ap~ppppaap
Ped:AB.plppppap  Z.aplppppap                        Z.ap~ppppap?
Ped:U            Z.aplppppapa                       Z.ap~ppppap?a
Ped:U.l          Z.aplppppapal                      Z.ap~ppppap?al
Ped:U.la         Z.aplppppapala                     Z.ap~ppppap?a~a
Ped:U.laa        Z.aplppppapalaa                    Z.ap~ppppap?a~aa
```

```
Ped:U.ra      Z.aplppppapara                        Z.ap~ppppap?a~a
Ped:U.raa     Z.aplppppaparaa                       Z.ap~ppppap?a~aa
Ped:DX4       [Z.aplppppapalaa/Z.aplppppaparaa] Z.ap~ppppap?a~aa
Ped:DX3       [Z.aplppppapalaa/Z.aplppppaparaa] Z.ap~ppppap?a~aa
```

The actual results of analyzing the worm's canonical paths rather than lineage paths were not closely examined in this thesis, and this is left as future work. However it seems that there is something to be learned from the symmetries of much of the worm's pedigree subtrees, and analyzing gene expression using canonical paths would be an interesting place to start.

Even though the canonical path results were not examined closely here, for most result datasets, filenames ending in both "lineagePath" and "canonicalPath" were separately generated, representing running the complete analysis with cells either treated separately or with gene expression collapsed into matching canonical paths.

5.6.2 *Z1/Z4 handling and symmetry inversion*

C. elegans contains two unusual lineage subtrees, rooted at postembryonic cells Z1 and Z4, that together form both the male and hermaphrodite gonad. The complicated situation in these cells is best described in Kimble and Hirsch [23], the abstract of which states:

"The ancestry of the cells in the hermaphrodite and male gonadal somatic structures of C. elegans *has been traced from the two gonadal somatic progenitor cells (Z1 and Z4) that are present in the newly hatched larvae of both sexes. The lineages of Z1 and Z4 are essentially invariant. In hermaphrodites, they give rise to a symmetrical group of structures consisting of 143 cells, and in males, they give rise to an asymmetrical group of structures consisting of 56 cells. The male gonad can be distinguished from the hermaphrodite gonad soon after the first division of Z1 and Z4. However, the development of Z1 and Z4 in hermaphrodites shares several features in common with their development in males suggesting that the two programs are controlled by similar mechanisms. In the hermaphrodite lineage, a variability in the positions of two cells is correlated with a variability in the lineages of four cells. This variability suggests that cell-cell interaction may play a more significant role in organisms that develop by invariant lineages than has hitherto been considered. None of the somatic structures (e.g., uterus, spermatheca, vas deferens) develops as a clone of a single cell. Instead, cells that arise early in the Z1–Z4 lineage generally contribute descendants to more than one structure, and individual structures consist of descendants of more than one lineage."*

Figure 21 shows the early stages of differentiation in these two lineages.

In the final linked pedigree, the Z1 and Z4 lineages are unified and symmetrically reflected in the two genders in as reasonable a manner

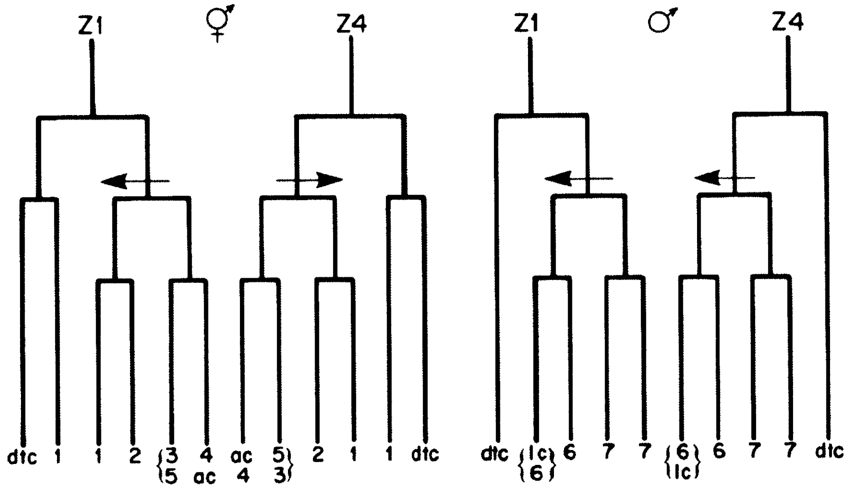

Z1
⚥
Z4
Z1
♂
Z4
dtc 1 1 2 {3 4 ac 5} 2 1 1 dtc
{5 ac 4 3}
dtc {lc 6} 6 7 7 {6 lc} 6 7 7 dtc

Figure 21: A reproduction of Figure 18 from Kimble and Hirsch [23], showing comparing the male and hermaphrodite Z1 and Z4 lineage early division patterns.

as possible according to the above principles, with the goal that a cell in the linked pedigree should represent a physical cell identity, even if that cell behaves differently in male and hermaphrodite. In particular, the 5R lineage is reflected in both anterior/posterior sense and Z1/Z4 root relative to the 5L lineage, such that for example, Z1.ppap is actually mapped onto the Z4.aapa terms in the anatomy ontology. This mapping appeared to produce the most symmetrical mapping onto what appear to be underlying actual cell identities. Examples of 5L and 5R linked pedigree nodes are shown below – notice the inversion in the 5R versions of the lineage:

```
Name     Ped:Z4.aapa(5L)
Lineages         Z4.aapa
LineagePath      Z.paaappaapaapa
CanonicalPath    Z.paaappaapaapa
Direction        a
AceDBCell        Z4.aapa
TermsCell        WBbt:0007734/Z4.aapa
TermsNuc         WBbt:0008006/Z4.aapa nucleus male       |       WBbt
    :0005704/Z4.aapa(5L) nucleus
Parent  Ped:Z4.aap(5L)
Daughters        Ped:Z4.aapaa(5L)        Ped:Z4.aapap(5L)

Name     Ped:Z4.aapa(5R)
Lineages         Z1.ppap
LineagePath      Z.paapppaapppap
CanonicalPath    Z.paapppaapppap
Direction        p
```

```
AceDBCell       Z1.ppap
TermsCell       WBbt:0005057/gon_male_sves[Z1ppap]      WBbt:0007654/
    Z1.ppap    |       WBbt:0007654/Z1.ppap
TermsNuc        WBbt:0007961/Z1.ppap nucleus male       |       WBbt
    :0005690/Z1.ppap(5R) nucleus
Parent  Ped:Z4.aap(5R)
Daughters       null    |       Ped:Z4.aapaa(5R)        Ped:Z4.aapap
    (5R)

Name    Ped:Z1.ppap(5R)
Lineages        Z4.aapa
LineagePath     Z.paaappaapaapa
CanonicalPath   Z.paaappaapaapa
Direction       a
AceDBCell       Z4.aapa
TermsCell       WBbt:0007734/Z4.aapa    |       WBbt:0007708/Z4.aap[a
    /l]
TermsNuc        WBbt:0008006/Z4.aapa nucleus male       |       WBbt
    :0003485/Z4.aap[a/l](5R) nucleus
Parent  Ped:Z1.ppa(5R)
Daughters       Ped:Z1.ppapa(5R)        Ped:Z1.ppapp(5R)
```

6

DATA PREPROCESSING (II): MAPPING GENE EXPRESSION ONTO THE CELL PEDIGREE

Understanding which cells and tissues that genes are differentially expressed in can give us a great deal of insight into the origin and function of genomic elements [13]. The gene association database was combined with the rich linked cell pedigree to try to understand the genetic regulators of cell phenotype.

6.1 OVERVIEW

As previously described, the reason for merging the two available pedigrees into one linked pedigree was that the AceDB "Old Cell Pedigree" has much richer per-cell metadata, but the gene association database links to terms in the anatomy ontology, which contains its own cell nucleus pedigree. By combining both into one pedigree, we produce a pedigree with rich cell metadata that may also be correlated with gene expression patterns.

However, the available gene expression patterns map not just to cell or nucleus terms in the anatomy ontology, but to terms throughout the ontology that correspond to tissues, organs, anatomical regions etc. Therefore to map gene expression at per-cell resolution, we need to find which terms in the anatomy ontology are comprised of which terms corresponding to cells or nuclei.

Figure 22 shows the basic procedure. Genes are associated with specific terms via the gene association database if they have been observed to be active in the corresponding part of the worm's anatomy. These associations are propagated down through the anatomy ontology by following ontology relationships in reverse (moving from B to A if A "is_a" B), until cell or nucleus terms are reached. Equivalently, a depth first search can be performed as described in Section 5.5, starting at each cell and nucleus term that was unified into a cell in the linked pedigree, and determining which genes are associated (through the gene association database) with terms that can be reached (through the anatomy ontology).

6.2 LINEAGE-SPECIFIC VS. TISSUE-SPECIFIC GENE ASSOCIATIONS

As shown in Figure 22, we keep track of whether gene associations were mapped to cells *through* anatomical terms that correspond to tissues or anatomical parts (these are referred to as *"tissue-specific associations"*) or

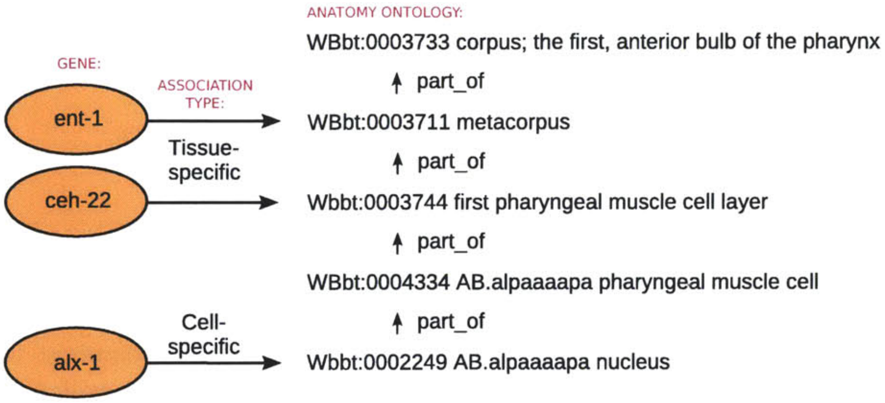

ANATOMY ONTOLOGY:
WBbt:0003733 corpus; the first, anterior bulb of the pharynx
GENE:
ASSOCIATION TYPE:
part_of
ent-1
WBbt:0003711 metacorpus
Tissue-specific
part_of
ceh-22
Wbbt:0003744 first pharyngeal muscle cell layer
part_of
WBbt:0004334 AB.alpaaaapa pharyngeal muscle cell
Cell-specific
part_of
alx-1
Wbbt:0002249 AB.alpaaaapa nucleus

Figure 22: Projection of genes through tissues and organs in anatomy ontology onto individual cells

directly to cells / nuclei (these are referred to as *"cell-specific associations"*, or more generally, "lineage-specific associations"[1]). This is because there is a lot more data for tissue-specific associations, but ultimately we can have lower confidence in it when our goal is to infer per-cell gene expression, because it is likely that in many cases, genes are recorded as expressed in all cells in a given tissue, whereas not all cells in that tissue may necessarily express the gene.

This thesis primarily focuses on lineage-specific results, but (in cross product with "lineagePath" vs. "canonicalPath" results, as discussed in Section 5.6.1) all results were generated separately for both lineage-specific and tissue-specific gene associations.

6.3 THE GENE-ANNOTATED PEDIGREE

The combination of the linked pedigree with gene association data that has been projected through the anatomy ontology produces a pedigree annotated with gene expression at per-cell resolution:

```
P0 : apx-1/C cav-1/C cdc-42/C ced-10/C chd-3/C ubxn-1/T ubxn-2/T ...
|AB : F54C8.4/C apx-1/C bub-1/C cav-1/C ced-10/T ced-3/T clh-5/T ...
||Aba : alh-9/C apx-1/C cav-1/C cdc-42/C ced-10/C chd-3/C ced-3/T ...
|||ABar : apx-1/C cav-1/C cdc-42/C ced-10/C chd-3/C B0272.3/T ...
|||| [...]
```

[1] "Lineage" is used rather than "cell" in many places is this thesis, because many cells may be combined into a single cell profile in certain circumstances. For example, in the "canonicalPath" version of analysis, multiple cells representing equivalent fates may be merged into one "lineage", i.e. the cell attributes and gene predicates for those cells will be merged. Also during contingency testing, if multiple cells have exactly the same gene expression profile, then they are collapsed as duplicates.

Note in the annotated tree, the distinction between cell-specific and tissue-specific associations (/C and /T) has been preserved, as described above.[2]

6.4 PROPAGATING GENE METADATA THROUGH GENES ONTO THE GENE-ANNOTATED PEDIGREE

The GFF3 gene metadata file from WormBase.org (described in Section 4.1.5) contains start and end genome coordinates for all *C. elegans* genes, as well as coordinates for each feature of a gene, such as introns, exons and UTRs. The GFF3 file also contains data about a large number of other genome landmarks such as predicted RNA binding sites. These features were retrieved and associated with the corresponding gene, and then additionally added to the gene-annotated pedigree, so that now the genes that are active at each cell in the pedigree have additional information about each of their features.

6.4.1 *Definition of* gene features

Gene features that were added to each gene include:

Upstream sequence For each gene, 1kb of sequence upstream of the first 5′ UTR was retrieved. (This size was picked arbitrarily.) Regulatory elements can be located upstream of the start of the gene.

5′ UTR The 5′ UTR (untranslated region) was retrieved for each gene; some genes have multiple alternative 5′ UTRs. (In *C. elegans,* 5′ UTRs can be quite short.)

3′ UTR The 3′ UTR was also retrieved for each gene, and again some genes have multiple alternative 3′ UTRs. (In *C. elegans,* 3′ UTRs are generally much longer than 5′ UTRs.)

Introns The introns for each gene were retrieved. These potentially overlapped with exons, because multiple alternative transcripts were recorded for many genes in the GFF3 file.

Exons The exons for each gene were also retrieved. This research is primarily concerned with finding gene regulatory elements,

[2] Coarse expression level and certainty information is also available in the anatomy association file for a lot of associations ("partial", "expressed", "uncertain"). This is also annotated on the tree, but is not used at the moment. (This information, because it comes from someone's reading of a scientific paper's results, is likely to be subjective and noisy anyway, but may be useful to improve the gene expression SNR.)

but it is conceivable that regulatory elements could be multiplexed into coding sequence.

SNPs, binding sites etc. The location of other genome landmarks were also recorded relative to the above genome features, e.g. if a known SNP or RNA binding site is recorded in the GFF3 gene metadata file at a location within an intron of a gene, that landmark is also associated with the gene feature

For each feature, genome sequence was extracted from the *C. elegans* genome file (in FASTA format). The sequence was then reverse-complemented if on the antisense strand (marked as "−" in the GFF3 file, vs. "+" for the sense strand), so that each gene feature's sequence was ordered in the direction of transcription.

7 THE *C. ELEGANS* EXPRESSOME

7.1 THE LINKED, GENE-ANNOTATED *C. elegans* PEDIGREE: AN UNPRECEDENTED RESOURCE

As a result of the work described in previous chapters in linking together cell metadata from two different sources and unifying the pedigree, as well as correlating anatomy-based gene expression with the individual cells that comprise each anatomical feature, we have produced a clean, fully connected, easily traversable and machine-analyzable cell fate pedigree for *C. elegans* with a full gene expression profile *for each cell in the pedigree.* We term this the *C. elegans expressome.*

We have produced the C. elegans expressome, *an extensive gene expression profile at individual cell resolution for all cells across all stages of development.*

This is the first time a whole-organism, per-cell gene expression profile has been available across all stages of development of an organism – and furthermore, each cell in this model is annotated with a rich list of phenotypic attributes.

C. elegans *is the first organism for which the expressome is available.*

The production of such a resource is unprecedented – the closest research, by Xiao et al. [27], studies only 93 genes in 363 specific cells from worms in only the L1 stage of development. In the model produced here, we associate an order of magnitude more genes with a significantly greater percentage of all cells in the cell pedigree; furthermore, our research considers a significantly greater number of phenotypic attributes for each cell than Xiao et al.

7.2 VARYING DATA QUALITY OF WORMBASE.ORG DATA

It is worth noting that the data in WormBase.org is the result of curation and submission of datasets from a large number of different data sources that were produced in different ways and under different conditions. As a result the quantity of data at WormBase.org is large but the quality and density of the data is not uniformly high. Furthermore, it is not possible in most cases to know which data is of high quality and what data is not. Therefore, compared to the research by Xiao et al. [27] (where gene activity for the entire dataset was uniformly curated), in our version of the final cell pedigree, we have significantly larger amounts of data but predict that the data quality is probably on average lower.

7.3 ON LEVERAGING THE UNREASONABLE EFFECTIVENESS OF DATA

There is a saving grace however to our large-scale integrative approach: when dataset sizes are very large and sources of error are variegated, *in aggregate* it can become easy to distinguish many of the most important signals in the data from the background noise, which tends to wash out (related to the phenomenon termed as "the unreasonable effectiveness of [big] data" *[16]*).

By employing and integrating all available data for C. *elegans* that is related to cell phenotype, cell division, tissue differentiation, gene expression and gene attributes, it is hoped that, in aggregate, a strong signal that describes genomic regulators of cell phenotype will emerge from the noise, and ultimately this signal will enable the identification of genomic features that are involved in regulating macro-scale organism structure.

7.4 USEFULNESS OF THE FINAL LINKED, ANNOTATED CELL PEDIGREE

The final cell pedigree is an immensely useful resource, and there is an almost unlimited number of things that the pedigree could be used for to gain insight into the developmental process.

In this particular research, we focus on the most immediately obvious use for the data in the annotated pedigree: looking for correlations between the *phenotypic traits of each cell* (cell attributes, see Section 9.1.1) and *shared properties of the genes that are expressed in each cell* (gene predicates such as promoter motifs that may be cell-specific, see Section 9.2.1), especially properties that are *unique to that cell* compared to other cells. Finding these correlations amounts to running contingency tests between the cell attributes and gene predicates for each cell in the pedigree (Section 10.1). Once we have the linked, gene-annotated pedigree, running analyses such as this becomes quite a straightforward process.

Since the hard work of data normalization and linking has now been done, it will be interesting to see what other uses will be found for this integrated dataset in the future.

Part III

SEARCHING FOR GENOMIC INFLUENCERS OF CELL PHENOTYPE

8

INTRODUCTION: SEARCHING FOR GENOMIC INFLUENCERS OF CELL PHENOTYPE

Eugene Wigner wrote a famous essay on the unreasonable effectiveness of mathematics in natural sciences [34]. He meant physics, of course. There is only one thing which is more unreasonable than the unreasonable effectiveness of mathematics in physics, and this is the unreasonable ineffectiveness of mathematics in biology. – attributed to Israel Gelfand

Biological complexity exhibits a massive fan-out effect at each successively larger scale of abstraction, but is also then tightly packaged and topologically separated from the next and more complex layer of abstraction by a narrow set of interaction channels, typically by means of a physical membrane of some sort.

It is far more tractable to look for genomic influencers of cell phenotype than to look for genomic influencers of organism structure, systemic disease, or especially an organism's behavior.

It is incredibly difficult to try to understand emergent complexity by looking at its constituent building blocks when those building blocks are many layers lower in the abstraction hierarchy. This is almost certainly why disease-oriented GWAS has so far proved to be somewhat of a failure at finding actual genomic causes of disease (beyond just identifying chromosomal regions with SNP features correlated with disease) [14][18][24][17][33]. In particular, statistically-significant correlations can inevitably be found for any disease because of the patient sample size relative to the size of the genome, but even if some of the associations measured as statistically-significant are indeed also biologically relevant, much more work remains to establish the functional basis for the observed associations [29]. Behavioral genomics has so far turned out to be even more of a failure, because behavior is even further removed from genotype than disease phenotypes [9][22]. For this reason, it is likely to be immensely difficult to locate the macro-scale structural blueprint for an entire organism in its genome. The goal of locating genomic features that influence cell phenotype is only slightly less ambitious, given the many layers of complexity between the genome sequence and the cell as a unit – nevertheless, hopefully this task is still far more tractable, given that the complexity of an individual cell's traits are a far cry from the complexity that emerges when many cells act together as an organism.

In this research, therefore, we first seek to find correlations between genotypic traits and cell phenotypic traits. The "Rosetta stone" for unlocking this correlation is the gene association database, which correlates gene activity (which is localized on the genome) with anatomical parts (which is therefore correlated with specific cells).

Much work in computational biology is currently focused on understanding complexity at one or two specific levels of complexity. We aim to bridge a taller jump in degrees of complexity. In the end, one of the major results of this work is a graph of genes that are co-expressed in cells that have similar phenotypic profiles. This is *not* the same thing as a gene-gene interaction network or a protein-protein interaction network – it is a correlation between low-level observations of gene activity and high-level phenotypic attributes of cells, spanning of many layers of complexity. This permits the study of the effects of genes at many cascading layers of complexity downstream from actual direct gene-gene interactions.

9 CELL ATTRIBUTES AND GENE PREDICATES

9.1 CELL ATTRIBUTES

9.1.1 *Definition of cell attribute*

A *cell attribute* is defined as the phenotypic trait of a specific cell[1]. Examples include whether a cell is a nerve cell or a muscle cell, whether a cell dies through apoptosis, whether a cell divides, differentiates or ceases to divide, whether a cell exhibits sex dimorphism (different forms in male and hermaphrodite), and whether a cell becomes part of a syncitium or not.

9.1.2 *List of cell attributes*

Cell attributes were extracted from both the anatomy ontology and the AceDB pedigrees, but more especially from the AceDB pedigree, because the cell metadata there was much richer (which was the reason for linking the two cell pedigrees together in the first place). The cell attributes that were extracted from the two pedigrees are given below. (Counts shown are for number of attributes of the form shown, not number of cells exhibiting the attribute.)

Cell Attribute Prefix *(examples in italics)*	**Count**	**Description**
anat:*pharynx*; anat:*Sex specific entity*		Cell is part of the specified anatomical structure. Specifically implies that the given anatomical term is reachable from the cell along a path through the anatomy ontology, meaning the cell is of the given type or is part of the given anatomical feature.
brief-id:*anterior inner longitudinal muscle*; brief-id:*hermaphrodite specific neuron*		Derived from the Brief_id field of AceDB record for cell

[1] At least, a phenotypic trait which WormBase.org has recorded as manifest in a specific cell.

Cell Attribute Prefix ***(examples in italics)***	**Count**	**Description**
cell-group:*Z4 lineage;* cell-group:*death_in_herms*		Derived from the Cell_group field of AceDB record for cell
cell-type:*gland;* cell-type:*neuron*		Derived from the Cell_type field of AceDB record for cell

is-leaf:*both;*
is-leaf:*herm-only;*
is-leaf:*male;*
is-leaf:*male-only;*
is-leaf:*one-or-both;*
is-leaf:*terminal*
		This cell is a leaf cell (i.e. no longer divides) in, respectively: both genders; herm only (when the cell does not exist in male); in male (whether or not the herm is a leaf cell); in male only (does not exist in herm); in one or both genders (if the cell exists for both genders, only one of them needs to be a leaf cell); and, for "terminal", in all genders that this cell exists for (i.e. if the cell exists for both genders, then both need to be leaf cells). (N.B. is-leaf:*herm* is not needed.)

life-stage:*3-fold embryo;* life-stage:*gastrulating embryo*		Derived from the Life_stage field of AceDB record for cell
lineage-birth:*111;* lineage-birth:*127*		The time point at which the cell appears in the lineage
lineage-birth-approx:*250;* lineage-birth-approx:*410*		The approximate time point at which the cell appears in the lineage, binned into 10 hour intervals
lineage-depth:*7;* lineage-depth:*13*		Number of cell divisions from the zygote
lineage-prefix:Z.*aalaapaaa;* lineage-prefix:Z.*aarppappppapapap*		Lineage prefix – all prefixes from Z.a and Z.p down to the full lineage path to the cell are added as attributes
lineage-suffix:*dlpap;* lineage-suffix:*aav*		Lineage suffix – all suffixes of lineage path of the cell, from suffix length 1 to 5

Cell Attribute Prefix ***(examples in italics)***	**Count**	**Description**
lineage-prefix-canonical: *Z.ap~ppppapaa~aald;* lineage-prefix-canonical: *Z.ap~ppppap?p~v*		Canonical lineage prefix (see text). All prefixes from Z.a and Z.p down to the full canonical path are added as attributes to each cell
lineage-suffix-canonical:*aaa~a;* lineage-suffix-canonical:*aap~*		Canonical lineage suffix (see text). All suffixes of the full canonical path are added as attributes to each cell
program:*death;* program:*division in hermaphrodite*		Derived from the Program field of AceDB record. Takes values: death, death in hermaphrodite, differentiation, differentiation in hermaphrodite, differentiation in male, division, division in hermaphrodite, division in male
syn-in:*both*; syn-in:*herm;* syn-in:*herm-only;* syn-in:*male;* syn-in:*male-only*; syn-in:*one-or-both-genders*		The cell fuses with one or more other cells and forms a syncitium in, respectively: both genders; in the herm (whether or not the cell exists in male); in the herm only (when the cell does not exist in male); in the male (whether or not the cell exists in herm); in the male only (when the cell does not exist in herm); in either male or herm or both.
syn-name:*hyp7 post-embryonic* herm; syn-name:*m3*		The name of the syncitium that this cell merges into, as used in the anatomy ontology. (Just a label, may or may not correspond with the name of an anatomy ontology Term.)
syn-term:*pm2DL-pm2DR*; syn-term:*seam*		The ID of the Term in the anatomy ontology that this cell merges into.
Subtotal		Sum of the above counts of attributes for individual cells

Cell Attribute Prefix *(examples in italics)*	**Count**	**Description**
Special cell attribute prefixes		
male-only:*anat:hypodermal cell;* male-only:*brief-id:male specific neuron*		"male-only:" is prepended to a second copy of each cell attributes present at each cell that exists only in male
herm-only:*program:differentiation;* herm-only:*lineage-suffix:ppa*		"herm-only:" is prepended to a second copy of each cell attributes present at each cell that exists only in herm
daughter:*brief-id:differentiates in male;* daughter:*anat:sensory neuron;* daughter:*herm-only:anat:gonad*		"daughter:" is prepended to all cell attributes of each daughter cell, and the union of those attributes is then added to the current cell, to enable inference to be performed on how a cell's gene activity affects its daughters.
parent:*anat:somatic neuron;* parent:*cell-group:MS lineage*		"parent:" is prepended to all cell attributes of the parent cell, and they are then added to the current cell, to enable inference on how gene activity in the current cell might be affected by its parent.
ALL CELL ATTRS:		Total count, including both attributes of individual cells and versions prefixed with [parent: | daughter:][male-only: | herm-only:]

9.1.3 *Example cell attributes*

An example of the cell attributes extracted for a randomly-chosen cell, Z.papplaa, is given below. Note the parent and daughter attributes listed at the end, which (as described in the table above) are cell attributes for the parent cell and the union of attributes for the daughter cells.

```
anat:Eplaa nucleus
anat:Organ system
anat:alimentary system
anat:body region
```

```
anat:digestive tract
anat:int4V
anat:intestinal cell
anat:intestine
anat:organ
cell-group:E lineage
cell-group:all enclosing embryo cells
cell-group:int_emb
direction:a
gender:both
program:division
life-stage:1.5-fold embryo
life-stage:2-fold embryo
life-stage:3-fold embryo
life-stage:bean embryo
life-stage:comma embryo
life-stage:elongating embryo
life-stage:embryo
life-stage:enclosing embryo
life-stage:fully-elongated embryo
life-stage:gastrulating embryo
life-stage:late cleavage stage embryo
life-stage:proliferating embryo
lineage-birth-approx:220
lineage-birth:226
lineage-depth:7
lineage-prefix-canonical:Z.p
lineage-prefix-canonical:Z.pa
lineage-prefix-canonical:Z.pap
lineage-prefix-canonical:Z.papp
lineage-prefix-canonical:Z.pappl
lineage-prefix-canonical:Z.pappla
lineage-prefix:Z.p
lineage-prefix:Z.pa
lineage-prefix:Z.pap
lineage-prefix:Z.papp
lineage-prefix:Z.pappl
lineage-prefix:Z.pappla
lineage-suffix-canonical:a
lineage-suffix-canonical:aa
lineage-suffix-canonical:laa
lineage-suffix-canonical:plaa
lineage-suffix-canonical:pplaa
lineage-suffix:a
lineage-suffix:aa
lineage-suffix:laa
lineage-suffix:plaa
lineage-suffix:pplaa

parent:anat:Epla
```

```
parent:anat:Epla nucleus
parent:anat:embryonic cell
parent:cell-group:E lineage
parent:direction:a
parent:gender:both
parent:life-stage:embryo
[...truncated...]
parent:lineage-suffix:la
parent:lineage-suffix:pla
parent:lineage-suffix:ppla
parent:program:division

daughter:anat:E.plaaa
daughter:anat:E.plaaa nucleus
daughter:anat:post-embryonic cell
daughter:cell-group:int_post
daughter:cell-type:endoderm
daughter:direction:a
daughter:direction:p
daughter:gender:both
daughter:is-leaf:both
daughter:is-leaf:one-or-both
daughter:is-leaf:terminal
daughter:lineage-depth:8
daughter:lineage-prefix-canonical:Z.p
daughter:lineage-prefix-canonical:Z.pa
[...truncated...]
daughter:lineage-suffix:laap
daughter:lineage-suffix:p
daughter:lineage-suffix:plaaa
daughter:lineage-suffix:plaap
daughter:program:differentiation
```

Note also that the "daughter:direction:p" and "daughter:direction:a" attributes are both present in this cell as a result of the union operation: one daughter is the posterior cell and one daughter is the anterior cell after division. Cell attributes have to therefore be interpreted correctly according to context: inference on "direction:a" compares this attribute to "direction:p", "direction:l", "direction:r", "direction:d" and "direction:v", whereas inference on "daughter:direction:a" (which will be exactly coincident with "daughter:direction:p") compares the attribute to only prevalence of "daughter:direction:l" (which is exactly coincident with "daughter:direction:r") and "daughter:direction:d" (which is exactly coincident with "daughter:direction:v").

9.2 GENE PREDICATES

9.2.1 *Definition of gene predicate*

A *gene predicate* is a Boolean attribute of the *genotypic profile of a specific cell* (as contrasted with a cell attribute, which is a Boolean "true" flag set on a cell if a given *phenotypic* trait is observed). The genotypic profile here refers to *all genes that have been observed to be active in the given cell*[2]. They are predicates because they are used to condition likelihoods, i.e.

P(cell has attribute X | gene predicate Y is true)

Gene predicates take one of two forms:

1. **Natural predicates**, e.g.
   a) "gene X is active in the given cell"
   b) "the sequence motif M (e.g. agcgttag) is found in the genome feature F (e.g. the 3′ UTR) of *at least one gene* that is active in the given cell"
2. **Thresholded predicates**, where some statistical quantity is measured for each gene in the genotypic profile of a cell, and then the distribution of measured values is thresholded using Otsu's method (see Section 9.2.2 below), which finds the best possible approximation to bimodality for that statistic. Values below the Otsu threshold are assigned the predicate value of "false" and those above, "true". Examples of thresholded predicates include:
   a) "the genes active in cell Y have longer-than-normal introns"
   b) "the genes active in cell Y have a higher-than-normal number of SNPs in their introns"

9.2.2 *Otsu thresholding for binarization of continuous domains into predicate form*

Otsu's thresholding method [3] is frequently used in computer graphics to binarize grayscale images, i.e. to threshold the intensity of pixels to form two intensity classes. The method aims to maximize the between-class variance, or alternatively minimize the sum of within-class variances for the two classes (Figure 23). Otsu showed in his original paper that

[2] Within the limits of observed data available at WormBase.org. There will be a great deal of gene activity that has not been observed or recorded: thus recorded activity usually implies the gene is indeed active in the cell, but lack of recorded activity does not necessarily imply the gene is not active in the cell.

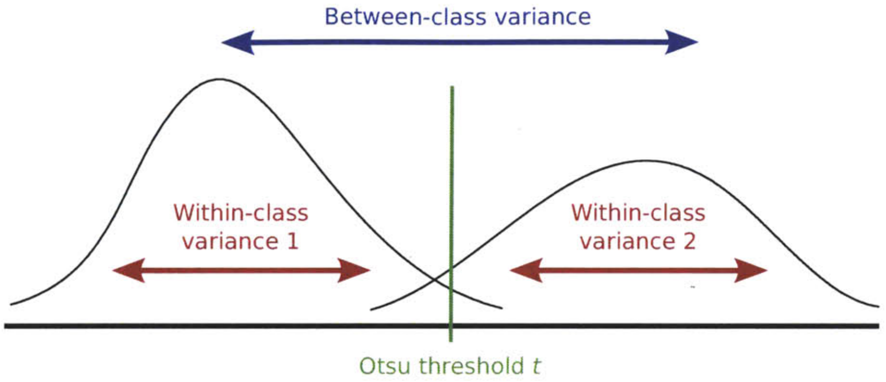


Figure 23: Otsu thresholding finds the threshold t which simultaneously (and equivalently) maximizes between-cluster variance (blue) and minimizes the sum of within-cluster variance (red).

maximizing between-class variant is are actually equivalent to minimizing the sum of within-class variances, because both of these sum to the total variance of the un-thresholded data.

Note that non-thresholded gene statistics may still be invaluable for learning about development, and a visualization tool was built to display these statistics across the cell fate pedigree and across the development timeline (see Chapter 12).

9.2.3 *List of gene predicates studied*

A number of different gene features were studied, and a large number of statistics were generated for each gene feature. Most statistics were thresholded using the Otsu thresholding algorithm, as described in Section 9.2.2, to turn them into binary predicates. A few measures were generated directly as binary predicates (notably, whether a gene was expressed in a given cell, and whether any gene active in the given cell had a specific motif in a given gene feature). In the table below, all predicates have been Otsu thresholded except where noted.

A few other things to note:

- "Gene feature" implies the upstream sequence (1kb upstream of the first 5′ UTR), a 5′ UTR, a 3′ UTR, an intron or an exon.
- "Within a gene" implies that a SNP or other genomic landmark fell within the region from the start of the 1kb upstream sequence before a gene to the end of the last 3pUTR.
- "Within a gene feature" implies that a genomic landmark fell inside the sequence of a specific gene feature, e.g. within the intron of a gene.

- Motifs are currently 8-mers, though the code supports shorter or longer motifs. (Longer motifs are generally too specific to have enough data to support inference, and are often longer than those actually employed by the cell machinery; shorter motifs are generally not biologically relevant.)

All gene predicates generated for *C. elegans* are given below, along with a short definition. (More precise definitions can be found by reading the code that extracts gene predicates.)

Gene Predicate Prefix ***(examples in italics)***	**Count**	**Definition**
content-*gc*-avg-*3pUTR-Coding_transcript*; content-*c*-avg-*intron-ncRNA*	120	Average nucleotide content of the form content-*N*-avg-*X*-*Y*, where *N* is nucleotide type (a, g, c, t, gc, agPurine), *X* is location (3pUTR, 5pUTR, intron, exon, upstreamSeq, insertion, deletion, ncRNA, CDS, SNP, tandem_repeat, binding_site, transposable_element_insertion_site) and *Y* is an extra location type parameter (Coding_transcript, Allele, miRNA, ncRNA, tandem, Mos_insertion_allele, or PicTar/miRanda for binding_site). This is an average because a certain feature may appear multiple times (e.g. there are typically multiple exons per gene).
count-*5pUTR-Coding_transcript*; count-*deletion-Allele*	21	Given the predicate count-*X*-*Y*, The number of occurrences of *X* (3pUTR, 5pUTR, intron, exon, binding_site, insertion, deletion, substitution, tandem_repeat, ...) in *Y* (Coding_transcript, Allele, wholegene, miRNA, ncRNA, ...) in the given gene.

Gene Predicate Prefix ***(examples in italics)***	**Count**	**Definition**
count-lengthWeighted-*SNP-Allele-3pUTR-Coding_transcript-int*	185	The same statistic as count-*X*-*Y* but weighted by the size of the feature X to give a sequence density, and augmented with optional location details to give count-*X*-*Y[-A-B[-C]]*, where *A* is the nearest gene feature (e.g. 3pUTR, 5pUTR, upstreamSeq, ...), *B* is the extra location information for *A* (e.g. "Coding_transcript" for an exon, as distinguished from a miRNA exon), and C is "int" or "ext" depending on whether the item being counted, *X*, is internal to or external to the closest gene feature *A* (i.e. if *X* is within *A* then "int" will be appended). Counts are generated both with and without optional features to give marginals.
density-SNP-wholegene; density-SNP-wholegene-inv	2	The density of SNPs in the whole gene (count over whole gene length), and one divided by that value.
frac-SNP-*a-exon-miRNA*; frac-SNP-*g-intron-Coding_transcript*	36	The fraction of SNPs of each nucleotide type in each gene feature, designated by frac-SNP-*N*-*X*-*Y*, where *N* is the SNP nucleotide, *X* is the gene feature (3pUTR, 5pUTR, intron, upstreamSeq etc.), and *Y* is the feature type (Coding_transcript, miRNA).

Gene Predicate Prefix *(examples in italics)*	**Count**	**Definition**

frac-SNP-missense;
frac-SNP-nonsense;
frac-SNP-splicesite;
frac-SNP-frameshift;
frac-substitution-a;
frac-substitution-c;
frac-substitution-g;
frac-substitution-t
	8	The fraction of SNPs within the gene of each of these types.

gene-expressed-*ahr-1*; gene-expressed-*C05D11.10*	1594	True if the given gene is expressed at all in a given cell or tissue. (Currently the extra information about gene expression level and certainty in the anatomy association database - "Partial", "Certain", "Uncertain" - is ignored.) Note that the count 1594 is the number for "lineage-specific-lineagePath", i.e. genes marked as active in specific cells. There are more genes active in tissue-specific associations. Not Otsu-thresholded as gene expression is binary as used here.
len-avg-*intron-Coding_transcript*; len-avg-*5pUTR-Coding_transcript*; len-avg-*binding_site-miRanda*	20	Of the form len-avg-*X*-*Y*. The average length of gene feature *X* type *Y* within the gene.
len-wholegene	1	The total length of the gene in base pairs (from TSS to the end of the final 3pUTR).

Gene Predicate Prefix *(examples in italics)*	**Count**	**Definition**
motif-density-avg-*intron-acgctatg*; motif-density-avg-*upstreamSeq-gcgctagc*	297,484	Of the form motif-density-avg-*X*-*M*. The average density of the motif *M* in the gene feature *X*, i.e. the count of motifs in the gene feature divided by the length of the gene feature. Gene features are 5pUTR, 3pUTR, intron, exon and upstreamSeq. Note that only 297484 out of a potential $4^8 \times 5 = 327680$ gene predicates (for 8-mers with five gene features) are actually present for the lineage-specific-lineagePath dataset; other data sets have different total counts (e.g. tissue-specific will have a higher count due to the greater number of available gene associations). This statistic is Otsu-thresholded as opposed to the "motif-present" statistic below.
motif-present-*3pUTR-ctgctggc*; motif-present-*exon-agtagacc*	297,484	Same as motif-density, except this predicate is not Otsu-thresholded, but the predicate is true if the total count of motifs in the given gene feature is greater than zero, i.e. if the motif is present at all.
repeat-avgScore-tandem_repeat-tandem	1	The average repeat score for tandem repeats in the gene (the score is higher for higher repeat counts and longer repeat units)

Gene Predicate Prefix ***(examples in italics)***	**Count**	**Definition**

repeat-*lengthWeighted-tandem_repeat-tandem-3pUTR-Coding_transcript-ext*;
repeat-*numRepeats-tandem_repeat-tandem-exon-miRNA-ext*;
repeat-*repeatLen-tandem_repeat-tandem-CDS-Coding_transcript-int*
	41	Statistics of tandem repeats in the gene, of the form repeat-*T*-tandem_repeat-tandem*[-A-B[-C]]*, where *T* is one of lengthWeighted (density of microsatellite repeat regions in gene feature, i.e. number of microsatellites divided by length of gene feature), numRepeats (number of repeat units) or repeatLen (length of each repeat unit), *A* and *B* designate the repeat location, and *C* is "int" or "ext" as described above for count-*X-Y[-A-B[-C]]*.

rnaBindingSite-anywhere-*let-7*; rnaBindingSite-anywhere-mir-*239a*; rnaBindingSite-anywhere-*mir-87*	117	Count of number of locations the given microRNA binds within any of the gene's features
rnaBindingSite-binding_site-*PicTar-3pUTR-Coding_transcript-ext-lsy-6*; rnaBindingSite-binding_site-*miRanda-3pUTR-Coding_transcript-mir-34*	*1144*	Binding sites as predicted by the PicTar and miRanda algorithms. Of the form rnaBindingSite-binding_site-*R-A-B[-C]*, for miRNA prediction method *R* (PicTar or miRanda) at gene feature *A* type *B*, optionally with information on whether the binding site was within ("int") or outside and just near ("ext") the gene feature.
rnaTotBindingSites-binding_site-*PicTar-3pUTR-Coding_transcript-ext*; rnaTotBindingSites-binding_site-*miRanda-exon-Coding_transcript-int*	26	Total number of binding sites of any RNA within the given gene feature, of form rnaTotBindingSites-binding_site-*R-A-B-C*, with fields as described above.

Gene Predicate Prefix ***(examples in italics)***	**Count**	**Definition**
ALL GENE PREDICATES:	598,283	(Note: this number, and many of the above, are specifically for the "lineage-specific-lineagePath" dataset)

9.3 LIST OF GENES STUDIED

There are over 20,000 genes recorded for *C. elegans* in WormBase.org (see Section 2.2). Of those, genes that were not associated with at least 10 cells were discarded to avoid problems with small-count statistics. There were 3244 genes in the "gene association database" that were recorded as active in at least that minimum number of cells. However depending on whether looking at "lineage-specific" or "tissue-specific" associations, and depending on whether looking at "lineagePath" or "canonicalPath" cell identities, the number of remaining valid genes varied dramatically:

Dataset	**Number of valid genes**
lineage-specific-lineagePath	298
lineage-specific-canonicalPath	349
tissue-specific-lineagePath	3227
tissue-specific-canonicalPath	3223

Taking the union of all genes found to pass the threshold minimum number of associations in any of the four datasets yields a list of 3244 unique genes, shown below. It is worth actually listing the name of each gene here, because these are the only genes that were actually able to be analyzed in this research. For all other *C. elegans* genes, there was insufficient data in WormBase.org (or the gene was too rarely expressed) to reliably test for correlations with cell phenotype.

AC7.1, AH6.3, B0024.10, B0034.1, B0035.1, B0222.3, B0228.7, B0284.1, B0285.3, B0303.2, B0303.3, B0334.4, B0336.3, B0336.7, B0361.6, B0393.6, B0416.5, B0495.9, B0507.1, B0511.6, B0513.9, C01B10.3, C01B12.3, C01F1.6, C01G10.7, C01G5.8, C01G6.4, C01H6.4, C01H6.9, C02C2.4, C02F12.10, C02F12.5, C03A3.2, C03H5.2, C04B4.2, C04C3.3, C04F12.5, C04F5.8, C05C10.3, C05C8.1, C05C8.2, C05C8.6, C05D11.10, C05D11.5, C05D11.7, C05D2.6, C05G5.1, C06A8.1, C06C3.4, C06G1.5, C06G3.5, C06G8.1, C07E3.5, C07E3.6, C07H6.2, C07H6.3, C08B11.3, C08B6.8, C08E3.13, C08F11.1, C08G9.2, C09E7.8, C09E9.1, C09F9.3, C09G12.1, C10E2.6, C10G11.7, C10G8.8, C11D2.4, C11D9.1, C11E4.1, C11E4.6, C11H1.3, C12D8.1, C13B9.3, C13C4.4, C13C4.5, C13F10.4, C13F10.7, C14A4.11, C14C10.5, C15C7.1, C16A3.10, C16C10.1, C16C10.4, C17E4.3, C17G10.1, C17G10.2, C17G10.9, C17H11.6, C17H12.13, C17H12.2, C18A3.5, C18B12.6, C18B2.3, C18E9.6, C18F10.2, C18F10.7, C23G10.10, C23G10.8, C23H4.6, C24A1.2, C24A1.3, C24A8.1, C24B5.4, C24D10.1, C25A1.5, C25A8.4, C25E10.12, C25G4.10, C25H3.9, C26E1.3, C26E6.3, C27C12.4, C27D6.4, C27D8.4, C27F2.10, C27H5.3, C28C12.4, C28H8.11, C29F5.1, C29F7.3, C29G2.1, C29H12.2, C30F12.1, C30F12.2, C30F12.6, C30H7.2, C31C9.2, C32B5.6, C32E8.9, C32F10.8, C33A12.1, C33A12.4, C33D12.2, C33E10.10, C33H5.18, C34B2.5, C34B2.6, C34B7.2, C34C6.4, C34D1.1, C34D1.2, C34D4.1, C34F11.3, C34G6.1, C35D10.12, C35E7.4, C36C9.1, C36E8.4, C37C3.2, C37C3.3, C37C3.7, C37E2.1, C37H5.6, C38C10.2, C39F7.1, C40H1.6, C41C4.8, C42C1.4, C42D4.3, C44C1.2, C44E4.4, C44H4.1, C44H4.4, C45G9.13, C45G9.5, C46C11.1, C46E10.8, C46F11.2, C46H11.6, C48B6.2, C49C3.5, C49F8.2, C49H3.6, C50C3.5, C50D2.2, C50D2.7, C50F2.3, C50F2.8, C50F4.14, C50F4.4, C50F7.4, C50H2.6, C52E12.4, C53A3.2, C54D10.10, C55C3.5, C55H1.1, C56G2.3, C56G2.4, C56G2.7, C56G2.9, CC4.2, CD4.4, D1005.3, D1014.5, D1046.5, D1054.1, D2007.1, D2007.2, D2007.3, D2007.4, D2030.7, D2045.9, D2085.3, D2085.5, D2096.11, D2096.6, DH11.1, E01A2.6, E02D9.1, E04D5.1, EEED8.6, EGAP798.1, F01D5.9, F01F1.15, F01F1.2, F01G4.6, F07A11.2, F07A11.4, F07F6.4, F08B6.3, F08C6.2, F08D12.7, F08F3.9, F08G12.1, F08G5.6, F08H9.3, F08H9.4, F09C3.2, F09D1.1, F09E10.6, F09E5.3, F09F7.7, F09G2.8, F10A3.4, F10D7.2, F10E7.5, F10E7.9, F10E9.11, F10F2.4, F11A10.2, F11A10.5, F11C1.5, F11F1.1, F13A2.4, F13B12.1, F13D12.3, F13E6.1, F13H10.1, F14B4.3, F15B10.1, F15C11.2, F15D3.7, F16A11.1, F16A11.2, F16B4.8, F16D3.4, F17A9.3, F18F11.1, F19B6.1, F20D1.9, F20D12.2, F20D6.11, F20H11.4, F20H11.5, F21D12.5, F21D5.2, F21D5.9, F21G4.1, F22B7.9, F22D6.2, F22F7.1, F22G12.5, F23B12.5, F23C8.6, F23F1.5, F23F12.8, F23H11.4, F23H11.5, F25B3.6, F25B4.1, F25B4.2, F25B5.2, F25B5.6, F25H2.5, F26A1.14, F26A1.6, F26A1.7, F26A1.8, F26A3.4, F26A3.5, F26E4.12, F26F4.6, F26H9.2, F26H9.5, F27D4.4, F27D4.6, F28A10.1, F28C6.1, F28C6.2, F28D1.2, F28F5.6, F28F8.5, F29G6.2, F31C3.3, F31E3.4, F32B6.2, F32B6.9, F32D8.15, F33H1.4, F33H2.3, F34D6.2, F34H10.4, F35A5.4, F35D11.3, F35F11.1, F35G2.1, F35H12.4, F36A2.8, F36D4.5, F36F2.1, F36H2.3, F37A4.6, F37B4.10, F37C12.1, F37H8.5, F38A1.5, F38A1.8, F38A5.1, F38A5.2, F38E9.1, F38E9.5, F39B2.1, F39B2.8, F39H11.1, F40E10.6, F40E3.2, F40F11.2, F40F8.1, F40F8.8, F40F9.10, F40F9.6, F40G9.2, F41C3.5, F41C3.8, F41E6.5, F41E7.1, F41E7.3, F41E7.6, F41E7.9, F41G4.1, F42A10.3, F42A8.1, F42D1.2, F42G9.1, F43C1.1, F43D9.1, F43E2.7, F43G9.1, F43G9.3, F44A2.3, F44A2.5, F44D12.1, F44D12.4, F44E7.2, F44E7.4, F44F4.1, F45C12.15, F45C12.2, F45C12.3, F45D11.14, F45F2.9, F46C5.6, F46F2.3, F46F6.2, F47B10.2, F47B7.2, F47B8.3, F47B8.8, F47D12.6, F47G4.6, F47G9.1, F47H4.1, F48B9.5, F48G7.10, F49C12.11, F49E2.5, F49H12.4, F52A8.6, F52C12.4, F52D10.2, F52E1.13, F52H3.5, F53A10.2, F53A9.4, F53B1.2, F53B3.3, F53B6.2, F53C11.5, F53C3.2, F53E10.1, F53E4.1, F53H1.4, F53H2.3, F53H4.5, F53H8.3, F54A5.1, F54C8.4, F54C8.7, F54C9.11, F54C9.7, F54D7.2, F54D8.6, F54F2.9, F54G2.1, F55A12.8, F55A4.1, F55C12.4, F55C5.8, F55G1.9, F55G11.7, F55H12.4, F56B6.6, F56C11.3, F56C9.10, F56D1.3, F57B10.9, F57C9.4, F58A4.5, F58B3.4, F58B4.3, F58E10.3, F58H1.2, F59A3.1, F59B10.3, F59B2.12, F59B2.13, F59C6.2, F59F4.1, F59F4.2, F59G1.4, F59H5.1, F59H6.6, H01G02.2, H03A11.2, H05L14.2, H06H21.3, H10E21.1, H14E04.1, H17B01.1, H19N07.1, H27A22.1, H28O16.1, H37N21.1, H38K22.4, H43I07.3, JC8.3, JC8.5, K01A2.5, K02D10.1, K02G10.1, K02G10.3, K03A1.4, K04F10.2, K05D4.9, K06H7.8, K07C11.10, K07C11.4, K07E3.4, K07G5.3, K07H8.3, K08B12.1, K08C7.4, K08D8.5, K08E3.5, K08E7.1, K08E7.8, K08F11.5, K08F8.1, K09A9.6, K09C8.2, K09H11.1, K10B2.4, K10B3.1, K10C2.4, K10C3.4, K10C8.3, K10D6.2, K10G6.4, K11C4.2, K11G9.2, K12B6.4, K12H4.5, K12H6.12, M01A10.3, M01E11.2, M01F1.3, M01F1.5, M01G5.3, M02B7.5, M02G9.1, M03B6.2, M03C11.3, M03D4.4, M03E7.1, M03F8.3, M04F3.4, M05B5.2, M106.3, M142.1, M176.2, M176.5, M18.1, M18.8, M6.1, M6.3, M60.6, M7.1, M70.5, M79.4, M88.3, M88.5, R01E6.5, R02D3.7, R02E12.4, R02E4.1, R02F2.1, R03D7.1, R03D7.4, R03E1.4, R04A9.5, R04B5.6, R04F11.5, R05D3.2, R05F9.1, R05F9.6, R05G6.1, R05G6.7, R05H10.2, R06F6.6, R07B1.9, R07E5.4, R07E5.7, R07G3.8, R07H5.8, R08D7.5, R09A8.5, R09F10.5, R102.2, R10F2.5, R11.3, R11A8.7, R11F4.1, R12C12.1, R12C12.6, R12E2.1, R12E2.13, R12E2.7, R13.2, R13A5.10, R13A5.9, R13F6.2, R13H4.5, R13H7.2, R144.10, R144.11, R151.2, R166.2, R53.3, R53.5, R74.6, R74.7, R90.5, T01C8.5, T01D3.1, T01G1.2, T01H3.3, T02E9.3, T03F6.3, T04A6.1, T04A6.2, T04A8.11, T04A8.6, T04B2.5, T04B8.5, T04C9.1, T04H1.2, T05E11.3, T05H10.1, T05H10.3, T06G6.5, T07C12.9, T07F10.1, T07F8.4, T07H6.5, T08A11.1, T08B2.11, T08B2.7, T08G11.1, T09A12.2, T09B4.8, T09B4.9, T09F3.2, T10B11.2, T10B11.6, T10B5.2, T10C6.5, T10F2.2, T11A5.6, T12D8.6, T12D8.8, T12D8.9, T12G3.1, T13C5.4, T13H5.6, T14B1.1, T14E8.1, T17H7.1, T18D3.7, T19B10.2, T19B4.1, T19D7.4, T20G5.10, T20H4.2, T21D12.9, T21H8.1, T22B11.5, T22C1.6, T22D1.3, T22E7.2, T22H9.4, T23B12.6, T23C6.5, T23G5.2, T23H2.4, T24H10.1, T26A5.5, T26A5.8, T26C12.1, T26E3.4, T27A3.1, T27A8.2, T27E4.3, T27E4.8, T27F2.4, T27F6.6, T27F7.3, T28B8.1, T28C12.4, T28C12.5, T28C6.7, T28D6.4, T28D6.7, T28D9.1, VC27A7L.1, W01A11.1, W01A11.2, W01F3.1, W01G7.4, W02B12.10, W02B12.11, W02B12.9, W02B3.4, W02D9.3, W03F8.6, W03F9.10, W04B5.5, W04G3.1, W04G3.5, W05B10.4, W05B2.4, W05H12.1, W06H8.6, W08E12.7, W08F4.3, W09C5.1, W09G3.2, W09H1.5, Y102A11A.2, Y102A11A.6, Y105C5A.15, Y105E8A.1, Y106G6H.4, Y108G3AL.2, Y110A2AL.13, Y110A7A.20, Y110A7A.9, Y111B2A.8, Y113G7B.16, Y116A8C.3, Y119C1B.5, Y11D7A.8, Y18D10A.1, Y19D10A.10, Y23B4A.2, Y25C1A.5, Y32G9A.6, Y32H12A.1, Y32H12A.8, Y34D9A.1, Y37A1B.5, Y37D8A.17, Y37D8A.22, Y37E11B.5, Y38E10A.2, Y38F1A.6, Y39A1A.1, Y39A1C.1, Y39A3CR.3, Y39G10AR.8, Y40B1B.7, Y41C4A.9, Y41D4A.4, Y41G9A.3, Y42G9A.3, Y42G9A.4, Y43F4A.1, Y43F4B.7, Y44E3A.3, Y45G12C.10, Y46G5A.4, Y46H3A.2, Y47A7.1, Y47G6A.7, Y48B6A.6, Y48C3A.1, Y48C3A.14, Y48C3A.16, Y48E1B.2, Y48G10A.1, Y49E10.20, Y49G5B.1, Y4C6B.5, Y50D7A.4, Y50D7A.8, Y53C10A.3, Y53C10A.5, Y53C12B.1, Y53C12C.1, Y53G8AL.2, Y54E10BR.4, Y54E5A.1, Y54E5A.7, Y54F10AL.1, Y54G11A.2, Y54G11A.7, Y54G2A.17, Y54G2A.18, Y54G2A.4, Y55D5A.1, Y55D9A.2, Y55F3AM.3, Y56A3A.18, Y56A3A.19, Y57G11C.20, Y57G11C.37, Y58A7A.4, Y59C2A.2, Y5F2A.4, Y60A3A.9, Y61A9LA.1, Y61A9LA.10, Y62E10A.13, Y66A7A.5, Y66D12A.7, Y66H1B.2, Y67D8A.3, Y69A2AR.18, Y6B3B.1, Y70G10A.3, Y71F9AL.14, Y71F9AL.17, Y71F9AL.9, Y71G12B.16, Y71H10A.1, Y71H10B.1, Y73B6BL.30, Y73B6BL.4, Y73C8B.1, Y73E7A.3, Y73F4A.2, Y73F8A.24, Y73F8A.27, Y73F8A.5, Y75B8A.7, Y76A2B.5, Y76A2B.6, Y82E9BR.14, Y94H6A.8, Y97E10AR.2, Y97E10AR.7, ZC116.3, ZC123.1, ZC123.3, ZC13.1, ZC155.2, ZC204.12, ZC204.2, ZC21.3, ZC317.1, ZC328.3, ZC373.1, ZC373.5, ZC376.4, ZC395.10, ZC506.1, ZC84.3, ZK1067.4, ZK1098.7, ZK112.3, ZK177.8, ZK262.2, ZK287.1, ZK337.1, ZK337.2, ZK370.4, ZK384.3, ZK484.4, ZK512.2, ZK6.4, ZK632.12, ZK632.4, ZK643.1, ZK643.2, ZK643.5, ZK652.8, ZK686.2, ZK688.2, ZK757.4, ZK792.1, ZK792.4, ZK792.7, ZK795.3, ZK816.4, ZK822.5, ZK829.4, ZK829.7, ZK856.14, ZK856.8, ZK899.2, ZK938.2, ZK973.1, ZK973.11, aak-2, aakb-1, aap-1, aat-1, aat-9, abcf-1, abcf-2, abch-1, abf-1, abf-2, abi-1, abl-1, abt-4, abt-5, abts-1, abts-2, abts-3, abts-4, abu-1, acbp-6, acc-4, acdh-10, acdh-11, acdh-3, acdh-7, ace-1, ace-2, ace-3, ace-4, acl-14, acl-2, acl-4, acl-5, acl-8, acn-1, aco-1, acr-12, acr-14, acr-15, acr-16, acr-5, acr-8, acs-11, acs-13, acs-19, acs-2, acs-20, acs-4, acs-5, act-1, act-2, act-3, act-4, act-5, acy-1, acy-2, adbp-1, add-1, adm-4, adr-1, adt-1, aex-1, aex-2, aex-3, aex-6, aff-1, age-1, agr-1, ags-3, aha-1, ahcy-1, ahr-1, aip-1, akt-1, akt-2, aldo-1, alh-1, alh-10, alh-6, alh-7, alp-1, alr-1, alx-1, aman-1, aman-2, amph-1, amt-3, ant-1.1, ant-1.2, ant-1.4, apa-2, apb-1, apb-3, apl-1, apm-1, apm-3, apr-1, aps-2, aptf-1, apx-1, apy-1, aqp-1, aqp-2, aqp-3, aqp-7, ard-1, arf-1.2, arf-3, arf-6, arg-1, ari-1, arl-1, arl-13, arl-3, arl-6, arl-8, arr-1, arx-1, arx-4, arx-7, asd-1, asd-2, asm-1, asna-1, ast-1, atf-2, atf-6, atg-13, atg-18, atg-2, atg-3, atg-4.1, atl-1, atn-1, atp-2, atp-3, atx-2, atx-3, avr-14, avr-15, bam-2, bar-1, bas-1, bath-15, bath-42, bath-44, bath-47, bbs-1, bbs-2, bbs-5, bbs-8, bcat-1, bec-1, bir-1, bli-3, bli-4, bli-5, blmp-1, bra-1, bre-1, bre-3, bro-1, cah-3, cal-4, calu-1, cam-1, cap-1, cap-2, car-1, cash-1, casy-1, cat-1, cat-2, cat-4, catp-1, cav-1, cav-2, cbp-1, cca-1, ccb-1, ccch-1, ccch-2, ccch-3, ccdc-47, ccdc-55, ccr-4, cct-1, cct-2, cct-4, cct-7, cdc-42, cdd-1, cdf-1, cdh-3, cdh-4, cdk-4, cdo-1, cdr-2, cdr-6, cdtl-7, ced-1, ced-10, ced-11, ced-12, ced-4, ced-9, ceh-10, ceh-13, ceh-14, ceh-16, ceh-17, ceh-18, ceh-19, ceh-2, ceh-20, ceh-22, ceh-23, ceh-24, ceh-26, ceh-27, ceh-28, ceh-30, ceh-31, ceh-32, ceh-33, ceh-34, ceh-36, ceh-37, ceh-38, ceh-39, ceh-41, ceh-43, ceh-44, ceh-45, ceh-48, ceh-49, ceh-5, ceh-51, ceh-6, ceh-60, ceh-7, ceh-9, cel-1, cep-1, ces-2, cex-1, cex-2, cey-1, cey-2, cey-3, cfi-1, cfim-1, cfim-2, cfz-2, cgh-1, cgt-1, cgt-3, cha-1, chd-3, che-11, che-12, che-13, che-14, che-2, che-3, chin-1, chn-1, cho-1, chs-2, cht-1, cids-1, cit-1.2, ckb-2, cki-1, cki-2, clc-1, clc-2, cle-1, clec-197, clec-38, clec-42, clec-5, clec-52, clec-78, clh-1, clh-2, clh-3, clh-6, clic-1, cln-3.1, cln-3.2, clp-1, cmk-1, cnb-1, cnc-2, cnc-4, cnd-1, cng-1, cnp-3, cnt-2, cnx-1, cog-1, cogc-3, cogc-5, col-121, col-130, col-144, col-19, col-43, col-89, coq-8, cpb-3, cpi-2, cpl-1, cpn-3, cpna-2, cpna-5, cpr-3, cpsf-2, cpsf-3, cpt-1, cpt-6, cpz-1, cra-1, cri-2, crm-1, crml-1, crn-2, crt-1, csb-1, csk-1, csn-2, csn-5, csq-1, cst-1, cth-2, ctl-3, cts-1, cua-1, cuc-1, cul-1, cul-2, cul-3, cul-5, cup-5, cut-6, cuti-1, cwn-1, cwn-2, cwp-1, cwp-2, cwp-3, cwp-4, cyd-1, cye-1, cyk-3, cyk-4, cyn-4, cyn-5, cyn-8, cyn-9, cyp-14A3, cyp-29A2, dab-1, dac-1, daf-1, daf-12, daf-14, daf-15, daf-18, daf-19, daf-2, daf-22, daf-28, daf-3, daf-36, daf-4, daf-5, daf-6, daf-9, dao-3, dao-5, dapk-1, daz-1, dbl-1, dcap-1, dcp-66, dct-13, dct-6, ddb-1, ddr-2, ddx-19, deb-1, deg-1, del-1, del-2, del-3, del-4, dep-1, des-2, dgk-1, dgk-3, dgk-4, dgn-1, dhc-1, dhp-1, dhs-19, dhs-21, dhs-23, dhs-31, dhs-5, dhs-9, die-1, dif-1, dig-1, din-1, dis-3, div-1, djr-1.2, dkf-1, dkf-2, dlc-1, dlg-1, dmd-3, dmd-4, dmd-5, dmd-7, dna-2, dnc-2, dnj-10, dnj-12, dnj-20, dnj-25, dnj-27, dnj-28, dnj-4, dnj-7, dnj-9, dod-22, dog-1, dop-1, dop-2, dop-3, dop-4, dpf-2, dpf-6, dpl-1, dpy-11, dpy-14, dpy-18, dpy-19, dpy-2, dpy-22, dpy-23, dpy-31, dpy-5, dpy-7, dre-1, drn-1, drp-1, dsc-1, dsh-1, dsh-2, dsl-1, dsl-2, dss-1, dyb-1, dyc-1, dyci-1, dyf-1, dyf-11, dyf-13, dyf-2, dyf-3, dyf-5, dyf-6, dyf-8, dylt-2, dyn-1, dys-1, eat-16, eat-20, eat-4, eat-5, eat-6, ech-5, ect-2, eel-1, eff-1, efl-1, efn-2, efn-4, eft-3, egl-10, egl-13, egl-15, egl-17, egl-18, egl-19, egl-2, egl-20, egl-21, egl-26, egl-3, egl-36, egl-38, egl-4, egl-44, egl-45, egl-46, egl-47, egl-5, egl-6, egl-8, eif-3.B, eif-3.D, eif-3.E, eif-3.H, eif-6, elo-1, elo-2, elo-3, elo-5, elo-6, elpc-1, elpc-3, elt-2, elt-3, elt-6, emb-5, emb-9, emr-1, eng-1, enol-1, ent-1, ent-2, eor-1, eor-2, epc-1, epn-1, eps-8, erd-2, eri-1, eri-6, eri-7, erm-1, erp-1, erv-46, esyt-2, etr-1, ets-4, ets-5, eva-1, evl-20, exc-4, exc-5, exc-7, exl-1, exos-1, exos-7, exp-2, eya-1, far-3, far-6, fat-2, fat-3, fat-5, fat-6, fax-1, fbl-1, fbp-1, fbxa-57, fce-1, feh-1, fhod-1, fkb-3, fkb-5, fkb-6, fkh-10, fkh-2, fkh-5, fkh-6, fkh-7, fkh-8, fkh-9, flh-2, fli-1, flp-1, flp-10, flp-11, flp-12, flp-13, flp-15, flp-17, flp-18, flp-20, flp-21, flp-22, flp-25, flp-3, flp-4, flp-5, flp-6, flp-7, flp-8, flt-1, fmo-1, fmo-3, fmo-4, folt-1, fos-1, fox-1, fozi-1, fpn-1.2, fre-1, frh-1, frk-1, frl-1, frm-2, frm-4, frm-8, fshr-1, fsn-1, ftn-2, ftt-2, fut-1, fzo-1, gale-1, gana-1, gap-2, gar-1, gar-2, gar-3, gas-1, gck-2, gck-3, gck-4, gcs-1, gcy-11, gcy-18, gcy-28, gcy-35, gcy-5, gei-1, gei-16, gei-18, gei-3, gem-4, ger-1, ggr-1, ggr-2, ggr-3, ggtb-1, git-1, gla-3, glb-1, glb-10, glb-11, glb-12, glb-13, glb-14, glb-16, glb-17, glb-18, glb-19, glb-2, glb-20, glb-21, glb-22, glb-23, glb-25, glb-27, glb-28, glb-29, glb-3, glb-30, glb-31, glb-32, glb-33, glb-4, glb-5, glb-6, glb-7, glb-9, glc-3, glc-4, gld-1, gld-2, glf-1, glh-1, glh-2, glo-4, glp-1, glr-1, glr-2, glr-4, glr-5, glr-8, glrx-10, glrx-21, glt-1, glt-4, glt-5, gly-1, gly-12, gly-13, gly-18, gly-2, gly-3, gna-1, gna-2, gnrr-1, goa-1, gob-1, gon-1, gon-14, gon-2, gon-4, gosr-1, gpa-1, gpa-10, gpa-12, gpa-13, gpa-14, gpa-15, gpa-16, gpa-2, gpa-3, gpa-6, gpa-7, gpa-8, gpa-9, gpb-1, gpb-2, gpc-1, gpc-2, gpdh-1, gpdh-2, gpi-1, gpn-1, grd-12, grd-13, grd-14, grd-16, grd-3, grd-5, grd-6, grd-7, grd-8, grh-1, grk-2, grl-1, grl-10, grl-11, grl-14, grl-15, grl-17, grl-2, grl-21, grl-25, grl-27, grl-29, grl-3, grl-4, grl-5, grl-6, grl-7, grl-8, grl-9, grp-1, gsa-1, gsnl-1, gsp-1, gsp-2, gspd-1, gst-10, gst-2, gst-24, gst-30, gst-33, gst-37, gst-38, gst-7, gsto-2, gta-1, gtl-1, gtl-2, gyg-1, haf-2, haf-6, haf-7, ham-2, har-1, hbl-1, hch-1, hcp-3, hda-1, hda-4, hel-1, hen-1, hex-1, hex-2, hex-5, hgo-1, hgrs-1, hid-1, him-17, him-4, hipr-1, his-13, his-2, his-4, his-44, his-48, his-61, his-64, hke-4.1, hlh-1, hlh-10, hlh-17, hlh-2, hlh-29, hlh-3, hlh-30, hlh-34, hlh-8, hmg-11, hmg-4, hmg-5, hmp-1, hmr-1, hnd-1, hot-4, hpd-1, hpl-1, hsb-1, hse-5, hsp-1, hsp-16.2, hsp-25, hsp-4, hsp-43, hsp-6, hst-2, hst-6, htp-3, hum-1, hum-5, hyl-1, icd-1, ida-1, idh-1, ifa-1, ifa-3, ifa-4, ifb-1, ifc-1, ife-1, iff-1, iff-2, ift-81, ifta-1, ifta-2, igcm-1, igcm-2, ima-3, imp-1, ina-1, inf-1, inft-1, ing-3, ins-1, ins-11, ins-17, ins-18, ins-2, ins-21, ins-22, ins-23, ins-3, ins-31, ins-4, ins-5, ins-6, ins-7, ins-8, ins-9, inx-1, inx-10, inx-11, inx-12, inx-13, inx-14, inx-18, inx-19, inx-2, inx-20, inx-3, inx-4, inx-5, inx-6, inx-7, inx-8, inx-9, ipla-1, irs-1, irx-1, isp-1, itr-1, itsn-1, ivd-1, jac-1, jkk-1, jmjd-2, jnk-1, jph-1, jud-4, jun-1, kal-1, kap-1, kat-1, kca-1, kcnl-2, kel-3, ketn-1, kin-1, kin-15, kin-16, kin-18, kin-19, kin-29, kin-32, kin-5, klc-2, klp-11, klp-12, klp-13, klp-15, klp-16, klp-20, klp-4, klp-6, klp-8, kqt-1, kqt-3, kri-1, krs-1, kvs-1, lad-2, lag-2, lagr-1, lam-1, lam-3, lap-1, larp-1, lat-1, lbp-1, lbp-5, ldb-1, ldh-1, let-19, let-2, let-381, let-413, let-418, let-502, let-60, let-607, let-653, let-7, let-711, let-721, let-75, let-754, let-756, let-767, let-805, let-92, lev-1, lev-11, lev-8, lfe-2, lgc-34, lgc-38, lgc-46, lgc-55, lgg-1, lig-4, lim-4, lim-6, lim-7, lim-8, lim-9, lin-1, lin-10, lin-11, lin-12, lin-13, lin-14, lin-17, lin-18, lin-22, lin-24, lin-25, lin-26, lin-28, lin-29, lin-3, lin-31, lin-35, lin-36, lin-39, lin-4, lin-40, lin-41, lin-42, lin-44, lin-45, lin-48, lin-52, lin-53, lin-58, lin-66, lip-1, lir-1, lir-3, lis-1, lit-1, lmp-1, lnp-1, lon-1, lon-2, lon-3, lon-8, lov-1, lpd-2, lpin-1, lpr-1, lpr-4, lrk-1, lrp-1, lrp-2, lrs-2, lsm-6, lst-1, lsy-22, lsy-6, ltd-1, lys-1, lys-7, lys-8, maa-1, mab-20, mab-21, mab-23, mab-3, mab-5, mab-7, mab-9, mag-1, magi-1, mai-2, mau-8, max-1, max-2, mbf-1, mbk-2, mboa-1, mboa-6, mboa-7, mbr-1, mca-2, mca-3, mce-1, mcm-2, mcm-7, mdh-1, mdl-1, mdt-15, mdt-6, mdt-8, mec-1, mec-12, mec-2, mec-3, mec-5, mec-6, mec-8, mec-9, med-1, med-2, mek-1, mel-11, mel-32, mep-1, mes-3, mes-6, mex-1, mex-5, mfb-1, mgl-1, mgl-2, mgl-3, mif-2, mif-3, mig-1, mig-10, mig-13, mig-15, mig-17, mig-2, mig-22, mig-23, mig-5, mig-6, mir-1, mir-2, mir-228, mir-230, mir-231, mir-234, mir-235, mir-236, mir-237, mir-238, mir-240, mir-241, mir-242, mir-243, mir-245, mir-247, mir-249, mir-250, mir-251, mir-252, mir-255, mir-265, mir-266, mir-268, mir-269, mir-270, mir-35, mir-359, mir-360, mir-392, mir-41, mir-42, mir-44, mir-45, mir-46, mir-47, mir-51, mir-52, mir-53, mir-57, mir-59, mir-61, mir-71, mir-72, mir-76, mir-784, mir-785, mir-786, mir-788, mir-79, mir-793, mir-797, mir-80, mir-81, mir-82, mir-83, mir-84, mir-90, miz-1, mkk-4, mlc-1, mlc-2, mlc-3, mlk-1, mlp-1, mls-1, mls-2, mlt-10, mlt-11, mlt-8, mlt-9, mml-1, mnm-2, moc-1, mod-1, moe-3, mom-2, mom-5, mpk-1, mps-1, mps-4, mpz-1, mrg-1, mrp-1, mrp-2, mrp-5, mrp-6, mrp-7, msp-10, msp-113, msp-142, msp-152, msp-19, msp-3, msp-31, msp-32, msp-33, msp-36, msp-38, msp-40, msp-45, msp-49, msp-50, msp-51, msp-53, msp-55, msp-56, msp-57, msp-59, msp-63, msp-64, msp-65, msp-74, msp-76, msp-77, msp-78, msp-79, msp-81, mtl-1, mtm-1, mtm-3, mtm-5, mtr-4, mtss-1, mua-3, mua-6, mup-4, mut-7, mxl-1, mxl-2, myo-2, myo-3, myo-5, nab-1, nac-1, nac-3, nas-1, nas-11, nas-14, nas-19, nas-22, nas-25, nas-27, nas-30, nas-31, nas-36, nas-37, nas-38, nas-39, nas-4, nas-7, nas-9, nasp-2, nca-1, ncam-1, nck-1, ncl-1, ncr-1, ncr-2, ncs-1, ndg-4, nduf-7, ndx-1, nekl-3, nep-1, nex-1, nex-2, nex-4, nfi-1, nfm-1, nft-1, nfya-2, ngp-1, nhl-1, nhl-2, nhr-1, nhr-108, nhr-112, nhr-114, nhr-119, nhr-123, nhr-131, nhr-132, nhr-133, nhr-137, nhr-148, nhr-15, nhr-152, nhr-156, nhr-166, nhr-168, nhr-17, nhr-178, nhr-179, nhr-181, nhr-183, nhr-185, nhr-187, nhr-188, nhr-196, nhr-2, nhr-20, nhr-206, nhr-207, nhr-208, nhr-213, nhr-214, nhr-219, nhr-22, nhr-228, nhr-23, nhr-231, nhr-235, nhr-237, nhr-25, nhr-258, nhr-271, nhr-28, nhr-3, nhr-30, nhr-37, nhr-4, nhr-40, nhr-41, nhr-42, nhr-43, nhr-44, nhr-47, nhr-48, nhr-50, nhr-51, nhr-52, nhr-54, nhr-56, nhr-57, nhr-58, nhr-62, nhr-63, nhr-64, nhr-65, nhr-66, nhr-67, nhr-68, nhr-69, nhr-71, nhr-76, nhr-79, nhr-80, nhr-83, nhr-84, nhr-85, nhr-90, nhr-91, nhr-98, nhx-1, nhx-3, nhx-5, nhx-8, nid-1, nipi-3, nlg-1, nlp-1, nlp-10, nlp-11, nlp-12, nlp-13, nlp-14, nlp-16, nlp-18, nlp-19, nlp-2, nlp-20, nlp-21, nlp-23, nlp-24, nlp-26, nlp-27, nlp-29, nlp-3, nlp-30, nlp-31, nlp-32, nlp-5, nlp-6, nlp-7, nlp-8, nlp-9, nlr-1, nmr-1, nmt-1, nmy-1, nmy-2, nnt-1, nob-1, nol-6, nph-1, nph-4, npp-11, npp-15, npp-18, npp-19, npr-1, npr-13, npr-16, npr-3, nra-1, nra-2, nra-4, nrd-1, nrf-6, nsf-1, nspd-9, nsy-1, ntl-2, nucb-1, nud-1, nud-2, nuo-2, nuo-3, nurf-1, obr-3, obr-4, ocr-2, oct-1, odc-1, odr-2, odr-3, oga-1, oig-1, oig-2, oig-3, old-1, olrn-1, oma-1, oma-2, orai-1, osm-1, osm-10, osm-11, osm-12, osm-3, osm-5, osm-6, osm-7, osm-9, osr-1, ost-1, otub-1, paa-1, pabp-2, pac-1, pad-1, pad-2, pag-3, pah-1, pak-1, pak-2, pal-1, pam-1, pap-1, paqr-1, paqr-2, par-1, par-2, par-3, par-4, par-5, par-6, pas-4, pat-10, pat-2, pat-3, pat-4, pbrm-1, pbs-6, pcca-1, pcm-1, pde-1, pde-2, pde-3, pdfr-1, pdhk-2, pdi-1, pdi-3, pdk-1, pdl-1, pdr-1, peb-1, pef-1, pen-2, pept-2, pept-3, pes-1, pes-22, pes-4, pes-7, pes-8, pes-9, pfd-1, pfd-3, pfd-5, pfd-6, pfn-1, pfn-2, pgp-1, pgp-10, pgp-12, pgp-13, pgp-14, pgp-2, pgp-3, pgp-4, pgp-6, pgp-7, pgp-8, pgp-9, pha-1, pha-4, phb-2, phf-5, phg-1, php-3, phy-2, pig-1, pink-1, pkc-1, pkd-2, plc-1, plc-3, pld-1, plp-1, plx-1, plx-2, pme-3, pme-4, pmp-4, pmp-5, pmr-1, pnk-4, pod-2, pop-1, pos-1, ppk-1, ppk-2, ppk-3, pps-1, pptr-1, pptr-2, pqbp-1.1, pqe-1, pqm-1, pqn-35, pqn-44, pqn-47, pqn-67, pqn-75, pqn-92, pqn-95, prdx-2, prk-1, prkl-1, prmt-1, prp-8, prs-2, prx-11, prx-19, prx-6, pry-1, psa-3, psf-1, pssy-1, ptb-1, ptl-1, ptp-2, ptp-3, ptr-10, ptr-19, ptr-5, ptr-8, puf-8, pxd-1, pxf-1, pyk-1, pyp-1, pzf-1, qua-1, rab-1, rab-10, rab-11.1, rab-3, rab-30, rab-37, rab-6.2, rad-51, ral-1, rap-1, rap-2, rbf-1, rbg-3, rbx-1, rcn-1, rdy-2, rec-8, ref-1, ref-2, rer-1, ret-1, rev-1, rfc-2, rfl-1, rfp-1, rga-2, rga-3, rga-6, rgef-1, rgl-1, rgr-1, rgs-1, rgs-2, rgs-3, rgs-5, rgs-9, rhgf-1, rhgf-2, rhr-1, rhy-1, ric-19, ric-3, ric-4, ric-8, rict-1, rig-1, rig-3, rig-4, rig-6, rla-2, rle-1, rme-1, rme-2, rme-6, rme-8, rncs-1, rnf-113, rnr-1, rnr-2, rnt-1, rol-3, rol-6, rpa-0, rpa-1, rpa-2, rpb-2, rpia-1, rpl-10, rpl-11.2, rpl-17, rpl-19, rpl-23, rpl-24.1, rpl-28, rpl-31, rpl-32, rpl-34, rpl-5, rpl-7A, rpl-9, rpm-1, rpn-3, rpn-5, rpn-7, rpn-8, rps-0, rps-14, rps-17, rps-21, rps-22, rps-23, rps-27, rps-29, rps-4, rpt-2, rpt-3, rpt-4, rpt-5, rpy-1, rrf-3, rrt-2, rskd-1, rskn-2, rsks-1, rsp-3, rsp-8, rsr-1, rsy-1, ruvb-1, sad-1, sago-1, sams-3, sams-5, sas-4, sax-1, sax-2, sax-3, sax-7, sbp-1, sbt-1, sca-1, scc-1, scc-3, scd-1, scd-2, scm-1, scpl-1, scpl-2, scrm-3, sdha-1, sdhb-1, sdhd-1, sdn-1, sdpn-1, sdz-12, sdz-33, sdz-36, sdz-4, sea-1, sea-2, sec-23, sec-5, secs-1, sek-1, sel-10, sel-2, sel-5, sel-8, sel-9, sem-4, sem-5, sepa-1, ser-1, ser-2, ser-3, ser-4, ser-5, ser-7, set-1, set-16, set-17, set-18, set-2, seu-1, sft-4, sgca-1, sgcb-1, sgk-1, sgt-1, shc-1, shk-1, shl-1, shw-3, siah-1, sin-3, sip-1, sir-2.1, sir-2.2, skn-1, sknr-1, skp-1, skr-3, slo-1, slo-2, slt-1, sma-1, sma-2, sma-3, sma-4, sma-5, sma-6, smc-4, smd-1, smg-2, smg-4, smgl-1, smi-1, smk-1, smo-1, smp-1, smp-2, snap-1, snb-1, snf-1, snf-11, snf-3, snf-5, snf-6, sng-1, snr-1, snr-2, snr-3, snr-4, snr-5, snr-6, snr-7, snt-1, snt-4, snx-1, soc-2, sod-2, sod-4, sop-3, sox-2, sox-3, spas-1, spat-3, spc-1, spd-2, spe-39, sph-1, spl-1, spn-4, spp-10, spp-7, spt-5, sptf-2, sptf-3, sqt-3, sqv-1, sqv-4, sqv-5, sqv-7, sra-10, sra-11, sra-13, sra-14, sra-17, sra-2, sra-20, sra-21, sra-22, sra-23, sra-25, sra-28, sra-29, sra-35, sra-36, sra-38, sra-39, sra-4, sra-6, sra-7, srab-1, srab-10, srab-11, srab-12, srab-13, srab-14, srab-16, srab-20, srab-21, srab-23, srab-24, srab-3, srab-4, srab-6, srab-7, srab-8, srab-9, srb-12, srb-13, srb-16, srb-3, srb-5, srb-6, srb-7, srbc-52, srbc-58, src-1, src-2, srd-15, srd-33, srd-36, srd-39, sre-1, sre-11, sre-12, sre-16, sre-21, sre-22, sre-23, sre-24, sre-26, sre-27, sre-28, sre-29, sre-30, sre-31, sre-32, sre-37, sre-4, sre-42, sre-43, sre-44, sre-45, sre-47, sre-51, sre-52, sre-53, sre-56, sre-6, sre-7, srf-3, srg-13, srg-2, srg-30, srg-33, srg-5, srg-62, srg-7, srg-8, srg-9, srgp-1, srh-11, srh-128, srh-15, srh-18, srh-182, srh-2, srh-279, srh-28, srh-281, srh-79, srh-92, sri-19, sri-24, sri-28, sri-33, sri-34, srj-24, srp-2, srp-3, srp-6, srr-6, srsx-29, srsx-30, srsx-34, srsx-9, srt-13, srt-20, srt-26, srt-27, srt-28, srt-29, srt-36, srt-63, srt-65, srt-68, srt-7, srt-70, srt-8, sru-17, sru-19, sru-27, sru-31, sru-38, srv-1, srv-4, srw-103, srw-108, srw-112, srw-118, srw-138, srw-139, srw-140, srw-85, srx-102, srx-105, srx-108, srx-110, srx-113, srx-114, srx-12, srx-41, srx-47, srx-76, srx-9, srxa-1, srxa-14, srxa-15, srxa-6, srxa-7, srxa-8, srz-1, srz-103, srz-13, srz-16, srz-28, srz-42, srz-67, srz-74, srz-94, srz-99, sta-1, stam-1, stau-1, stdh-1, stdh-4, stg-1, stg-2, sti-1, stim-1, stn-1, sto-1, sto-4, str-108, str-111, str-112, str-163, str-168, str-90, sul-3, sulp-1, sulp-7, sun-1, sup-10, sup-12, sup-35, sup-9, sur-2, sur-5, sur-6, sur-7, sut-1, sut-2, swan-2, swd-3.1, swd-3.2, syd-1, syd-2, syd-9, syg-1, syg-2, sym-4, syn-2, syn-3, syn-4, tab-1, taf-5, tag-120, tag-123, tag-130, tag-135, tag-140, tag-141, tag-144, tag-163, tag-170, tag-172, tag-174, tag-175, tag-18, tag-180, tag-196, tag-197, tag-210, tag-218, tag-224, tag-231, tag-232, tag-233, tag-24, tag-242, tag-243, tag-246, tag-250, tag-256, tag-257, tag-277, tag-290, tag-297, tag-299, tag-304, tag-320, tag-332, tag-335, tag-349, tag-52, tag-60, tag-68, tag-93, tag-96, tag-97, tap-1, tat-3, tax-2, tax-4, tax-6, tba-1, tba-2, tba-6, tba-8, tbb-4, tbc-10, tbc-11, tbc-13, tbc-16, tbc-18, tbc-2, tbc-3, tbc-4, tbc-8, tbc-9, tbg-1, tbh-1, tbx-2, tbx-35, tbx-37, tbx-38, tbx-39, tbx-8, tbx-9, tcl-2, ten-1, tgt-2, tig-2, tir-1, tkt-1, tli-1, tlk-1, tlp-1, tnc-2, tni-3, tni-4, tnt-2, toc-1, toe-4, tom-1, tomm-7, top-1, top-3, tor-2, tpa-1, tph-1, tpk-1, tps-1, tps-2, tpst-1, tra-1, tre-2, tre-3, tre-4, tre-5, trim-9, trp-1, trp-2, trpa-1, trs-1, trx-1, trxr-2, tsg-101, tsn-1, tsp-7, tth-1, ttn-1, ttr-8, ttx-3, ttx-7, ttyh-1, tub-1, twk-1, twk-16, twk-17, twk-2, twk-20, twk-22, twk-23, twk-29, twk-3, twk-30, twk-32, twk-4, twk-46, twk-6, twk-9, tyra-2, tyra-3, tza-1, tza-2, uaf-2, uba-1, ubc-1, ubc-14, ubc-20, ubc-23, ubc-25, ubc-3, ubl-5, ubq-1, ubq-2, ubxn-3, ubxn-6, ucp-4, ucr-2.2, ufd-2, ufd-3, ugt-58, ugt-61, ugt-8, ugt-9, uig-1, unc-1, unc-103, unc-104, unc-108, unc-11, unc-112, unc-115, unc-116, unc-122, unc-129, unc-13, unc-130, unc-14, unc-15, unc-16, unc-17, unc-18, unc-2, unc-25, unc-26, unc-27, unc-3, unc-30, unc-31, unc-32, unc-33, unc-34, unc-36, unc-38, unc-39, unc-4, unc-40, unc-41, unc-42, unc-43, unc-44, unc-45, unc-46, unc-47, unc-49, unc-5, unc-50, unc-51, unc-52, unc-53, unc-54, unc-55, unc-57, unc-58, unc-59, unc-6, unc-60, unc-61, unc-62, unc-63, unc-64, unc-68, unc-69, unc-7, unc-70, unc-71, unc-73, unc-75, unc-76, unc-78, unc-79, unc-8, unc-80, unc-83, unc-86, unc-87, unc-89, unc-9, unc-93, unc-94, unc-95, unc-96, unc-97, unc-98, vab-1, vab-10, vab-15, vab-19, vab-2, vab-3, vab-7, vab-8, vab-9, vang-1, vav-1, vbh-1, vem-1, ver-3, vha-1, vha-11, vha-12, vha-13, vha-15, vha-16, vha-17, vha-2, vha-3, vha-4, vha-5, vha-6, vha-7, vha-8, vhp-1, vig-1, viln-1, vps-11, vps-18, vps-2, vps-32.1, vps-35, vps-36, vps-54, vrk-1, wdr-23, wht-2, wip-1, wnk-1, wrk-1, wrn-1, wrt-1, wrt-10, wrt-2, wrt-3, wrt-4, wrt-5, wrt-6, wrt-8, wrt-9, wsp-1, wts-1, wwp-1, xbx-1, xbx-3, xbx-4, xbx-5, xbx-6, xbx-7, xnd-1, xnp-1, xpc-1, xpo-3, xrn-2, yop-1, zag-1, zfp-1, zfp-2, zig-1, zig-2, zig-3, zig-4, zig-5, zig-6, zig-8, zip-4, zip-5, zoo-1, ztf-1, ztf-11, ztf-12, ztf-16, ztf-17, ztf-18, ztf-2, ztf-22, ztf-4, ztf-6, ztf-7, ztf-8, zyg-12, zyg-8, zyx-1

10 CONTINGENCY TESTING

10.1 OVERVIEW

As noted in Section 7.4, having a rich set of data about gene expression across the entire organism enables us to link genotypic features to cell phenotype, because genes can be located within the genome and gene expression levels can be mapped from tissues onto the individual cells that comprise those tissues. *Gene predicates* are the result of bimodality testing for the presence/absence, size or frequency of specific genomic features within the genes that are active in a given cell, and *cell attributes* give a set of attribute tags that are present or absent for each cell based on cell metadata (cell type, anatomical role, cell fate, cell program, timing of division, etc.).

This allows for the calculation of the conditional probability of that a cell has a specific attribute given that a gene predicate is true for genes active in that cell. Furthermore we can test for *correlations* between gene predicates and cell attributes in an attempt to find genes or properties of genes that may be implicated in a cell having a certain attribute.

The simplest such correlation test is an odds ration. For example, if we want to test *how many more times likely it is that a cell* X *dies through apoptosis when a gene* Y *is expressed, compared to when* Y *is not expressed,* then the odds ratio of the cell attribute "cell X dies" can be tested against the gene predicate "gene Y is expressed" as:

$$\begin{aligned}&\mathrm{OR}(\text{cell X dies; gene Y expressed}) = \\ &\quad \frac{\mathrm{P}(\text{X dies} \mid \text{Y expressed})/\mathrm{P}(\neg\text{X dies} \mid \text{Y expressed})}{\mathrm{P}(\text{X dies} \mid \neg\text{Y expressed})/\mathrm{P}(\neg\text{X dies} \mid \neg\text{Y expressed})}\end{aligned} \tag{10.1}$$

10.2 CONTINGENCY TEST METHODS

The odds ratio as calculated above is an interesting measure, particularly if it is already known that there exists causal relationship between the two, as in our case where changes in gene expression do in fact cause changes in cell functioning. The way the probabilities are conditioned allows the odds ratio to be used to gain insight into the strength of the causal relationship from gene properties to cell attributes, as opposed to just giving insights into correlations between the two.

At the same time, however, the odds ratio defined above suffers from instability when counts in any one of the four numbers are small: at the limit, if zero instances of one of the four contingency classes are counted, then the odds ratio can end up with a calculated value of zero or infinity. Adding Laplacian pseudocounts to numerators and/or denominators can help the situation but biases the odds ratios. Since the WormBase data can be quite sparse, the odds ratio is probably not the best choice for this dataset. Consequently, it is worth considering other contingency test methods.

A survey of exact inference for contingency tables is given in [4]. Contingency testing methods all have their own strengths and weaknesses, and contingency testing is an area of ongoing debate and research. Consequently, to avoid lending too much credence to any single contingency testing method, a number of different contingency tests were performed on each cell attribute / gene predicate combination, including Fisher's exact test, Pearson's Chi squared test.

1. Bayesian contingency testing (defined in Section 10.3), resulting in a z-score that effectively gives the number of standard deviations difference between the maximum likelihood estimate of $\mathrm{P}(\text{cell attr A} \mid \text{gene pred } B)$ and $\mathrm{P}(\text{cell attr A} \mid \neg\text{gene pred } B)$.

2. The log likelihood ratio between the right-tailed and left-tailed Fisher's exact test: this is the same as the two-tailed Fisher's exact test, except that you get a signed number as a result, which tells you whether the correlation is positive or negative. (Fisher's exact test is argued to be conservative [8], i.e. that its actual rejection rate is below the nominal significance level; it is also quite hard to compute quickly, but a gamma function approximation was used for speed.)

3. The log odds ratio, as calculated in 10.1 above. (Can over-estimate or under-estimate actual odds ratio when counts are small.)

4. The Chi-squared test. (Not reliable for counts less than 10.)

5. The log of the ratio between the likelihood of the cell attribute given the gene predicate is true, over the likelihood of the cell attribute given that the gene predicate is not true:

   $$\mathrm{P}(\text{cell attr X} \mid \text{gene Y}: \text{ pred Z})/\mathrm{P}(\text{cell attr X} \mid \text{gene Y}: \neg\text{pred Z})$$

   (This amounts to a simple odds ratio test for a change in likelihood of the cell attribute as a result of some gene property being manifest; it is the odds ratio equivalent of the Bayesian test listed above.)

10.3 BAYESIAN CONTINGENCY TESTING

A very promising alternative to traditional frequentist contingency testing is to perform a simple (standard) Bayesian test of the posterior distribution of the difference in two random variables. This is roughly equivalent to Fisher's exact test, but works in the Bayesian space.

10.3.1 *Contingency Table*

Consider the following contingency table:

	Cell dies through apoptosis	**Cell does not die through apoptosis**	**TOTAL**
Gene *Y* is expressed	31 (y_1)	297	328 (n_1)
Gene *Y* is not expressed	2 (y_2)	221	223 (n_2)
TOTAL	33	518	551

10.3.2 *Hypothesis*

Let

$$\theta_1 = y_1/n_1 \tag{10.2}$$

represent the proportion of cells that die through apoptosis that also express gene *Y*, and

$$\theta_2 = y_2/n_2 \tag{10.3}$$

represent the proportion of cells that die through apoptosis but do *not* express gene *Y*. The statistical hypothesis we wish to investigate is whether $\Theta_1$is greater than, equal to, or less than Θ_2, i.e. given

$$\delta = \theta_1 - \theta_2 \tag{10.4}$$

we wish to compare the difference δ to zero. A difference of zero indicates that gene *Y* does not affect cell death; a difference greater than zero indicates that gene *Y* may upregulate cell death; and a difference less than zero indicates that gene *Y* may downregulate cell death.

10.3.3 *Beta-Binomial model*

As a first approximation[1], it is reasonable to model y_1 and y_2 using the binomial distribution:

$$\begin{aligned} y_1 &\sim \text{Binom}(n_1, \theta_1) \\ y_2 &\sim \text{Binom}(n_2, \theta_2) \end{aligned} \qquad (10.5)$$

Also as a first approximation, we can assume simple uniform priors on the proportions:

$$\begin{aligned} \theta_1 &\sim \text{Unif}(0,1) &= \text{Beta}(0,1) \\ \theta_2 &\sim \text{Unif}(0,1) &= \text{Beta}(0,1) \end{aligned} \qquad (10.6)$$

Given the conjugacy of the beta for the binomial, we can analytically compute the posteriors:

$$\begin{aligned} p(\theta_1 \,|\, y_1, n_1) &= \text{Beta}(\theta_1 \,|\, y_1 + 1, n_1 - y_1 + 1) \\ p(\theta_2 \,|\, y_2, n_2) &= \text{Beta}(\theta_2 \,|\, y_2 + 1, n_2 - y_2 + 1) \end{aligned} \qquad (10.7)$$

The difference of two distributions is the convolution of the distributions, so the analytical solution to the posterior density δ is

$$p(\delta \,|\, y, n) = \int_{-\infty}^{\infty} \text{Beta}(\theta_1 \,|\, y_1 + 1, n_1 - y_1 + 1)\, \text{Beta}(\theta_2 \,|\, y_2 + 1, n_2 - y_2 + 1) \qquad (10.8)$$

Rather than try to solve this integral analytically (which may not even yield a closed form), it is simple to instead estimate this integral by simulation, by sampling from θ_1 and θ_2 and building a sample distribution out of the differences. However, a reasonably large number (tens or hundreds of thousands) of samples may be required to ensure high accuracy in estimating a single difference $\overline{\delta}$, which makes this method intractable when billions of contingency tests must be run, as is the case for the cross product of all cell attributes with all gene predicates.

10.4 THE Z-SCORE OF A CONTINGENCY TEST

The mean value of the posterior density δ (minus zero, which is what we are comparing the difference to) divided by its standard deviation yields a *z-score* for the contingency test

$$z = \frac{\mu}{\sigma}, \qquad (10.9)$$

[1] [Following the working on the LingPipe blog post at `http://goo.gl/hVbP`]

which is a measure of the number of standard deviations in difference between the maximum likelihood estimates of the prior distributions of θ_1 and θ_2. Converting the mean of the posterior density of the difference is a way of normalizing differences so that different contingency tests can be compared.

Note that z-scores assume a normal distribution. Due to the central limit theorem, it is reasonable to assume that the convolution of two beta distributions together is somewhat normal in shape (at least closer to normal than either of the individual beta distributions). Also, when estimating posterior density using simulation, a large number of points must be sampled to reduce error in estimating the z-score.

10.5 NEW PARAMETRIC ESTIMATOR FOR BAYESIAN CONTINGENCY TEST

While it may be hard or potentially impossible to find a closed form solution for Eq. 10.8, if all that is needed is the z-score (normalized difference between the mean and zero) of the posterior, we can use the closed form of the mean and variance of the beta distribution to estimate the mean and variance of the difference between two beta-distributed random variables.

The mean and standard deviation of the beta distribution Beta(α, β) with hyperparameters α and β are, respectively

$$\mu = \frac{\alpha}{\alpha+\beta} \tag{10.10}$$

$$\sigma = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{10.11}$$

In our case, referring to Eq. 10.7, we have

$$\begin{aligned} \alpha_1 &= y_1+1, & \beta_1 &= n_1-y_1+1 \\ \alpha_2 &= y_2+1, & \beta_2 &= n_2-y_2+1 \end{aligned} \tag{10.12}$$

The mean and standard deviation of the difference in two distributions are given by

$$\mu_\delta = \mu_1 - \mu_2 \tag{10.13}$$

$$\sigma_\delta = \sqrt{\sigma_1^2 + \sigma_2^2 + 2\operatorname{cov}(\theta_1, \theta_2)}. \tag{10.14}$$

In Magnussen [28, Eq. 4], the first-order Taylor series approximation for the covariance between two beta distributions is given as

$$\operatorname{cov}(\theta_1, \theta_2) = \frac{\alpha_1 \times \alpha_2 + (1 + \beta_1) \times (1 + \beta_2)}{(\alpha_1 + \beta_1) \times (\alpha_2 + \beta_2) \times (1 + \alpha_1 + \beta_1) \times (1 + \alpha_2 + \beta_2)}. \tag{10.15}$$

However it turns out that for most pairs of beta distributions, this term is very small compared to the terms $\sigma_1^2$and σ_2^2 in Eq. 10.14, and in cases where the term is non-negligible (with extreme counts, i.e. when y_1 is close to 0 or n_1, or when y_2 is close to 0 or n_2), the relatively large size of the term is actually due to error from using only a first-order Taylor series approximation to the covariance. If we assume the covariance term is negligible and remove it from Eq. 10.14, yielding

$$\sigma_\delta \approx \sqrt{\sigma_1^2 + \sigma_2^2}, \tag{10.16}$$

then test this closed-form estimate against the simulated estimate of the standard deviation derived by sampling differences in large numbers of values drawn from pairs of beta distributions, with $n_1, n_2 \in [5, 1000]$ and $y_i \in [0, n_i]$, then we observe that the correlation between the closed-form estimate and the empirically simulated estimate is very close (the correlation coefficient R is very close to 1.0). This shows that the covariance term is indeed negligible in practice for parameters in this range, and can be safely ignored, simplifying our calculations.

Substituting Eq. 10.12 into Eq. 10.10 and Eq. 10.11 yields formulas for μ_1, μ_2, σ_1 and σ_2 in terms of α_1, β_1, α_2 and β_2. Substituting μ_1, μ_2, σ_1 and σ_2 into Eq. 10.13 and Eq. 10.16 yields formulas for μ_δ and σ_δ in terms of α_1, β_1, α_2 and β_2. Finally, substituting μ_δ and σ_δ into Eq. 10.9 and simplifying yields:

$$z_\delta \approx \frac{\frac{\alpha_1}{\alpha_1+\beta_1} + \frac{\alpha_2}{\alpha_2+\beta_2}}{\sqrt{\frac{\alpha_1\beta_1}{(\alpha_1+\beta_1)^2(\alpha_1+\beta_1+1)} + \frac{\alpha_2\beta_2}{(\alpha_2+\beta_2)^2(\alpha_2+\beta_2+1)}}}. \tag{10.17}$$

Note that z_δ may be negative, indicating inverse correlation.

Substituting in Eq. 10.12 and simplifying, we can also express Eq. 10.17 directly in terms of y_1, y_2, n_1 and n_2:

$$z_\delta \approx \frac{\frac{y_1+1}{n_1+2} - \frac{y_2+1}{n_2+2}}{\sqrt{\frac{(y_1+1)(n_1-y_1+1)}{(n_1+2)^2(n_1+3)} + \frac{(y_2+1)(n_2-y_2+1)}{(n_2+2)^2(n_2+3)}}}. \tag{10.18}$$

Effectively Eq. 10.17 illustrates that this method at its core examines difference-of-proportions, i.e. $y_1/n_1 - y_2/n_2$, only with built-in pseudocounts, and a normalizer in the denominator (the standard deviation) that

Algorithm 10.1 Fast novel estimator for Bayesian posterior difference between two binomially distributed variables, as described in Eq. 10.17

```
// Directly compute the approximate parameters of the
// posterior of the difference in the two beta distributions;
// much faster than empirical sampling.

float a1 = y1 + 1, b1 = x1 + 1;
float a2 = y2 + 1, b2 = x2 + 1;
float sum1 = n1 + 2; // == (a1 + b1);
float sum2 = n2 + 2; // == (a2 + b2);
float mean1 = a1 / sum1;
float mean2 = a2 / sum2;
float var1 = a1 * b1 / (sum1 * sum1 * (sum1 + 1));
float var2 = a2 * b2 / (sum2 * sum2 * (sum2 + 1));
float meanAnalyt = mean1 - mean2;
float stddevAnalyt = (float) Math.sqrt(var1 + var2);
float zScore = meanAnalyt / stddevAnalyt;
```

is maximized at proportions near 0.5 (where the beta distribution has the highest variance).

(For a Java implementation of Eq. 10.17, see Algorithm 10.1.)

Eq. 10.18 represents a closed-form estimate to the z-score that yields almost identical results to the applying the empirical Bayesian sampling method and taking a large number of samples to estimate the z-score. The near identical results are due to only the negligible covariance term being estimated as zero in the closed form, as described above. This formula appears to be novel.

This z-score estimate z_δ is significantly faster to compute than the empirical Bayesian sampling method for testing difference, and yet is a closed form estimate. Due to its origin in the beta distribution, z_δ should also produce reasonable results for small count statistics (z_δ is always defined even when y_1 is close to or equal to 0 or n_1, or when y_2 is close to or equal to 0 or n_2).

The main source of bias in this estimator arises from assuming that the convolution of two beta functions is approximately normal (allowing for the calculation of the z-score statistic using Eq. 10.9) – this is only approximately true for y values close to 0, and as a result the z-score may be less accurate or less meaningful.

11

PERFORMING CONTINGENCY TESTS BETWEEN CELL ATTRIBUTES AND GENE PREDICATES

11.1 METHOD

Contingency tests were performed between all cell attributes (Section 9.1.1) and all gene predicates (Section 9.2.1). Contingency tests with counts that were too small to produce reliable results in some of the contingency test statistics (i.e. counts smaller than 10) were eliminated, producing a sparse table of contingency test results of size ~600k rows – the number of gene predicates – by 21k columns – the number of cell attributes (see Figure 24(a)). This table is effectively produced by using the cell axis to join the table of gene predicates for each cell with the table of cell attributes for each cell. For each gene predicate, we look up the list of all cells for which the gene predicate is true, then find the list of all cell attributes possessed by those cells, in order to build the contingency table counts $n(\text{cell attr } X \mid \text{gene predicate } Y)$, $n(\text{gene predicate } Y)$, $n(\text{cell attr } X \mid \neg\text{gene predicate } Y)$ and $n(\neg\text{gene predicate } Y)$.

The table shown in Figure 24(a) is output as a large text file, with one table entry per row, and includes the row name (cell attribute) and column name (gene predicate) in each row to allow for easy filtering of test scores by simply grepping for cell attribute and/or gene predicate.

An example output line is given below:

```
10.514316       106.19144       198.83801       152.11592
    5.024643        129.11458       95      625     1       1268
    cell-group:embryonic_death      gene-expressed-flt-1
```

The order of these fields is:

1. Fast approximation to Bayesian difference (expressed as a z-score, i.e. number of standard deviations difference in the likelihood of the cell attribute given that the gene predicate is true, compared to the likelihood of the cell attribute given that the gene predicate is not true).

2. Log of the p-value of the signed result of the two-tailed Fisher's Exact Test.

3. The Chi-Squared test statistic.

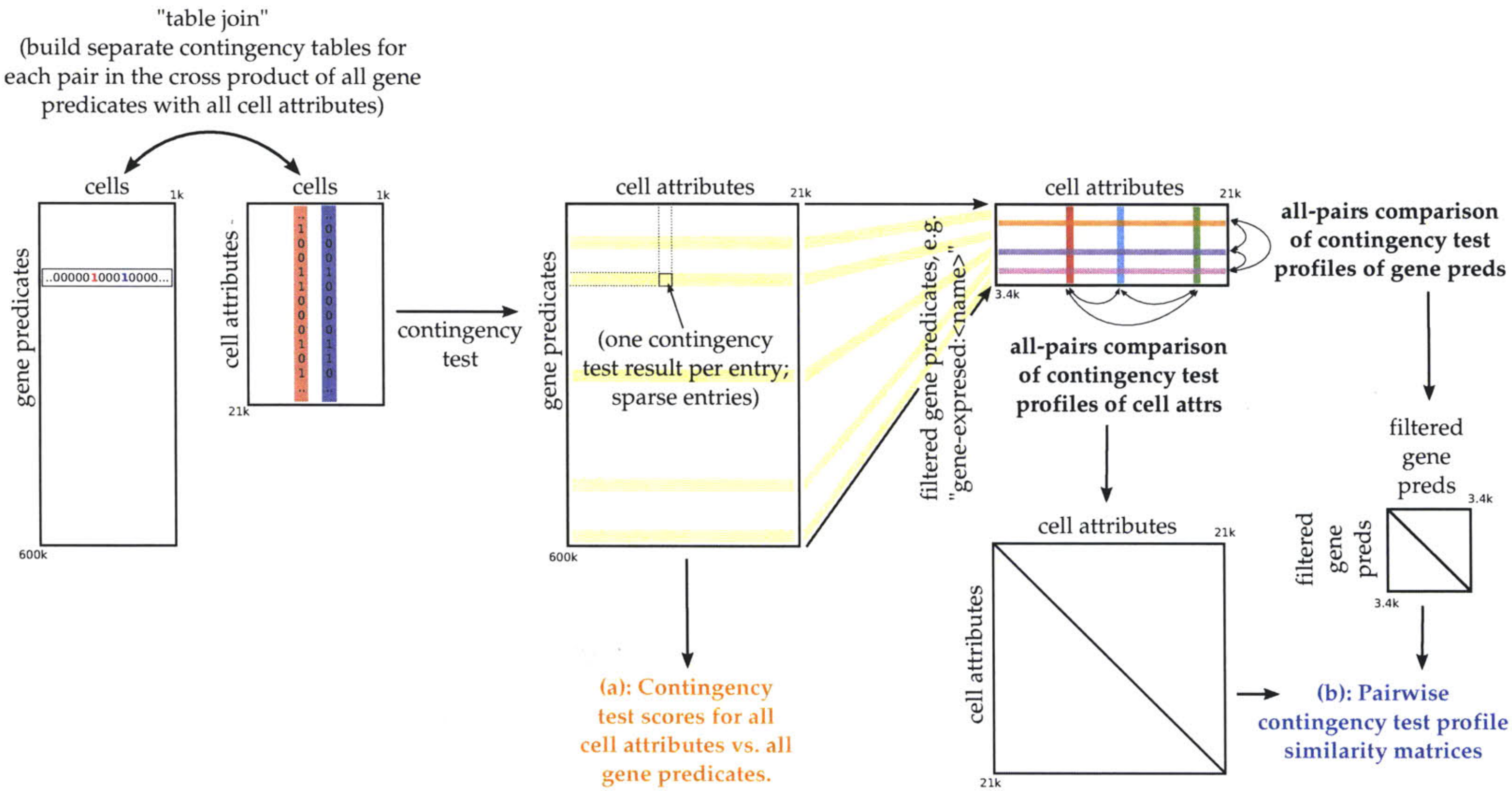

"table join"
(build separate contingency tables for each pair in the cross product of all gene predicates with all cell attributes)
cells
1k
gene predicates
..000001000010000...
600k
cells
1k
cell attributes
21k
contingency test
cell attributes
21k
gene predicates
(one contingency test result per entry; sparse entries)
600k
filtered gene predicates, e.g. "gene-expresed:<name>"
cell attributes
21k
3.4k
all-pairs comparison of contingency test profiles of gene preds
all-pairs comparison of contingency test profiles of cell attrs
cell attributes
21k
cell attributes
21k
filtered gene preds
3.4k
filtered gene preds
3.4k
(a): Contingency test scores for all cell attributes vs. all gene predicates.
(b): Pairwise contingency test profile similarity matrices

Figure 24: (a) Contingency testing between all cell attributes and all gene predicates; (b) comparison of pairs of entire rows and columns of contingency test scores to produce a pairwise distance matrix for all pairs of cell attributes that are correlated with similar gene predicate contingency score profiles, and all pairs of gene predicates with similar cell attribute contingency score profiles.

4. The odds ratio given in Section 10.1.

5. The log of the odds ratio.

6. The likelihood of the cell attribute given the gene predicate is true over the likelihood of the cell attribute given that the gene predicate is not true, i.e. the difference given in "1." above, expressed as a ratio rather than the z-score of the difference.

7. The number of cells which have the cell attribute and in which the gene predicate is true.

8. The number of cells in which the gene predicate is true.

9. The number of cells which have the cell attribute and in which the gene predicate is NOT true.

10. The number of cells in which the gene predicate is NOT true.

11. The cell attribute.

12. The gene predicate.

Fields 7-10 are the contingency table values y_1, n_1, y_2, and n_2. Field 1, the z-score of the Bayesian difference (generally referred to as just "the z-score" of the contingency test below) is the statistic that was primarily focused on, but the other statistics were calculated in case they are helpful.

The entire process was repeated for the cross product of lineage-specific vs. tissue-specific gene associations (Section 6.2) and lineagePath vs. canonicalPath (Section 5.6.1), resulting in four different output files. Depending on whether highly cell-specific results are desired (i.e. using lineage-specific associations, with the tradeoff that the data is sparser) or much denser and therefore richer results are desired (with the tradeoff that tissue-specific mappings are less specific and so likely introduce some spurious gene association mappings), either version of the results can be used. The canonicalPath versions can be used to investigate cell symmetries (left as future work).

The top few z-score results that include the keyword "death" in the cell attribute and "gene" in the gene predicate are shown in Figures 25 and 26 for lineage-specific and tissue-specific associations respectively.

Note that positive z-scores indicate that a gene predicate is positively implicated in the given cell attribute, but negative z-scores are also possible, which indicate that a gene predicate is negatively implicated in the given cell attribute, i.e. that it may be a repressor of that specific attribute of cell phenotype. Examples of the largest negative z-score results for tissue-specific associations are given in Figure 27.

```
5.142792        23      107     5       406     daughter:brief-id:programmed cell death gene-expressed-elt-6
5.142792        23      107     5       406     daughter:death-in:herm  gene-expressed-elt-6
5.142792        23      107     5       406     daughter:death-in:male  gene-expressed-elt-6
5.142792        23      107     5       406     daughter:death-in:one-or-both-genders   gene-expressed-elt-6
5.142792        23      107     5       406     daughter:program:death  gene-expressed-elt-6
4.948151        23      124     5       389     daughter:brief-id:programmed cell death gene-expressed-egl-18
4.948151        23      124     5       389     daughter:death-in:herm  gene-expressed-egl-18
4.948151        23      124     5       389     daughter:death-in:male  gene-expressed-egl-18
4.948151        23      124     5       389     daughter:death-in:one-or-both-genders   gene-expressed-egl-18
4.948151        23      124     5       389     daughter:program:death  gene-expressed-egl-18
4.614289        18      133     0       380     daughter:cell-group:embryonic_death     gene-expressed-hcp-3
4.537436        17      102     1       411     daughter:cell-group:embryonic_death     gene-expressed-flt-1
4.504727        17      107     1       406     daughter:cell-group:embryonic_death     gene-expressed-elt-6
4.406427        17      124     1       389     daughter:cell-group:embryonic_death     gene-expressed-egl-18
4.238692        26      292     2       221     daughter:brief-id:programmed cell death density-SNP-wholegene-inv
4.238692        26      292     2       221     daughter:death-in:herm  density-SNP-wholegene-inv
4.238692        26      292     2       221     daughter:death-in:male  density-SNP-wholegene-inv
4.238692        26      292     2       221     daughter:death-in:one-or-both-genders   density-SNP-wholegene-inv
4.238692        26      292     2       221     daughter:program:death  density-SNP-wholegene-inv
4.050139        18      102     10      411     daughter:brief-id:programmed cell death gene-expressed-flt-1
4.050139        18      102     10      411     daughter:death-in:herm  gene-expressed-flt-1
4.050139        18      102     10      411     daughter:death-in:male  gene-expressed-flt-1
4.050139        18      102     10      411     daughter:death-in:one-or-both-genders   gene-expressed-flt-1
```

Figure 25: Lineage-specific z-score results: results yielded by running "grep death odds-ratios-lineage-specific-lineagePath | grep gene | cut -f 1,7-12", i.e. filtering lineage-specific results for rows that include the keywords "death" (apoptosis) and "gene", and then stripping out all columns but the z-score, the contingency table values, the cell attribute and the gene predicate.

```
4.583439    18    149    0     465    daughter:cell-group:embryonic_death     gene-expressed-hcp-3
4.212135    18    289    0     325    daughter:cell-group:embryonic_death     gene-expressed-F25B5.2
4.161621    18    305    0     309    daughter:cell-group:embryonic_death     gene-expressed-F53C3.2
3.888885    17    267    1     347    daughter:cell-group:embryonic_death     gene-expressed-flt-1
3.863977    18    374    0     240    daughter:cell-group:embryonic_death     gene-expressed-cdc-42
3.825413    20    149    12    465    daughter:brief-id:programmed cell death gene-expressed-hcp-3
3.825413    20    149    12    465    daughter:death-in:herm  gene-expressed-hcp-3
3.746264    18    393    0     221    daughter:cell-group:embryonic_death     gene-expressed-lin-53
3.732592    18    395    0     219    daughter:cell-group:embryonic_death     gene-expressed-chd-3
3.725656    18    396    0     218    daughter:cell-group:embryonic_death     gene-expressed-set-2
3.688352    17    306    1     308    daughter:cell-group:embryonic_death     gene-expressed-elt-6
3.682583    18    402    0     212    daughter:cell-group:embryonic_death     gene-expressed-lin-35
3.673741    12    149    0     465    cell-group:embryonic_death      gene-expressed-hcp-3
3.577360    20    149    15    465    daughter:program:death  gene-expressed-hcp-3
3.517140    17    333    1     281    daughter:cell-group:embryonic_death     gene-expressed-egl-18
3.495626    20    149    16    465    daughter:death-in:male  gene-expressed-hcp-3
3.414355    20    149    17    465    daughter:death-in:one-or-both-genders   gene-expressed-hcp-3
3.161757    18    455    0     159    daughter:cell-group:embryonic_death     gene-expressed-cav-1
3.118549    24    289    8     325    daughter:brief-id:programmed cell death gene-expressed-F25B5.2
3.118549    24    289    8     325    daughter:death-in:herm  gene-expressed-F25B5.2
2.964183    12    374    0     240    cell-group:embryonic_death      gene-expressed-cdc-42
2.946309    11    267    1     347    cell-group:embryonic_death      gene-expressed-flt-1
2.876965    24    305    8     309    daughter:brief-id:programmed cell death gene-expressed-F53C3.2
```

Figure 26: Tissue-specific z-score results: results yielded by running "grep death odds-ratios-tissue-specific-lineagePath | grep gene | cut -f 1,7-12", i.e. filtering tissue-specific results for rows that include the keywords "death" (apoptosis) and "gene", and then stripping out all columns but the z-score, the contingency table values, the cell attribute and the gene predicate.

```
-2.208811    0    104   18    510    daughter:cell-group:embryonic_death     gene-expressed-sax-2
-2.208811    0    104   18    510    daughter:cell-group:embryonic_death     gene-expressed-spl-1
-2.208811    0    104   18    510    daughter:cell-group:embryonic_death     gene-expressed-srb-12
-2.208811    0    104   18    510    daughter:cell-group:embryonic_death     gene-expressed-sre-28
-2.208811    0    104   18    510    daughter:cell-group:embryonic_death     gene-expressed-tag-242
-2.208811    0    104   18    510    daughter:cell-group:embryonic_death     gene-expressed-zig-6
-2.209630    0    323   6     291    daughter:cell-group:death_in_both_sexes gene-expressed-gpa-7
-2.209630    0    323   6     291    daughter:cell-group:death_in_both_sexes gene-expressed-isp-1
-2.209630    0    323   6     291    daughter:cell-group:death_in_both_sexes gene-expressed-let-92
-2.209630    0    323   6     291    daughter:death-in:herm-only     gene-expressed-gpa-7
-2.209630    0    323   6     291    daughter:death-in:herm-only     gene-expressed-isp-1
-2.209630    0    323   6     291    daughter:death-in:herm-only     gene-expressed-let-92
-2.209630    0    323   6     291    daughter:death-in:herm-only     gene-expressed-tag-123
-2.209630    0    323   6     291    daughter:death-in:male-only     gene-expressed-gpa-7
-2.209630    0    323   6     291    daughter:death-in:male-only     gene-expressed-isp-1
-2.209630    0    323   6     291    daughter:death-in:male-only     gene-expressed-let-92
-2.209630    0    323   6     291    daughter:death-in:male-only     gene-expressed-tag-123
-2.210138    0    273   7     341    daughter:brief-id:death in hermaphrodite        gene-expressed-C37E2.1
-2.210138    0    273   7     341    daughter:brief-id:death in hermaphrodite        gene-expressed-F38A1.8
-2.210138    0    273   7     341    daughter:brief-id:death in hermaphrodite        gene-expressed-T10C6.5
-2.210138    0    273   7     341    daughter:brief-id:death in hermaphrodite        gene-expressed-cct-7
-2.210138    0    273   7     341    daughter:brief-id:death in hermaphrodite        gene-expressed-ceh-38
-2.210138    0    273   7     341    daughter:brief-id:death in hermaphrodite        gene-expressed-gly-12
```

Figure 27: The largest negative z-score results from tissue-specific gene associations.

- Genes that are implicated in a cell phenotypic attribute of interest (apoptosis; neurogenesis; syncitia; becoming a leaf cell, i.e. ceasing to divide; etc.) can easily be found in decreasing order of z-score by grepping through the results file for a cell attribute and the gene predicate substring "gene-expressed".
- Motifs that may regulate a given cell phenotype can be found by grepping through the results file for the cell attribute string and the gene predicate substring "motif-present".
- More complex analyses can be performed, such as looking for genes that are more highly implicated in a cell moving anterior compared to posterior, by looking for genes whose z-score when tested with the "direction:a" cell attribute is much higher than the z-score when tested with "direction:p".
- z-scores can be grouped by cell attribute or gene predicate and analyzed as vectors, as described in the next section.

11.3 PAIRWISE COMPARISON OF Z-SCORE PROFILE TO COMPUTE A "PHENOTYPE INFLUENCE NETWORK"

The z-scores for all cell attributes for a given gene predicate "gene-expressed-<name>" can be taken together as a vector that describes how a gene affects phenotypes across all cells (some positively, some negatively, some not at all). Then all pairs of gene predicate z-score vectors can be compared pairwise, and a distance score can be calculated between each pair of z-score vectors to produce a pairwise distance matrix between gene predicates, indicating gene predicates that have a similar effect on cell phenotype (Figure 24(b)). Similarly, vectors of z-scores for a given gene predicate can be pairwise-compared to find cell attributes that have similar gene-regulatory origins. The result is a pairwise gene-gene distance matrix or a pairwise cell-cell distance matrix respectively.

In the case of a pairwise gene-gene distance matrix, the resultant "phenotype influence network" is significantly different in interpretation than a standard gene-gene interaction network, because the distances represent similarity in cell phenotype when a given gene is expressed, i.e. the network indicates genes that have a similar effect several layers of complexity downstream from the direct action of most genes. This is a very valuable artifact because it allows a much higher-level analysis of the effects of genes. It is left as future work, however, to correlate phenotype influence networks with gene interaction networks to find similarities and differences.

If "motif-present-########" gene predicates are chosen in the building of a phenotype influence network, there are $4^8 = 65,536$ different 8-mer motifs and so the all-pairs distance matrix represents a completely-connected graph of 65,536 nodes and ${}^{65536}C_2 = 2,147,450,880$ edges. This graph is very hard to visualize effectively (with about 1000 times more edges than there are pixels on an average screen), and even harder to lay out, especially using a graph layout algorithm that scales superlinearly in the number of edges (as do most graph layout algorithms that try to avoid edge crossings and/or edge/node overlaps). Consequently the graph was thinned out to include only the closest edges above some threshold significance level, or such that the maximum degree of a node was fixed (so that only the most significant / shortest edges were connected to each node), producing a graph with the name number of nodes but far fewer edges. This graph was then laid out using the "GEM (Frick)" Graph Expectation Maximization algorithm of the Tulip graph layout and visualization program[1].

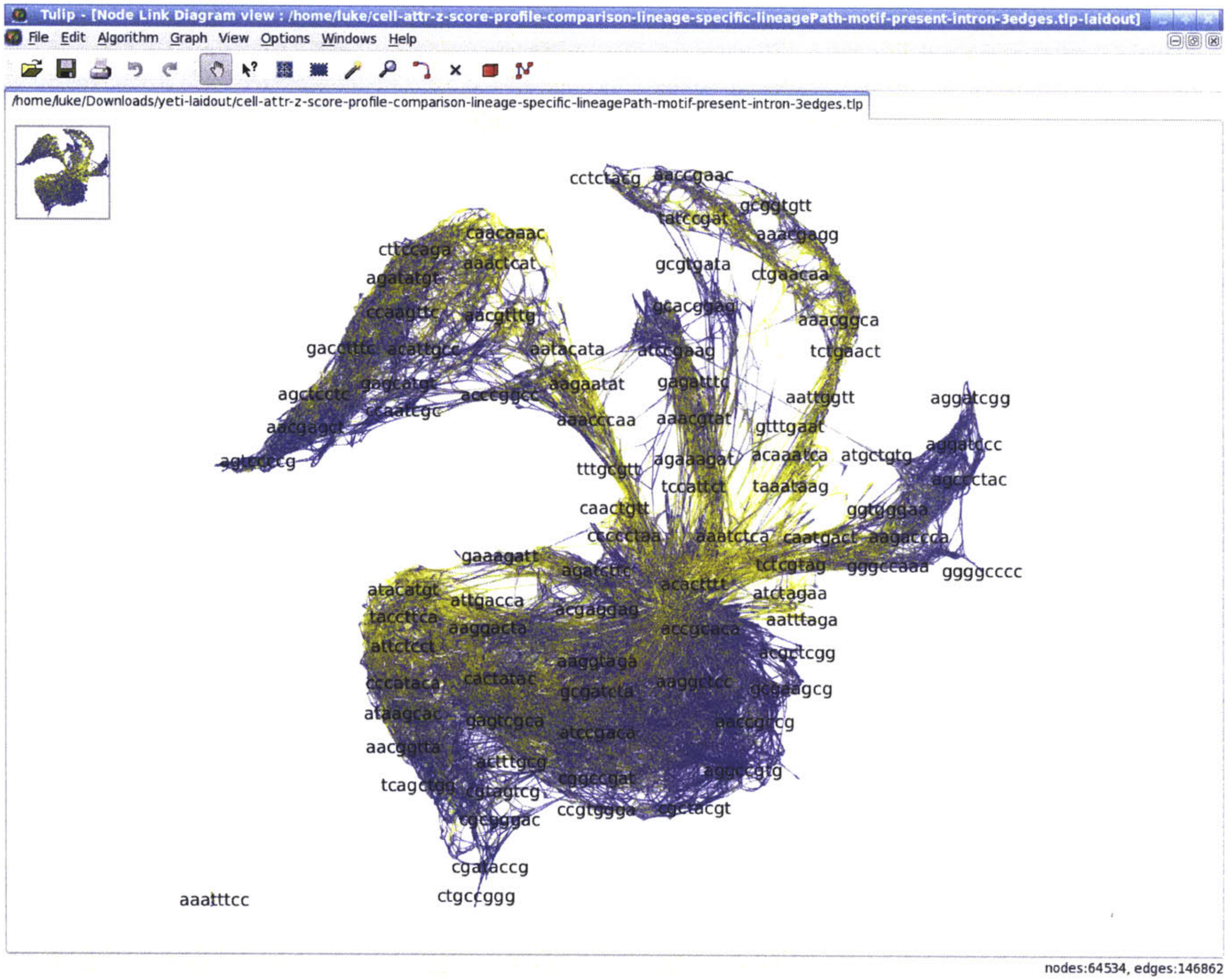
The result of laying out the phenotype influence network for intronic motifs (thinned by limiting node degree to a maximum of 3) is shown in Figure 28 on the facing page. Clearly the graph is highly non-random in structure, indicating that the space of all possible intronic 8-mer motifs is divided into different regions that are present in the introns of genes expressed in cells with different phenotypic traits. Just exactly which functional motifs these motif-space regions correspond to, and how the motifs within different subclusters in the laid-out graph are related to each other, would require extensive further research, but these would be very interesting questions to ask (not just for the introns, but also for similar graphs of motifs in the 3′ and 5′ UTRs, the sequence upstream of genes, and even for exon motifs, which are likely to have regulatory roles in some cases).

A phenotype influence network for "gene-expressed" gene predicates rather than "motif-present" gene predicates is also shown in Figure 30 on page 111.

11.4 MULTIRESOLUTION ANALYSIS OF CLUSTERS IN PHENOTYPE INFLUENCE NETWORKS

These phenotype influence networks exhibited different types of structure at different scales, and various top-down clustering methods (spectral clustering [32] and Power Iteration Clustering (PIC) [26]) as well as bottom-up clustering methods (hierarchical agglomerative clustering, min/mean/-median/max weighted) were used to try to extract useful clusterings. However it proved quite difficult to automatically extract clusters that naturally fell out of graph layout algorithms like Tulip's "GEM (Frick)".

[1] http://tulip.labri.fr/


Tulip - [Node Link Diagram view : /home/luke/cell-attr-z-score-profile-comparison-lineage-specific-lineagePath-motif-present-intron-3edges.tlp-laidout]

/home/luke/Downloads/yeti-laidout/cell-attr-z-score-profile-comparison-lineage-specific-lineagePath-motif-present-intron-3edges.tlp

nodes:64534, edges:146862

Figure 28: Pairwise "phenotype influence network" for intronic motifs. (Only a few motifs are actually labeled.)

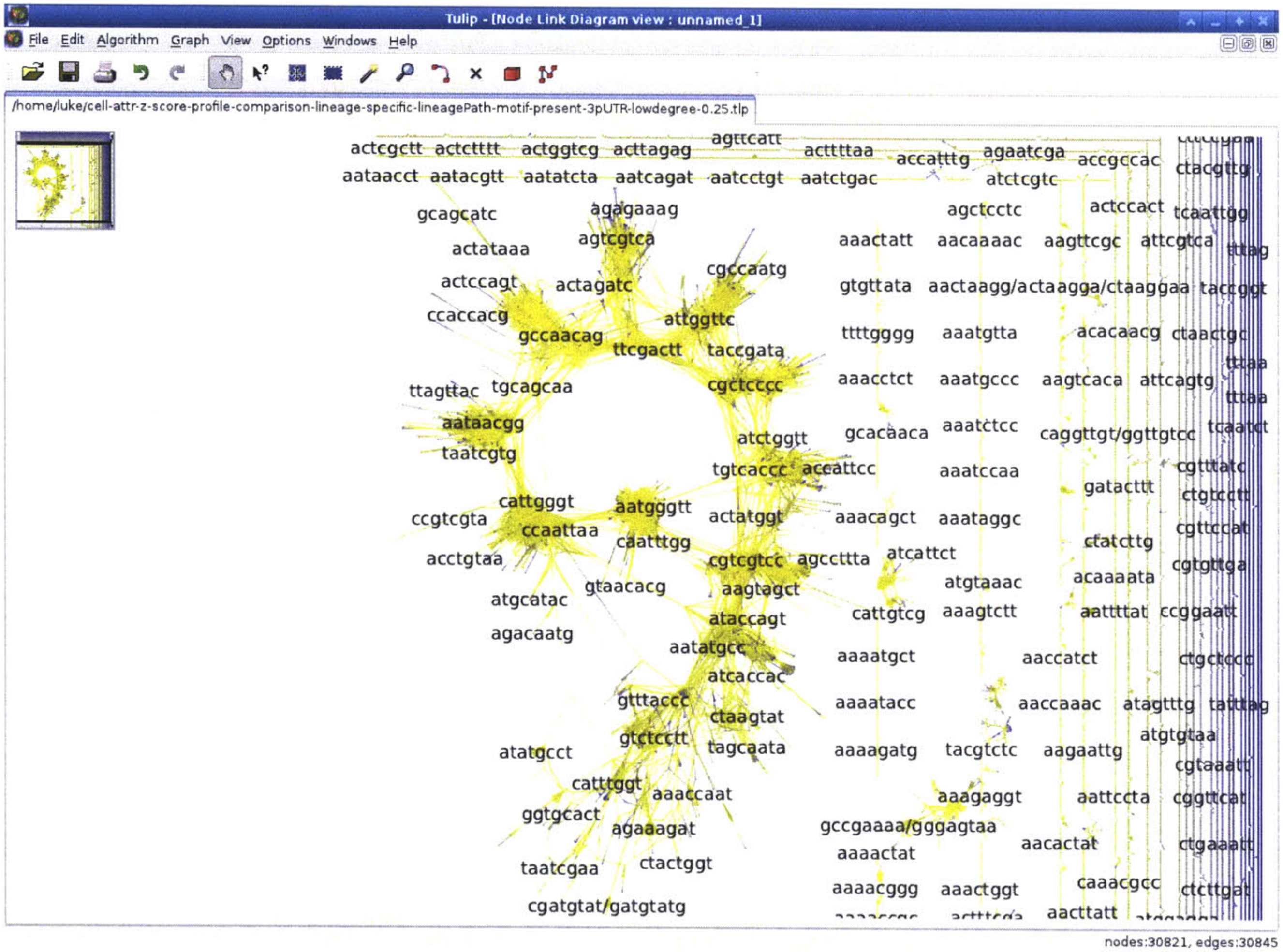
Changing the significance level at which edges in the all-pairs distance matrix were added to the graph before laying the graph out in Tulip, however, allowed for different types of substructure to be observed. With a low significance threshold, a fascinating pattern appeared in the phenotype influence network for motifs in the 3′ UTRs (Figure 29): a mostly-linear chain of clusters that were in most cases closely related only to the neighboring clusters in the chain at this significance level. This seems to indicate a smooth continuum of influence on phenotype, perhaps as a time series (similar to how the HOX genes work).

To try to understand the phenotypes described by each cluster, a tool was built that would allow the user to select a group of nodes (gene predicates) representing a cluster, and then average together the z-score profiles for each gene predicate, i.e. the z-scores obtained by testing any gene predicate in the cluster with a given cell attribute were all averaged together, producing an average z-score for each cell attribute. The resulting per-cluster profile of average z-score for each cell attribute could then be used to compare the effective aggregate phenotypic profile represented by each cluster. The cell attributes with the top average z-score are given below, with clusters numbered starting with zero in the middle and moving concentrically outwardly in clockwise order. The top attributes seem predominantly related to relatively early development and in many cases to neuronal development.

More work needs to be done to see if there is some sort of linear correlation between cluster number and the average cell birth time, average physical distance along the worm, or similar.

```
Cluster 0:

4.895   anat:ray neuron
4.895   cell-group:ray neurons
4.895   male-only:anat:ray neuron
4.895   male-only:cell-group:ray neurons
4.817   male-only:anat:sensory neuron
4.801   male-only:cell-group:neuron
4.730   anat:digestive tract
4.692   male-only:lineage-suffix-canonical:a
4.692   male-only:lineage-suffix:a
4.676   male-only:direction:a
4.535   anat:pharyngeal cell
4.535   anat:pharyngeal segment
4.535   anat:pharynx
4.489   anat:ray
4.489   cell-group:male sensory rays
4.489   male-only:anat:ray
4.489   male-only:cell-group:male sensory rays
4.458   male-only:anat:post-embryonic cell
```


Tulip - [Node Link Diagram view : unnamed_1]

/home/luke/cell-attr-z-score-profile-comparison-lineage-specific-lineagePath-motif-present-3pUTR-lowdegree-0.25.tlp

nodes:30821, edges:30845

File
Edit
Algorithm
Graph
View
Options
Windows
Help
actcgctt actctttt actggtcg acttagag agttcatt actttaa accatttg agaatcga accgccac ctacgttg
aataacct aatacgtt aatatcta aatcagat aatcctgt aatctgac atctcgtc
gcagcatc agagaaag agctcctc actccact tcaattgg
actataaa agtcgtca aaactatt aacaaaac aagttcgc attcgtca tttag
actccagt actagatc cgccaatg gtgttata aactaagg/actaagga/ctaaggaa taccggt
ccaccacg gccaacag attggttc ttttgggg aaatgtta acacaacg ctaactgc
ttcgactt taccgata tttaa
ttagttac tgcagcaa cgctcccc aaacctct aaatgccc aagtcaca attcagtg tttaa
aataacgg atctggtt gcacaaca aaatctcc caggttgt/ggttgtcc tcaatct
taatcgtg tgtcaccc accattcc aaatccaa cgtttatc
gatacttt ctgtcctt
cattgggt aatgggtt actatggt aaacagct aaataggc cgttccat
ccgtcgta ccaattaa caatttgg ctatcttg
acctgtaa cgtcgtcc agcctttta atcattct cgtgttga
gtaacacg aagtagct atgtaaac acaaaata
atgcatac ataccagt cattgtcg aaagtctt aattttat ccggaatt
agacaatg aatatgcc aaaatgct aaccatct ctgctccc
atcaccac
gtttaccc aaaatacc aaccaaac atagtttg tatttag
ctaagtat atgtgtaa
atatgcct gtctcctt tagcaata aaaagatg tacgtctc aagaattg cgtaaatt
catttggt aaaccaat aaagaggt aattccta cggttcat
ggtgcact agaaagat gccgaaaa/gggagtaa
aaaactat aacactat ctgaaatt
taatcgaa ctactggt aaaacggg aaactggt caaacgcc ctcttgat
cgatgtat/gatgtatg aacttatt

Figure 29: The 3′ UTR motif "phenotype influence network" with a low z-score edge-add threshold reveals a chain of clusters that are primarily closely related in phenotypic attributes to one or two linearly neighboring clusters. (Other motifs to the right and above were not joined to any or many other motifs at this significance level.)

```
4.412   male-only:anat:organ

Cluster 1:

9.043   anat:GABAergic neuron
4.069   life-stage:3-fold embryo
4.030   life-stage:comma embryo
4.020   life-stage:1.5-fold embryo
4.018   anat:Organ system
4.018   anat:digestive tract
4.006   life-stage:fully-elongated embryo
3.985   life-stage:2-fold embryo
3.894   program:differentiation
3.887   anat:pharyngeal nervous system
3.887   anat:pharyngeal neuron
3.799   life-stage:elongating embryo
3.778   is-leaf:one-or-both
3.650   is-leaf:both
3.630   anat:anterior pharyngeal ganglion (post)
3.614   anat:pharyngeal cell
3.614   anat:pharyngeal segment
3.614   anat:pharynx
3.570   anat:ciliated neuron

Cluster 2:

7.260   anat:GABAergic neuron
6.070   parent:lineage-prefix-canonical:Z.pap
6.070   parent:lineage-prefix:Z.pap
6.034   lineage-prefix-canonical:Z.pap
6.034   lineage-prefix:Z.pap
6.034   parent:cell-group:E lineage
4.601   cell-group:E lineage
4.456   daughter:lineage-prefix-canonical:Z.pappr
4.456   daughter:lineage-prefix:Z.pappr
4.440   lineage-prefix-canonical:Z.papp
4.440   lineage-prefix:Z.papp
4.424   daughter:lineage-prefix-canonical:Z.papp
4.424   daughter:lineage-prefix:Z.papp
4.160   anat:dopaminergic neuron
4.127   program:differentiation
4.032   daughter:lineage-prefix-canonical:Z.pap
4.032   daughter:lineage-prefix:Z.pap
3.980   anat:Organ system
3.857   cell-group:Q lineage

Cluster 3:

5.848   anat:GABAergic neuron
4.507   program:differentiation
```

```
4.256   is-leaf:both
4.175   anat:pharyngeal cell
4.175   anat:pharyngeal segment
4.175   anat:pharynx
4.123   anat:Organ system
3.931   anat:anterior pharyngeal ganglion (post)
3.927   life-stage:3-fold embryo
3.897   is-leaf:one-or-both
3.887   life-stage:fully-elongated embryo
3.886   anat:pharyngeal nervous system
3.886   anat:pharyngeal neuron
3.842   life-stage:2-fold embryo
3.833   life-stage:1.5-fold embryo
3.816   life-stage:comma embryo
3.725   cell-type:structural cell
3.725   male-only:cell-type:structural cell
3.725   male-only:lineage-suffix-canonical:app

Cluster 4:

5.379   anat:GABAergic neuron
5.152   anat:Organ system
4.879   program:differentiation
4.461   is-leaf:one-or-both
4.408   is-leaf:both
4.176   life-stage:fully-elongated embryo
4.154   life-stage:2-fold embryo
4.122   life-stage:1.5-fold embryo
3.955   life-stage:elongating embryo
3.793   life-stage:comma embryo
3.786   life-stage:3-fold embryo
3.733   is-leaf:terminal
3.538   parent:life-stage:late cleavage stage embryo
3.421   anat:body region
3.362   anat:hypodermal cell
3.362   anat:hypodermis
3.347   parent:life-stage:3-fold embryo
3.337   anat:ray neuron
3.337   cell-group:ray neurons

Cluster 5:

5.029   program:differentiation
4.690   anat:Organ system
4.482   cell-group:D lineage
4.465   lineage-prefix-canonical:Z.ppp
4.465   lineage-prefix:Z.ppp
4.441   is-leaf:both
4.406   anat:pharyngeal cell
4.406   anat:pharyngeal segment
```

```
4.406   anat:pharynx
4.241   life-stage:comma embryo
4.228   is-leaf:one-or-both
4.128   life-stage:3-fold embryo
4.061   life-stage:fully-elongated embryo
4.038   life-stage:2-fold embryo
4.000   parent:lineage-prefix-canonical:Z.ap~appa
4.000   parent:lineage-prefix:Z.ap~appa
3.995   anat:GABAergic neuron
3.982   lineage-prefix-canonical:Z.pppa
3.982   lineage-prefix:Z.pppa

Cluster 6:

5.309   program:differentiation
4.869   anat:Organ system
4.807   is-leaf:both
4.770   anat:GABAergic neuron
4.708   life-stage:3-fold embryo
4.635   life-stage:fully-elongated embryo
4.611   life-stage:2-fold embryo
4.536   is-leaf:one-or-both
4.453   life-stage:comma embryo
4.420   life-stage:1.5-fold embryo
4.385   life-stage:elongating embryo
4.256   anat:nerve ring neuron
3.925   anat:motor neuron
3.841   is-leaf:terminal
3.770   anat:male-specific
3.576   brief-id:ring motor neuron \/ interneuron
3.548   parent:lineage-prefix-canonical:Z.ap~appa
3.548   parent:lineage-prefix:Z.ap~appa
3.501   parent:life-stage:late cleavage stage embryo

Cluster 7:

5.282   anat:male-specific
4.890   anat:sensory neuron
4.852   program:differentiation
4.702   is-leaf:one-or-both
4.647   anat:Sex specific entity
4.626   is-leaf:both
4.525   brief-id:postembryonic blast cell
4.505   cell-group:post-embryonic blast cell
4.499   male-only:anat:sensory neuron
4.490   life-stage:elongating embryo
4.344   parent:male-only:lineage-suffix-canonical:p
4.344   parent:male-only:lineage-suffix:p
4.312   male-only:anat:organ
4.312   male-only:anat:sensillum
```

```
4.307   parent:male-only:direction:p
4.219   parent:male-only:anat:post-embryonic cell
4.219   parent:male-only:gender-dimorphism
4.219   parent:male-only:gender:male
4.219   parent:male-only:lineage-prefix-canonical:Z.a

Cluster 8:

5.807   program:differentiation
5.558   is-leaf:both
5.257   is-leaf:one-or-both
5.052   life-stage:2-fold embryo
5.013   life-stage:3-fold embryo
4.935   life-stage:fully-elongated embryo
4.822   life-stage:elongating embryo
4.793   is-leaf:terminal
4.706   life-stage:1.5-fold embryo
4.697   anat:Organ system
4.600   anat:male-specific
4.583   life-stage:comma embryo
4.288   parent:male-only:lineage-suffix-canonical:ap
4.288   parent:male-only:lineage-suffix:ap
4.230   cell-group:amphid neurons
4.192   parent:male-only:lineage-suffix-canonical:p
4.192   parent:male-only:lineage-suffix:p
4.153   parent:male-only:direction:p
4.063   cell-type:sensory cell

Cluster 9:

5.979   program:differentiation
5.349   is-leaf:both
5.131   is-leaf:one-or-both
4.922   anat:male-specific
4.748   life-stage:elongating embryo
4.734   life-stage:3-fold embryo
4.724   life-stage:2-fold embryo
4.664   is-leaf:terminal
4.641   life-stage:fully-elongated embryo
4.536   anat:Organ system
4.284   life-stage:1.5-fold embryo
4.194   life-stage:comma embryo
3.974   anat:GABAergic neuron
3.923   parent:life-stage:late cleavage stage embryo
3.906   male-only:cell-type:neuron
3.887   male-only:anat:neuron
3.838   parent:male-only:lineage-suffix-canonical:ap
3.838   parent:male-only:lineage-suffix:ap
3.827   parent:male-only:anat:post-embryonic cell
```

```
Cluster 10:

5.563   program:differentiation
5.014   is-leaf:one-or-both
4.929   is-leaf:both
4.856   life-stage:3-fold embryo
4.817   life-stage:fully-elongated embryo
4.791   life-stage:2-fold embryo
4.619   life-stage:1.5-fold embryo
4.558   life-stage:elongating embryo
4.331   anat:Organ system
4.244   life-stage:comma embryo
3.792   is-leaf:terminal
3.597   parent:life-stage:late cleavage stage embryo
3.525   lineage-prefix-canonical:Z.ap~ppppap
3.525   lineage-prefix:Z.ap~ppppap
3.483   anat:pharyngeal cell
3.483   anat:pharyngeal segment
3.483   anat:pharynx
3.483   parent:lineage-prefix-canonical:Z.ap~ppppa
3.483   parent:lineage-prefix:Z.ap~ppppa

Cluster 11:

6.336   program:differentiation
5.041   is-leaf:one-or-both
4.696   is-leaf:both
4.566   anat:Organ system
4.112   is-leaf:terminal
3.968   life-stage:fully-elongated embryo
3.960   life-stage:3-fold embryo
3.947   life-stage:2-fold embryo
3.838   life-stage:elongating embryo
3.836   anat:GABAergic neuron
3.758   cell-group:D lineage
3.734   lineage-prefix-canonical:Z.ppp
3.734   lineage-prefix:Z.ppp
3.481   life-stage:1.5-fold embryo
3.440   lineage-prefix-canonical:Z.pppa
3.440   lineage-prefix:Z.pppa
3.440   parent:cell-group:D lineage
3.440   parent:lineage-prefix-canonical:Z.ppp
3.440   parent:lineage-prefix:Z.ppp
```

11.5 BIOLOGICAL RELEVANCE OF RESULTS

The known functions local clusters of genes in a pairwise phenotype influence network (Figure 30) were compared in order to get a rough

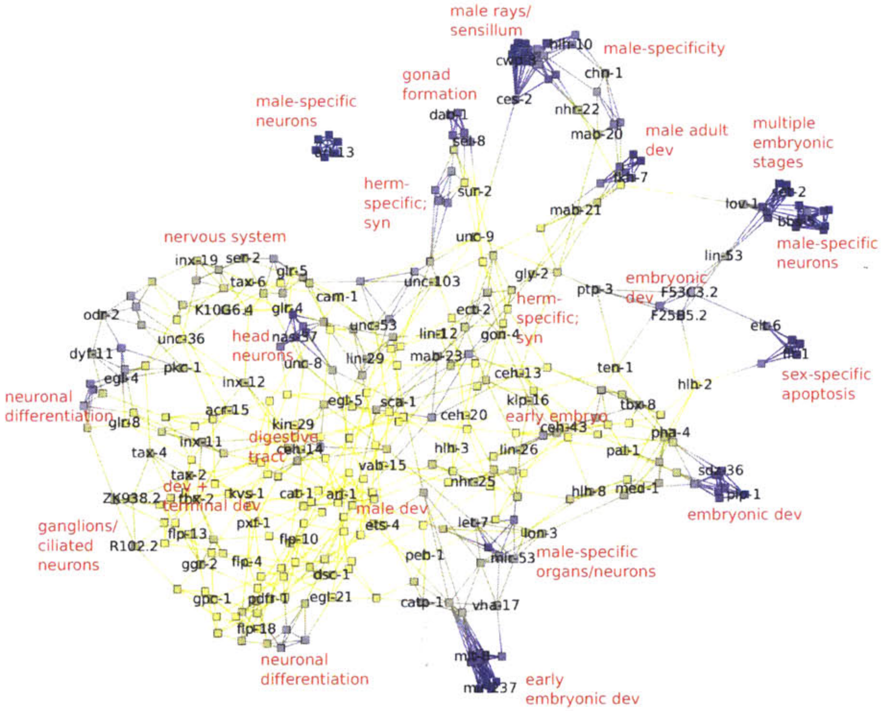


Figure 30: A pairwise phenotype influence network for genes, with some of the highest z-score correlated cell phenotypic traits manually labeled.

description of how the topology of the phenotype influence network corresponded with actual gene function. The result was interesting in that gender-specific and neurogenesis-specific gene functioning did appear localized in the graph, indicating the graph was serving the purpose of bringing together genes that are implicated in similar cell phenotypes.

A number of lists of genes of known common functional type (e.g. [15], [19]) were used to search the contingency test result file to determine cell phenotypic traits that were most strongly implicated for the expression of those gene sets. Preliminary results indicate a good correlation between known functions and high z-scores for corresponding cell-phenotypic traits, but full results are forthcoming and will be published separately.

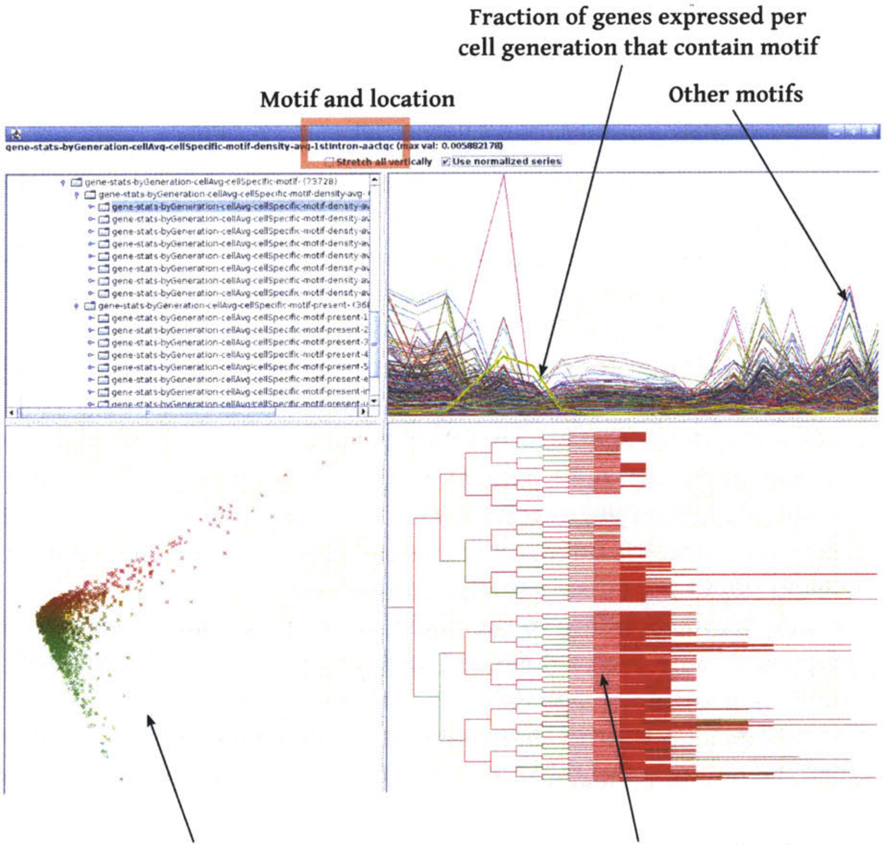
12 TOOL FOR VISUALIZATION OF GENE EXPRESSION PATTERNS ACROSS THE CELL PEDIGREE

Gene statistics (Section 9.2.1), prior to Otsu thresholding (Section 9.2.2) to turn them into gene predicates, may contain rich information that is useful in learning a significant amount about development, for example it may be useful to know if actual average intron length in active genes increases across the developmental timeline rather than just knowing that the proportion of introns increases that is categorized by Otsu threshold binarization as "long".

Figure 31 shows the tree visualization tool built for this purpose. A tree of all gene predicates is shown at top right (grouped hierarchically by common string prefix). The currently-selected node in the string prefix tree contains 4^6 descendant nodes, one corresponding to each 6-mer motif that fits the pattern "motif-density-avg-1stIntron-######"[1]. For each of the 4^6 nodes in the tree node selected in the top left pane, one line series is drawn in the top right pane (corresponding to the average value of that statistic across the developmental timeline, with the zygote on the lefthand size and the full adult worm on the righthand side, and one cell division depth per series value), and one point is drawn in the bottom left pane (corresponding to the 2D PCA projection of the graph series at top right, with each value used as one dimension for PCA, with red indicating a weighted average series value corresponding to early-stage development, i.e. a series that down-trends after early development, and green indicating a weighted average series value corresponding to late stage development, i.e. a graph that up-trends in the adult worm). The pedigree drawn at bottom right shows the particular cells in the pedigree in which the yellow-highlighted series are most strongly scored, with red for low average score and green for high average score across all genes in the pedigree at that point. (The horizontal axis of the pedigree chart at bottom left and the series plot at top right are the same, where the distance across the x-axis corresponds to the number of cell divisions since the zygote.) A series may be highlighted by moving the mouse over a series in the top right graph or over a point in the bottom left PCA plot.

This was developed as an exploratory tool, allowing large numbers of features to be visually examined at the same time, and may prove useful in finding genomic features that give rise to interesting developmental patterns.

[1] This particular screenshot was taken on a dataset built with 6-mer motifs as opposed to the 8-mers used elsewhere.



Fraction of genes expressed per cell generation that contain motif

Motif and location

Other motifs

gene-stats-byGeneration-cellAvg-cellSpecific-motif-density-avg-1stintron-aactqc (max val: 0.0058821 78)

□ Stretch all vertically ☑ Use normalized series

```
gene-stats-byGeneration-cellAvg-cellSpecific-motif- (73728)
  gene-stats-byGeneration-cellAvg-cellSpecific-motif-density-avg-
    gene-stats-byGeneration-cellAvg-cellSpecific-motif-density-av
    gene-stats-byGeneration-cellAvg-cellSpecific-motif-density-av
    gene-stats-byGeneration-cellAvg-cellSpecific-motif-density-av
    gene-stats-byGeneration-cellAvg-cellSpecific-motif-density-av
    gene-stats-byGeneration-cellAvg-cellSpecific-motif-density-av
    gene-stats-byGeneration-cellAvg-cellSpecific-motif-density-av
    gene-stats-byGeneration-cellAvg-cellSpecific-motif-density-av
    gene-stats-byGeneration-cellAvg-cellSpecific-motif-density-av
  gene-stats-byGeneration-cellAvg-cellSpecific-motif-present- (36
    gene-stats-byGeneration-cellAvg-cellSpecific-motif-present-1
    gene-stats-byGeneration-cellAvg-cellSpecific-motif-present-2
    gene-stats-byGeneration-cellAvg-cellSpecific-motif-present-3
    gene-stats-byGeneration-cellAvg-cellSpecific-motif-present-4
    gene-stats-byGeneration-cellAvg-cellSpecific-motif-present-5
    gene-stats-byGeneration-cellAvg-cellSpecific-motif-present-e
    gene-stats-byGeneration-cellAvg-cellSpecific-motif-present-in
    gene-stats-byGeneration-cellAvg-cellSpecific-motif-present-u
```

PCA of current motif series (treating each generation as one dimension). Color indicates mean generation of expression from early (red) to late (green) stage of development.

Expression distribution in cell pedigree. Green indicates a majority of genes active at cell contain the promoter motif; red indicates a lower proportion of active genes with the motif.

Figure 31: Screenshot of tree visualization tool

13

CONCLUSION

This thesis has presented a novel set of integrative genomics methods and data analysis techniques for building a model for *in silico* analysis of development in *C. elegans.* The highest-quality machine-readable and analytically-traversible cell pedigree for *C. elegans* that is so far available was produced by carefully merging available pedigrees into a new cell fate tree, and the anatomy ontology and gene association databases from WormBase.org were linked to this resource to yield the first whole-organism, single-cell-resolution gene expression map across all stages of development. We performed contingency testing between tens of thousands of cell phenotypic attributes and hundreds of thousands of thresholded gene statistics to produce a database that may be quickly queried to find the probable genetic bases of many cell phenotypic traits, and performed all-pairs analysis of pairs of cell attributes and pairs of gene attributes across this database to produce "phenotype influence networks", networks of genes and motifs that appear to co-influence phenotype. Finally A visualizer was introduced that allows for the simultaneous graphical display of a large number of gene expression statistics, permitting users to quickly visually identify interesting genomic features (including active genes, motifs, SNPs etc.) that appear to have unusual spatial distributions or temporal characteristics. The whole-organism expressome, related analysis tools and methodologies, and data visualization techniques together enable a new genre of direct *in silico* analysis of organism development.

BIBLIOGRAPHY

[1] Caenorhabditis elegans. *Wikipedia*. URL http://en.wikipedia.org/wiki/Caenorhabditis_elegans.

[2] WormBase. URL http://en.wikipedia.org/wiki/Wormbase.

[3] A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62, 1979. doi: {10.1109/TSMC.1979.4310076}.

[4] A. Agresti. A survey of exact inference for contingency tables. *Statistical Science*, 7(1):131–153, 1992. ISSN 0883-4237.

[5] S Brenner. The genetics of Caenorhabditis elegans. *Genetics*, 77(1): 71–94, 1974.

[6] Sydney Brenner. In the beginning was the worm. *Genetics*, 182(2): 413–5, 2009. doi: {10.1534/genetics.109.104976}.

[7] The WormBase Consortium. http://WormBase.org/.

[8] Ralph B. D'Agostino, Warren Chase, and Albert Belanger. The Appropriateness of Some Common Procedures for Testing the Equality of Two Independent Binomial Populations. *The American Statistician*, 42(3):198–202, August 1998. doi: {10.2307/2685002}.

[9] H. L. de Jong. Genetic Determinism: How Not to Interpret Behavioral Genetics. *Theory and Psychology*, 10(5):615, 2000. doi: {10.1177/0959354300105003}.

[10] U Deppe, E Schierenberg, T Cole, C Krieg, D Schmitt, B Yoder, and G von Ehrenstein. Cell lineages of the embryo of the nematode Caenorhabditis elegans. *Proc. Natl. Acad. Sci. U.S.A.*, 75(1):376–80, 1978.

[11] The C. elegans Sequencing Consortium. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. *Science*, 282(5396):2012, 1998. doi: {10.1126/science.282.5396.2012}.

[12] David S Fay. The cell cycle and development: lessons from C. elegans. *Semin. Cell Dev. Biol.*, 16(3):397–406, 2005. doi: {10.1016/j.semcdb.2005.02.002}.

[13] Shiri Freilich, Tim Massingham, Sumit Bhattacharyya, Hannes Ponsting, Paul A Lyons, Tom C Freeman, and Janet M Thornton. Relationship between the tissue-specificity of mouse gene expression and the

evolutionary origin and function of the proteins. *Genome Biol.*, 6(7): R56, 2005. doi: {10.1186/gb-2005-6-7-r56}.

[14] David B Goldstein. Common genetic variation and human traits. *N. Engl. J. Med.*, 360(17):1696–8, 2009. doi: {10.1056/NEJMp0806284}.

[15] E. S. Haag. The evolution of nematode sex determination: C. elegans as a reference point for comparative biology. *WormBook*, 10, 2005.

[16] A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, 24(2):8–12, 2009. ISSN 1541-1672.

[17] John Hardy and Andrew Singleton. Genomewide association studies and human disease. *N. Engl. J. Med.*, 360(17):1759–68, 2009. doi: {10.1056/NEJMra0808700}.

[18] Joel N Hirschhorn. Genomewide association studies–illuminating biologic pathways. *N. Engl. J. Med.*, 360(17):1699–701, 2009. doi: {10.1056/NEJMp0808934}.

[19] O. Hobert. Neurogenesis in the nematode Caenorhabditis elegans. *WormBook: the online review of* C. *elegans biology*, page 1, 2010.

[20] H. R. Horvitz. Worms, Life and Death. *Nobel Lecture*, 2002. URL `http://nobelprize.org/nobel_prizes/medicine/laureates/2002/horvitz-lecture.html`.

[21] Sanger Institute. C. elegans genome sequence. URL `http://www.sanger.ac.uk/Projects/C_elegans/Genomic_Sequence.shtml`.

[22] E. F. Keller, H. E. Longino, and J. M. Smith. Rethinking the meaning of genetic determinism. *The Tanner Lectures on Human Values*, February 1993.

[23] Judith Kimble and David Hirsch. The postembryonic cell lineages of the hermaphrodite and male gonads in Caenorhabditis elegans. *Developmental Biology*, 70(2):396, 1979. doi: {10.1016/0012-1606(79)90035-6}.

[24] Peter Kraft and David J Hunter. Genetic risk prediction—are we there yet? *N. Engl. J. Med.*, 360(17):1701–3, 2009. doi: {10.1056/NEJMp0810107}.

[25] Raymond Y N Lee and Paul W Sternberg. Building a cell and anatomy ontology of Caenorhabditis elegans. *Comp. Funct. Genomics*, 4(1):121–6, 2003. doi: {10.1002/cfg.248}.

[26] Frank Lin and William W. Cohen. Power Iteration Clustering. In *Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel*, 2010.

[27] Xiao Liu, Fuhui Long, Hanchuan Peng, Sarah J Aerni, Min Jiang, Adolfo Sánchez-Blanco, John I Murray, Elicia Preston, Barbara Mericle, Serafim Batzoglou, Eugene W Myers, and Stuart K Kim. Analysis of cell fate from single-cell gene expression profiles in C. elegans. *Cell*, 139(3):623–33, 2009. doi: {10.1016/j.cell.2009.08.044}.

[28] Steen Magnussen. An algorithm for generating positively correlated Beta-distributed random variables with known marginal distributions and a specified correlation. *Computational Statistics and Data Analysis*, 46(2):397–406, 2004. ISSN 0167-9473. doi: {DOI:10.1016/S0167-9473(03)00169-5}. URL http://www.sciencedirect.com/science/article/B6V8V-4961NM2-2/2/2be87b359cb3c31e9276fbe947788e3e.

[29] T. A. Manolio. Genomewide association studies and assessment of the risk of disease. *N Engl J Med*, 363(2):166–176, 2010.

[30] J Reboul, P Vaglio, N Tzellas, N Thierry-Mieg, T Moore, C Jackson, T Shin-i, Y Kohara, D Thierry-Mieg, J Thierry-Mieg, H Lee, J Hitti, L Doucette-Stamm, J L Hartley, G F Temple, M A Brasch, J Vandenhaute, P E Lamesch, D E Hill, and M Vidal. Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in C. elegans. *Nat. Genet.*, 27(3):332–6, 2001. doi: {10.1038/85913}.

[31] J E Sulston and S Brenner. The DNA of Caenorhabditis elegans. *Genetics*, 77(1):95–104, 1974.

[32] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[33] Nicholas Wade. Genes Show Limited Value in Predicting Diseases. *New York Times*, page A1, April 2009. URL http://www.nytimes.com/2009/04/16/health/research/16gene.html.

[34] E. Wigner. The unreasonable effectiveness of mathematics in the natural sciences. *Communications on Pure and Applied Mathematics*, 13 (1):1–14, 1960.