A Data Mining Approach for Detecting Evolutionary Divergence in Transcriptomic Data

by

Owen Zeno Woody

A thesis
presented by the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Biology

Waterloo, Ontario, Canada, 2019

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	Dr. Teresa Crease Professor, Department of Integrative Biology, University of Guelph
Supervisor(s)	Dr. Brendan J. McConkey Associate Professor, Department of Biology, University of Waterloo
Internal Member	Dr. Kirsten M. Müller Professor, Department of Biology, University of Waterloo
Internal-external Member	Dr. Dan Brown Professor, Cheriton School of Computer Science, University of Waterloo
Other Member(s)	Dr. Josh D. Neufeld Professor, Department of Biology, University of Waterloo

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

It has become common to produce genome sequences for organisms of scientific or popular interest. Although these genome projects provide insight into the gene and protein complements of a species including their evolutionary relationships, it remains challenging to determine gene regulatory behavior from genome sequence alone. It has also become common to produce "expression atlas" transcriptomic data sets. These atlases employ high-throughput transcript assays to survey an assortment of tissues, developmental states, and responses to stimuli that each may individually elicit or inhibit the transcription of genes.

Although genomic and transcriptomic data sets are both routinely collected, they are seldom analyzed in tandem. Here I present a novel approach to combining these complementary data with a software package called BranchOut. BranchOut uses genomic information to construct gene family phylogenies, and then attempts to map gene expression activity onto this phylogeny to allow estimation of ancestral expression states. This allows the identification of specific innovations due to gene duplications that resulted in fundamental diversification in the roles of otherwise closely related genes.

As a proof of concept, the BranchOut technique is first applied to a tangible small-scale example in *Apis mellifera*. Subsequently, the power of BranchOut to analyze complete genomes is shown for two mammalian genomes, *Sus scrofa* and *Bos taurus*. The transcriptomic data sets for these two mammals employ microarray and RNAseq platforms, respectively, for expression analysis, demonstrating BranchOut's applicability to both future and historic expression atlases. Potential refinements to the approach are also discussed.

Acknowledgements

I would like to thank my supervisor, Brendan McConkey, for his patience, kindness and insight throughout the preparation of this work.

I'd also like to thank my committee members for their suggestions and support over many years: Dan Brown, Kirsten Müller, and Josh Neufeld. It was thanks to your mentorship that this was all possible.

I am very grateful for the support I received through both NSERC and OGS for this project.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction – Sporks, Soups and Sausages

The work presented in this thesis addresses a specific aspect of evolutionary biology: the **evolution** of **novel function** at the **gene family level**. Biological organisms are capable of remarkably complex activities, and all this functionality is somehow encoded in the organism's genomic complement. In order to discuss the evolution of function – broadly, how an organism can become capable of "new" activities – it is important first to establish a working understanding of a few key concepts. First, I will describe the features and activities of a gene that make it a functional unit in the cell. Next, I will cover the means by which new genes can arise, and what it means for a gene to be part of a family. Lastly, I will describe what I define as a "novel function", and how novelty can be introduced into the genome through the process of gene origination and duplication.

1.1 Defining the gene as a functional unit

This work will focus on genes as fundamental units of inheritance and evolution. There are some intriguing examples of inheritable adaptation that do not require changes at the genetic level (RNA interference, for example (Spracklin, Fields et al. 2017)), but they will lie outside the scope of this discussion. The "Central Dogma of Biology" – that DNA genes encode (messenger) RNA molecules, which are in turn translated to form proteins – remains a powerful explanatory tool, and my research will make a similar assumption about the central role genes play in biological activity and inheritance.

A gene is a string of information in the genome with two main parts: the **coding sequence** and its **regulatory control** (MacCarthy, Bergman 2007). The **coding sequence** of a gene provides a blueprint for the construction of an active biomolecule – typically a protein, but sometimes an RNA molecule with catalytic properties (ribozyme). Because the number of building blocks for these biomolecules is limited (4 nucleotides or approximately 20 amino acids), the sequence and arrangement of these components is the fundamental characteristic that determines a biomolecule's role. Consequently, when the coding sequence is the focus, the Central Dogma can be restated as follows: (nucleotide) sequence directs (biomolecule) structure, which in turn directs function. Changes to the coding sequence can fundamentally alter the structure of the encoded biomolecule and may impact the biological capability of the resulting product.

A first draft for the definition of "gene function" might stop there. For example, consider the following statement:

> Gene Z is a protein coding gene. When translated, the resultant Z protein is a biomolecule that disrupts the integrity of cell membranes.

This action is a consequence of the protein's structure, which is a consequence of its sequence. We could then infer that other genes with similar sequences encode proteins with similar functions, and we could come up with a category of "membrane disruption" genes. This strategy is commonly applied to transfer annotation from well-studied genes to genes with unknown function (Sjolander 2004).

This definition of function is workable, but it is perhaps too coarse. Suppose we collected several genes that seemed to carry out the "membrane disruption" function and then studied them more closely. Here are two hypothetical variations on the statement that could ensue:

> "Gene Z1 is a protein coding gene. It is translated specifically during mitosis, and the resultant Z1 protein is a biomolecule that disrupts the integrity of cell membranes."

Contrast the previous statement with the following:

> "Gene Z2 is a protein coding gene. It is only expressed in the venom gland, and the resultant Z2 protein is a biomolecule that disrupts the integrity of cell membranes."

Although the **fundamental activity of the biomolecule** is consistent for these two cases, the extra details included now suggests two very different **functions**. The added information indicated "when" or "where" the gene was expressed. This fundamentally alters our perspective on the encoded protein – the former, being involved in mitosis, is likely to be ancient, essential, and precise. The latter, by contrast, suggests a more recent origin in a complex multicellular organism, and an encoded protein that will be made most effective by being active under a very broad range of conditions and on a variety of substrates. These disparate roles suggest that while the fundamental activity may be the same, there may be subtleties to the structure of the two proteins that encourage specificity versus generality.

For this reason the **function** of a gene is also investigated and defined by its **regulatory control**. If structure encodes the "what" and "how" of a gene, regulatory

control dictates the "when", "where", and "why" (and "how much"). The regulatory control is also a consequence of nucleotide sequence, but most of the key elements lie outside of the coding sequence elsewhere in the genome. Peripheral genomic structure can alter the availability of a gene for transcription either directly (*i.e.* via cis-regulatory elements) or indirectly (*i.e.* via genome folding and/or gene accessibility in proximity to histones (Schoenfelder, Fraser 2019)).

In summary, for the purposes of this work, **gene function** is an aggregate term. The term is used to describe the activity of a gene not only in terms of its (physical, biomechanical) interactions, but also in terms of its cellular regulation and control. Changes to either of these two aspects of a gene may affect the gene's function, and in cases where one aspect remains unchanged in a retained duplicate, the other is often different (Semon, Wolfe 2008, Arnaiz, Gout et al. 2010, Ren, Fiers et al. 2005). Hence, these two aspects of gene function are fundamentally linked, and both are subject to evolution through random mutation. They differ, however, in their **ease of accessibility to researchers**, which has often resulted in the two aspects being examined separately (Daugaard, Rohde et al. 2007). This has made the study of gene function, and the evolution thereof, a much more difficult task, and it is the specific goal of this work to reunite these two aspects into one framework.

1.2 Where do genes come from, and what does it mean to be in a member of a "gene family"?

Genomes expand or contract their genetic complement through a process of gene birth and death. The vast majority of gene births are the result of a duplication event, meaning that each gene has a "parent" from which it is derived, and this

ancestry that can be traced back to a distant precursor. Various duplication mechanisms are reviewed in (Woody, Doxey et al. 2008).

Despite the gene's unclear origin, the subsequent flow of genetic information through the tree of life is well understood. Following a duplication event, both the template (original) gene and its copy will be subject to random mutation and selection, producing two genome elements that are often broadly similar but distinct. As these two genes are subsequently and serially duplicated, a collection of genes that share a common ancestor will end up co-existing within the genome. This collection of related genes is commonly referred to as a **gene family**. Some well-studied gene families include the globin gene family (Storz, Opazo et al. 2011), and the FOXO family of transcription factors (Wang, M., Zhang et al. 2009). When the focus of a study is a particular species, the extent of the family is limited to the member genes that occur in its genome. It is also possible to consider the breadth of a family across multiple species (Khaitovich, Weiss et al. 2004, Whitehead, Crawford 2006a, Whitehead, Crawford 2006b, Nuzhdin, Wayne et al. 2004, Huminiecki, Wolfe 2004), but in this case the processes of gene duplication and speciation (analogously, species duplication) are conflated and analyses are more challenging.

The mechanism of duplication and the ultimate fate of duplicated genes are both intimately related to functional evolution. To help make this process tangible, I will use an analogy based on a human utensil: the "spork".

1.3 The story of the "spork", a utensil that was not meant to be

As a thought experiment, consider the "spoon" as a functional protein encoded by a gene. The spoon is ideally suited for soups, where it can serve as a

ladle, and is capable of assisting the ingestion of small food chunks. The spoon is not well-suited for cutting or skewering. Broadly, its function can be described as "the spoon is produced at meal times (regulatory control), and aids the ingestion of liquids (activity)". We can think of the regulatory control broadly as the **conditions for expression**.

Consider a genome carrying only a single "spoon" gene (and no other utensils). There are two ways in which the single gene could mutate: through alterations to either its regulatory control, or to its activity. Many alterations would be obviously bad. For example, introducing a pore in the well of the spoon would compromise its function and be detrimental to the host. Similarly, disruption of regulatory control could make the spoon unavailable as a "response" to mealtime, which is also a negative result. For single-copy genes, the majority of these negative changes are eventually purged by natural selection (Cvijovic, Good et al. 2018). If a gene duplicate undergoes this loss of function and becomes essentially inactive, the process is called **nonfunctionalization**.

There can be changes that are neither harmful nor beneficial. If a change to regulatory control caused spoons to be also available during times of play, the net result in fitness could be near neutral. Similarly, mild alterations to the stem/grip of the spoon are likely to be neutral.

Lastly, as a hypothetical example of a beneficial, "gain-of-function" mutation, consider the addition of prongs to the end of the spoon, resulting in a "spork". This is an alteration to the structure that impacts the activity of the utensil in two ways. The spork has acquired the ability to skewer, which increases its potential

functionality as an ingestion tool (e.g., now the host can eat sausages). However, the prongs on the front introduce an inter-prong space that is incapable of holding liquid; the 'original' function from the spoon has been compromised (but not lost). These function-altering events are particularly exciting to researchers, and the process of acquiring a new function this way is referred to as **neofunctionalization** (Tirosh, Barkai 2007).

Depending on the nutrient sources available, selective pressure could promote one function over the other: if there are no soups, the prongs could eventually overtake the well, and if there are no sausages, the prongs are likely to be lost to restore the integrity of the well. If both resources are available, the hybrid spork may be the best option.

Next, let us re-examine these three outcomes, but introduce the new element of gene duplication events. For now, we'll assume that duplicates retain both the regulatory control and activity of their templates (though this is not always the case in practice).

First, let us return to the debilitating mutation – the introduction of a pore in the well of the spoon – but let's precede this event with a gene duplication event. Now the host genome (which was previously viable with a single copy of the spoon gene) has two copies, but one of these copies has acquired a debilitating mutation that renders it incapable of its ancestral purpose. In this case, the pre-duplication state is restored – the genome returns to its ancestral configuration of having a single functional spoon – and the nonfunctionalized duplicate is no longer subject to natural selection. It is free (and likely) to acquire further debilitating mutations, and

will eventually lose the signals that drove its regulatory control. It will be no longer expressed, no longer functional, and in all senses dead. As it turns out, most duplicates are adaptively deleterious (Qian, Liao et al. 2010) and this process of duplicate loss is widespread in genomes, making nonfunctionalization (or non-viability) the most common outcome of gene duplication events. However, in the absence of deleterious outcomes, gene duplicates tend to persist for some time even in the absence of functional diversification (Skamnioti, Furlong et al. 2008).

Next, let us examine the case of the spork, and consider the potential outcomes of a duplicated spork gene. Prior to duplication, the spork served a dual role – both as skewer and ladle – and mutations that enhanced one aspect generally came at the expense of the other. Following duplication, this is no longer the case – if one "spork" subsequently loses its prongs and reverts to the ancestral "spoon" phenotype, the skewer-capable spork remains. Moreover, the remaining spork is under far less pressure to maintain the integrity of its well – the tines could indeed extend all the way to the shaft of the spork, producing a fork, without limiting the host's access to soup nutrients (thanks to its perfectly functional spoon). This process, termed **subfunctionalization**, best encapsulates the central role gene duplication can play in the evolution of function by providing a means of escaping from adaptive conflict between two competing goals (Barkman, Zhang 2009, Johnson, D. A., Thomas 2007, Freilich, Massingham et al. 2006).

This example focused on the evolution of gene activity, but gene regulation can also play a complementary role in the process of gene duplication and retention. While our example gene has a fairly broad regulatory condition – make the utensil

when food is available – one could imagine a finer, more modular set of rules producing this behavior. For example, this overall behavior could be the sum of three individual regulatory controllers acting in unison – one directing breakfast expression, one directing lunch expression, and one for dinner (the biological implementations of gene regulation are often similarly modular). Were a spork gene to duplicate, initially retaining identical regulatory control, the loss of the "make sporks available for breakfast" module on one copy would not be deleterious to the host genome – the other spork gene would still direct the production of the utensil. But suppose this intact spork gene then subsequently lost its "lunch" module – again, not immediately deleterious, as our "not for breakfasts" spork will still guide production of sporks for lunchtime. However, the status quo has now been changed in a subtle way – both copies of the ancestral template are now necessary, as they have non-overlapping sets of regulatory instruction. Loss of one gene will result in loss of spork production for at least one meal, so now both genes are subject to natural selection. This diversification of function into two (or more) broad categories within a single gene family has been noted in various studies (Viaene, Vekemans et al. 2010, Goettel, Messing 2010, Jarinova, Hatch et al. 2008, Johnson, D. A., Thomas 2007, Des Marais, Rausher 2008, Wang, R., Chong et al. 2006), and mining for these events on a large scale is one of the primary goals of the presented work.

We can summarize the ideas illustrated by this analogy of functional evolution as follows:

- Gene duplication events often precede mutations in gene activity

- ❖ Duplicates that differ in regulation are exposed to different environments but may or may not be under selection for similar activity
- ❖ Duplicates with identical activity are unlikely to have identical regulation (for long)

1.4 Mechanisms and scales of duplication events

There are several mechanisms by which gene duplicates can arise. Although duplication and mutation are indispensable aspects of evolution, they nonetheless are the consequence of errors in the (DNA) replication and maintenance process. Any introduced deviations in the DNA genome are propagated to cellular offspring following cell division. Variations that occur in the germline have the ability to impact the entire genome of progeny.

There are a few mechanisms by which the ancestral template DNA genome can become altered to introduce gene duplicates. These mechanisms differ in their scope dramatically. At one extreme, the duplicate gene may involve a single new short coding strand being inserted into a more-or-less random location in the genome. Intermediate-scale duplications can result from unequal crossing-over during meiosis (Redon, Ishikawa et al. 2006). Tandem duplications, where sequence regions containing a gene are duplicated in series, produce genes that tend to have conserved function (Wang, Z., Dong et al. 2010). At the other end of the spectrum, the duplication could affect the *entire genome*, duplicating every chromosome and all genes therein (Van de Peer, Maere et al. 2009, Blanc, Wolfe 2004, Vision, Brown et al. 2000). Whole genome duplications seem particularly

common in various plant species (Lockton, Gaut 2005). Duplications of individual chromosomes also happen, but they tend to be more destabilizing.

Large-scale duplications will duplicate the coding strand as well as any preceding and proceeding DNA strings. Any "nearby" regulatory control sequences will be included in the duplication (Cannon, Mitra et al. 2004). This can preserve almost all regulatory control, with a few exceptions (for example, those that depend on the folding of the chromosome to bring sequences into proximity). Depending on the fidelity and extent of the copied region, duplicates may begin by sharing both activity and regulatory control (Woody, Doxey et al. 2008).

There are, however, mechanisms that preserve the coding sequence but little to none of the regulatory control (Brenner, Johnson et al. 2000). As an extreme example, consider the process of reverse transcription. Reverse transcription involves editing the DNA genome by using an RNA template, a 'violation' of the Central Dogma that is employed by some viruses. The point of insertion used by DNA-editing enzymes does not depend on the RNA's origin. When duplicating a gene, these enzymes may introduce duplicates that are distantly separated, or on completely different chromosomes. Depending on the processing stage of the RNA molecule, very little of the original (transcriptional) regulatory sequences may remain (Wang, Z., Dong et al. 2010). Mature mRNA molecules, for example, will have completed intron/exon processing and will only contain regulatory sequences that guide translation. As a result, it is possible for some duplicates to retain few, if any, of their former regulatory control sequences. Regulation of these new genes will be completely dependent on whatever sequences happen to be located near

their point of insertion (Richardson, Salvador-Palomeque et al. 2014, Ohshima 2013).

For the purposes of this study, we can summarize the possible outcomes of duplication in terms of their impact on gene function. While the activity (coding sequence) is almost always preserved, the extent to which regulatory control is preserved varies from full conservation to none (Guan, Dunham et al. 2007). In the case where regulatory control is not preserved, the gene will be completely dependent on pre-existing regulatory sequences that occur near the point of duplicate insertion, and will not necessarily be "silenced".

1.5 Protein biomechanical functions are determined by coding sequences, which mutate slowly and follow patterns

Genes are rarely produced *de-novo*, so most extant genes must either be descended directly from a precursor gene or be related to another gene via duplication (Ruiz-Orera, Hernandez-Rodriguez et al. 2015). Changes to coding sequence may introduce mutations, and if these mutations lead to adaptive characteristics the gene will retain these characteristics that distinguish them from their ancestor. To some extent, this process follows a set of rules that can be summarized into a model of sequence evolution. These models are used in the process of sequence alignment, which in turn is used to measure the degree of sequence dissimilarity (and therefore evolutionary distance) between sequences. There are a number of ways to use sequence distance to support hypotheses of **homology**, the property of common sequence ancestry (Saripella, Sonnhammer et al. 2016).

Most models of gene families depend on sequence similarity directly. Since sequence also determines function (through the central dogma), gene activity and ancestry are closely linked. In other words, genes in the same family typically expected to have similar (biochemical) gene activity by definition (Rajashekar, Samson et al. 2007). Changes to coding sequences can lead to changes in biomolecular function, allowing functional adaptation and diversification to occur (Turunen, Seelke et al. 2009, Panchin, Gelfand et al. 2010).

1.6 Regulatory control is determined by non-coding sequence elements, which mutate quickly and unpredictably

Members of a gene family do not necessarily share common regulatory control. Although some newly duplicated genes will retain their regulatory control sequences as well, others may not, and very closely related genes (new duplicates) may not share any regulatory similarity whatsoever, as with reverse transcription. **Regulatory elements** can be discrete and observable (Huang, H. Y., Chien et al. 2006). For example, the binding site for a transcription factor may have a recognizable sequence motif. However, it is not as easy to model mutations to these elements, as changes to a single element of a regulatory element may be sufficient to turn the module "on" or "off". Gain or loss of a single regulatory module may not be detrimental to the gene's function, so these regulatory domains are subject to more "churn", and may appear and disappear rapidly on evolutionary time scales (Casneuf, De Bodt et al. 2006a, Chain, Evans 2006). A change in regulatory control may itself be sufficient to generate evolution of function, should changes to either

dosage or regulatory triggers generate adaptive cellular behaviors (Des Marais, Rausher 2008).

In addition to these discrete elements, other genome features affecting regulation are less tangible. **Spacing**, broadly encapsulating the physical (3d) location of regulatory elements relative to genes, can affect gene regulation in a way that is not readily quantifiable by observing a complete genome sequence alone (Tsankov, Thompson et al. 2010). This spacing could also change in unpredictable ways as a result of sequence mutations.

In summary, the regulation of a gene does not evolve according to a pattern that can be easily modeled. Gene regulation is unreliable for determining common ancestry, and similarity of regulation is not indicative of similar ancestry. Nonetheless, regulation stands out as one of the two key aspects of gene function, and gene regulation is intimately connected to gene activity. Changes in one aspect can be supportive of changes in the other.

1.7 Genome sequences offer complete access to gene complement, and can be used directly to infer gene family membership and activity

Genome sequences are quickly becoming routine, lower-cost, "first-step" features in studies of biological organisms. Genome sequences expose the complete gene complement of an organism to study. Since sequence directly determines gene activity and can be used to hypothesize homology, a single complete genome sequence is sufficient to provide extensive information about which gene families are present, and the broad biomechanical activities these genes perform. This

stands in contrast to studies of gene activity, where each additional assay can yield additional insight into regulatory control.

1.8 Gene regulation is complex but can be observed empirically

Genome sequences reveal the positions of some recognizable regulatory elements, but their impact on nearby genes is unclear, and often conflated with a number of other influencing factors that are not easily quantified (Beer, Tavazoie 2004, Jarinova, Hatch et al. 2008, McClintock, Kheirbek et al. 2002, Wang, D., Sung et al. 2007). For example, the 3D-archictecture of the genome itself may restrict or prohibit access to regulatory regions (Tsankov, Thompson et al. 2010). Conversely, distantly separated regulatory elements may be brought into close proximity through contortions around histones and other structure-imposing elements (Huang, P., Keller et al. 2017). The field of systems biology tries to address this regulatory complexity through mathematical modeling, but the accuracy of these models depends on extensive research (Jarinova, Hatch et al. 2008, Beer, Tavazoie 2004). At present, it is not feasible to rely on such models for genome-wide prediction of gene regulation (Comelli, Gonzalez 2009, Akitaya, Tsumoto et al. 2003).

As a result of this challenge, the present state of the art in studying gene regulation is to use high-throughput assays to observe the state of the cell directly (Ranz, Machado 2006). The goal of these assays is to provide information about the activity of a cell in response to a stimulus. In essence, the cell's instruction set (i.e., when, where, why genes should be expressed) can be studied by observing the outcome of these instructions and then trying to infer what the guidelines were.

Under the assumption that genes are only active (or activated) when they are needed, this information can additionally be used to infer the functional purpose of genes. For example, if a gene with unknown function is expressed at the same time (and under the same circumstances) as a suite of genes with a known association with a particular pathway, it may be reasonable to assume that this gene is also a component of this broad system response (Xing, Ouyang et al. 2007).

In summary, because gene expression cannot be readily predicted from genome sequence information alone, assays are used to query the cellular state to essentially witness genes in action. A gene's activation under a particular set of stimuli can be used as circumstantial evidence of gene function, and through this information a gene's regulatory control may be inferred.

1.9 The evolution of gene regulation is similarly unpredictable

The regulatory control of a gene can be modeled in a number of ways. At the biological level, gene function is the outcome of several discrete, countable events. Specifically, within a given time period a gene within a cell is transcribed a number of times. These transcripts are translated a number of times, and then a number of the protein products are relocated. Each of these events has a "how many?" associated with it that will influence the transcription/translation rate, and thus the effect of the gene in the cell.

Each of these events is directly or indirectly driven by instructions encoded in the genome, and are thus subject to the process of evolution (Drummond, Bloom et al. 2005). To model this process, simplifying assumptions are often made. Two broad approaches are possible.

1.9.1 Binary models of gene function

A considerably simplified model of gene activity would be to treat it like a light switch with two fundamental states: on (active) or off (inactive). If there were some threshold copy number that could be taken as evidence of gene activation, the state of each gene could be summarized in terms of whether it was active or not. This model could be expanded to include additional discrete states as appropriate.

While this model may seem overly simplistic, there is some empirical evidence suggesting that in many cases it is a reasonable choice. Because gene activity levels (this process of multiple discrete copying steps) are subject to evolution, it is biologically useful for the cell to have some buffering capacity so that small variations in gene product level do not impact cell function. This can be seen as an example of canalization. Canalization broadly refers to the phenomenon whereby a genome is able to evoke the same phenotype in spite of environmental changes and stresses (Sato 2018). If we consider the gene's activity state to be the phenotype, then canalization suggests this activation state should be somewhat robust to small changes both in the expression environment and in the genome itself.

1.9.2 Continuous models of gene function

For modeling reasons, it is sometimes useful to consider gene expression as a continuous value. This allows gene expression evolution to be estimated using a number of standard scientific models. Small changes in gene activity can be modeled using Brownian motion as a baseline, for example. This provides a "null hypothesis" background for evolution. If a gene's activity is not under selective

pressure, it should drift around some idealized value "at random". If a series of changes result in a shift away from this former ideal, to the point where the Brownian motion model seems improbable, one could instead infer that the activity of the gene had changed (Gu, X. 2004).

1.10 Mechanisms of gene duplication

There are several biological mechanisms that can result in the duplication of partial or complete genes. This is important for the evolution of novel function for one key reason: the extent to which regulatory control is duplicated differs considerably from one mechanism to the next. The following sections describe some of the mechanisms and the extent to which they preserve regulatory control.

1.10.1 Whole genome duplication

Failure to properly separate sets of chromosomes during meiosis can result in germ cells with an additional copy of the complete genome. As long as "gene stoichiometry" is preserved (Coate, Song et al. 2016), these gametes can produce viable offspring (Birchler, Veitia 2019). These events are hypothesized to have happened multiple times within several evolutionary lineages, and are believed to be important drivers of evolution (MacKintosh, Ferrier 2017). Each duplicate gene retains its full regulatory control structure, but is under weak selective pressure (and thus at some liberty to explore functional or non-functional intermediates).

As is the case for duplication events involving a single gene, the loss of function of one gene from the duplicate pair often restores the genome to a state similar to its pre-duplication ancestor. For whole genome duplications, however, this process is occurring on a genome-wide scale. A large number of duplicate pairs

are ultimately reverted to a "single-copy" state as genes are rendered dysfunctional by deleterious mutations. The rate at which this process happens varies with the organism's capacity and tolerance for excess genomic baggage. Plants, for example, retain relics of genomic duplications over large scales of time (Qiao, Li et al. 2019).

1.10.2 RNA-mediated gene duplication

There are genetic mechanisms that allow RNA molecules to be reverse-transcribed as insertions back into the genome (Schacherer, Tourrette et al. 2004). When this machinery uses an mRNA transcript (in some state of processing) as a template to edit the genome, it can result in the creation of a duplicate sharing the general character of the gene from which the mRNA template was derived. Duplicates that arise in this way share a few interesting characteristics owing to their origin as an mRNA molecule. mRNA molecules do not typically retain any of the genomic features required to attract, initiate, or terminate transcription. They may also exclude introns, depending on the extent of cellular processing applied to the mRNA molecule. In the case of genes with multiple splicing arrangements, this also means the newly introduced duplicate will be "frozen" with a particular selection and arrangement of exons.

1.11 Gene expression is quantitative but its outcome is non-linear

Most genes must be translated into protein before they can affect the state of the cell. Accordingly, measuring the presence and quantity of a protein product is often a reasonable proxy for the progenitor gene's involvement or activation. However, protein abundance is historically less experimentally accessible than RNA abundance (Zhao, Fang et al. 2017), so mRNA transcript abundance is often studied

in its place. Because mRNA transcripts represent an intermediate stage in the protein manufacturing process, it is often assumed that mRNA abundance can serve as a proxy for gene activity.

There are reasons to doubt this assumption. In particular, mRNA transcripts are not translated once and only once. They may not be translated at all, and a single mRNA molecule can be translated multiple times to generate several copies of the encoded protein (Vogel, Marcotte 2012). It is reasonable to assume that an increase in transcript production would result in an increase in protein production, but the rate and ratio is not uniform and can be gene-specific (Vogel, Marcotte 2012, Maier, Guell et al. 2009, Mehdi, Patrick et al. 2014).

There have been studies comparing mRNA and protein abundance from the same samples. Many showed low correlations in abundance (Nie, Wu et al. 2006, Mehdi, Patrick et al. 2014) . These studies are challenging because mRNA and protein abundance are evaluated with completely different experimental platforms, and it is extremely difficult to attribute differences in measured quantity to true biological effects alone.

In summary, it is worth keeping in mind that studies using protein abundance offer a distinct and often complementary perspective on experiments focusing on mRNA abundance.

1.12 Both expression and sequence are necessary but neither is sufficient

Because it must be the case that all the guidelines for when, where, and why to express a gene are encoded (in some form) in the genome, algorithms for predicting gene expression have yet to achieve the accuracy required to render

expression assays obsolete. For the foreseeable future, a complete and accurate genome sequence will still be insufficient to describe all aspects of gene activity.

Acknowledging this, several recent genome projects have included a complementary expression atlas. These atlas resources are constructed by conducting tens to hundreds of gene expression assays. These assays are also dependent on the genome sequence; microarray probes and RNAseq use the genome sequence as a template, and sequencing-based approaches can only confirm the existence of genes expressed in a particular sample (with most samples being unlikely to involve every gene). To get a complete picture of the inner workings of an organism, both aspects – genome sequence and the gene-expression assays – must be included.

Because it has become standard practice to parcel genome sequences and expression atlas projects together, there are relatively few tools available that exploit both these channels of information in a complementary way. The software presented in this thesis research, "BranchOut", provides a novel approach for combining this information.

1.13 Goals of this dissertation

My dissertation is intended to help researchers who seek meaning from large-scale quantitative biomolecular assays by helping them mine them for new hypotheses to explore. While gene expression will often be used as an example, the concepts apply to protein abundance (and other biomolecular markers) as well. The tools and algorithms I discuss provide novel perspectives on gene expression projects by making creative use of existing complementary information.

Following a further background discussion of preprocessing methodology included in Chapter 2, Chapters 3 and 4 will lay the groundwork for BranchOut, a suite of analysis tools prepared in R that integrate the results of a large-scale (gene) expression project with phylogenetic analyses. This integrated perspective will help identify changes in gene function that have been of pivotal importance to the evolution of functional diversity at the level of a gene family. Chapter 3 describes specific software implementation and interface details. Chapter 4 shows the BranchOut concept applied to a small-scale case study. Chapter 5 applies BranchOut to the entire *Sus scrofa* genome using a microarray-based expression atlas. Chapter 6 applies BranchOut to the entire *Bos taurus* genome using a transcriptomic sequencing-based expression atlas. Lastly, Chapter 7 will cover closing thoughts and future directions for similar work.

Chapter 2: Preprocessing Methodology for Genomic and Transcriptomic Data

This chapter will focus on some fundamental issues common to quantitative biomolecule analysis platforms. Because some decisions made in BranchOut (and elsewhere in this dissertation) will make assumptions about the way data has been pre-processed and standardized, I feel it worthwhile to discuss these issues in a second preliminary chapter here.

More specifically, this chapter describes the decisions and options involved with preparing the two streams of data that are broadly required for my software to work. The first stream is composed of sequence data, including gene sequences, gene families, and gene phylogenetic trees. The second stream is composed of gene activity records, which will be measures of gene expression behavior from microarrays or RNA sequencing data for most of this study.

A discussion of various prior efforts to study the evolution of gene expression can be found in a previous review article (Woody, Doxey et al. 2008).

2.1 Preparing gene expression data

Both microarray data and high throughput sequencing data have sources of noise and error that need to be accounted for through normalization. For microarrays, for example, the robust multiarray average (RMA) algorithm (Bolstad, Irizarry et al. 2003) is commonly used for probe signal processing. RMA has a history as a reliable and publishable standard with well-known strengths and

weaknesses (Irizarry, Wu et al. 2006). Among the strengths, RMA is very quick to compute, which is a useful feature for genome level investigations of gene function. Foremost among the weaknesses is a tendency to introduce bias to extreme signals (Irizarry, Wu et al. 2006). A popular variant of RMA, GC-RMA (Seo, Hoffman 2006), attempts to additionally model GC-content as a biasing factor in probe signal intensity. GC-RMA would also be a reasonable choice, but recent microarrays have been designed with GC-content compensation in mind, causing GC-RMA to expend effort fitting a model parameter that ideally should have greatly reduced influence in modern arrays.

Depending on the ultimate goal, it is worth considering whether to use expression values directly (i.e. a numerical expression value) , a binary presence/absence call (Yang, Su et al. 2005), or a summary that indicates a general trend in expression behavior (e.g. membership to an expression cluster, on/off designations, presence/absence calls) (Doxey, Yaish et al. 2007a, Sahoo, Dill et al. 2007, Woody, Doxey et al. 2008).

2.2 Associating gene sequences into families

In order to establish an evolutionary history, genes are first organized into families. Ideally, a gene family consists of genes that share a common ancestor, with all family members being the consequence of gene duplication events. Typically, a proxy for family membership based on sequence similarity is used. One such family assignment scheme is maintained by the PFAM database, which assigns all sequences one (or more) identifiers based on conserved domains (Punta, Coggill et

al. 2012). PFAM uses a hidden Markov model-based approach (Potter, Luciani et al. 2018) to determine family affiliation, but this is just one of many possible sequence similarity indices that could be used.

2.3 Methods for inferring family relationships (Phylogenetic trees)

Once family membership is determined, the next step is to construct a hypothetical evolutionary history depicting the relatedness of sequences from the family. This process has two main steps: multiple sequence alignment, a process which attempts to align conserved regions of family members, and then phylogenetic tree construction, which attempts to transform this sequence alignment-based relatedness into a phylogenetic (typically bifurcating) tree. The question of which algorithms work best for each of these stages is still open to debate. However, the two dominant methods for phylogeny estimation, maximum likelihood and Bayesian inference, produce reasonably similar hypotheses when provided with high-quality data (Anisimova, Gil et al. 2011).

2.4 Models for transforming expression into hypotheses of function

There are a number of reasonable ways to interpret gene expression values (be they sequencing abundance or microarray signal intensity) as evidence of gene activity. The direct use of raw abundance counts has the greatest potential accuracy, but relies on a detailed understanding of systems biology for the organism in question to be easy to interpret.

Gene expression atlas projects often make use of a cartoon diagram of the organism. This diagram provides a visual map that can be annotated with the

expression activity of a gene. For example, the BAR resource center hosted by the University of Toronto provides a tool for *Arabidopsis* gene expression (Toufighi, Brady et al. 2005). Once a gene is selected, diagrams depicting the various organs of the plant (as well as several developmental stages) are colored using heatmap colors to show relative expression levels of the selected gene scaled against that gene's mean expression level across all tissues. This provides a quick and effective means of identifying tissues where the gene in question is under active regulation, and whether that regulation is promoting or inhibiting production of the associated protein product.

This approach works well for individual genes but cannot be readily scaled up to include a set of genes. To compare the expression behavior of two related genes, the user would have to toggle between two different visualizations and identify (color) differences. Although this could readily identify a number of tissues where the behavior may differ, this approach requires a fair amount of user interaction and interpretation.

It is also worth asking whether a gene's expression behavior is best interpreted in isolation. A more systems-aware perspective might strive to include a number of related genes in the study of gene expression. There are two obvious candidates for relatedness in this context.

First, the relatedness could be determined by shared involvement and/or participation in a biological system. For example, genes affiliated with the same biological process (as determined through gene ontology notation or otherwise) could show similar patterns of activation/inhibition across an expression atlas

(Ashburner, Ball et al. 2000, Gu, X., Zhang et al. 2005). This provides some context for interpreting the behavior of a single poorly understood gene. If its behavior is unknown but well-correlated with a set of genes involved in a specific biological process, it seems reasonable as a first guess to assume a similar involvement for the gene in question (Gu, Z., Nicolae et al. 2002). On the other hand, if a set of genes involved in a process show highly correlated behavior with the exception of one member whose behavior is largely consistent but different in a few conditions, this could be indicative of genetic novelty (in the form of a new function or re-purposed role) (Casneuf, De Bodt et al. 2006b, Ha, Li et al. 2007). For collections of genes that share some but not all of their expression characteristics, an approach more akin to biclustering, *i.e.*, identifying related genes and expression triggers together, may be more appropriate (Prelic, Bleuler et al. 2006).

2.5 Inferring ancestral expression/function

Another category of summarization techniques employ some form of clustering to associate genes with similar expression profiles. Consider a set of genes selected based on either shared ancestry or inferred participation in some biological role. If the expression information for these genes were arranged in a matrix, with one row per gene, one column per expression assay, one could treat each row of the matrix as a profile, and then define a **distance measure** that could be used to establish pair-wise measures of similarity with each other profile included in the set. Clustering methods take these dissimilarity scores and attempt to group profiles with a high degree of similarity into a cluster. These clusters have

a tangible interpretation – if the distance measure reports that two expression profiles are quite similar, then the genes that generated these profiles are regulated in a similar fashion. This would presumably mean they share regulatory control signals, and are turned on and turned off by similar stimuli. From here, it is a relatively small step to suggest that the genes may be involved in a similar pathway and be deployed to accomplish a similar function (Okamura, Obayashi et al. 2015). Genes that fall into different clusters, by contrast, would exceed a threshold dissimilarity measure and have profiles that do not suggest a shared function. There are a few variants on clustering that differ in the distance metric used and/or in the type of clustering applied. Approaches that do not employ clusters are also available but will not be pursued here (Gu, X. 2004, Gu, X., Su 2007, Oakley, Gu et al. 2005, Rossnes, Eidhammer et al. 2005).

2.6 Decisions made for BranchOut Preprocessing

For microarray expression data, BranchOut uses RMA normalized scores. For sequencing-based expression, BranchOut uses normalized read counts produced by the "cuffnorm" program (Trapnell, Williams et al. 2010). In determining gene family membership, BranchOut makes use of an identifier mapping table provided by the PIRSF website (Nikolskaya, Arighi et al. 2007), which includes PFAM assignments together with a number of other transcript, gene, and protein identifiers. In some circumstances, PFAM identifiers were available through other more closely related annotation sources, but ultimately PFAM identifiers were the criteria by which gene family membership was determined.

Following member identification, each gene family was then subjected to multiple sequence alignment by MUSCLE (Edgar 2004) then used to construct a tree using the PhyML (Guindon, Dufayard et al. 2010, Guindon, Gascuel 2003) software package with default settings. Although PhyML was the default for most large-scale analyses, early versions of BranchOut used trees produced by MrBayes instead, and ultimately either tool is acceptable as a source of trees. One tree was generated per gene family.

BranchOut employs a novel scheme for converting expression assay measurements into clusters denoting gene activity. For each gene family, a clustering approach is used on a tissue-by-tissue basis to determine which genes seem to show elevated expression when compared to other members of the family. By applying this process sequentially to each individual tissue, it is possible to identify specific gene family/tissue divergences in behavior that might otherwise be overlooked if expression behavior were clustered based on expression across all tissue samples simultaneously (a common approach with its own merits (Doxey, Yaish et al. 2007b)).

The process for assigning expression categories is detailed in Chapter 3. In brief, the categorization of normalized expression values into expression states is done by the MCLUST (Scrucca, Fop et al. 2016) software package. MCLUST treats a set of values as having been drawn from a mixture of an (unknown or specifiable) number of normal distributions, and uses a machine learning approach to assign values into their most probable parent distributions. There are a number of

parameters to the MCLUST software that are of particular relevance to this study. For example, by fixing the number of distributions at 2, the implied biological states are “low” and “high”. Fixing the number of distributions at 3 would naturally encourage “low”, “medium” and “high” configurations.

Reconstruction of ancestral states was conducted using the “Analyses of Phylogenetics and Evolution” (APE) package (Paradis, Claude et al. 2004a) for R (R Development Core Team 2010). A maximum likelihood criterion was used where ancestral states were inferred based on descendant nodes only.

Chapter 3: BranchOut Software Specifications

This chapter provides describes the development of the BranchOut software, which decisions were made, and focus exposition on the statistics, summarizations, and visualizations developed specifically for this software.

The software used to implement my approach is called "BranchOut" to reflect both the tree-based nature of the approach and the focus of the software on identifying subsets of the tree (ideally monophyletic subclades) that seem to have distinguished themselves from the remainder of the family by adopting a new function within the host organism. BranchOut is a collection of scripts for use in the R software (R Development Core Team 2010). BranchOut has several dependencies, most of which are other packages available either through R's public package repository or through the Bioconductor suite of bioinformatics-oriented R packages (Gentleman, Carey et al. 2004) owned and maintained outside the primary R database (Bioconductor release 3.2). Some additional software packages were used for multiple sequence alignment and phylogenetic tree estimation, but these software packages reflected particular analysis decisions and could be swapped out as necessary depending on the user's preferences.

3.1 Outline of software

The BranchOut software can be considered as the sum total of three separate components: **Input**, **Processing**, and **Output**. **Input**, the first component, deals with reading in, organizing, and pre-processing the raw biological assay metrics

being used. This first component is the least rigid by design, because users will be "jumping in" to the software at sometimes different starting points. The design philosophy for this component was to encompass a very broad approach (described below) but the software is not overly laborious to customize for small-scale applications (like the one shown in Chapter 4).

BranchOut assumes that the user is interested in a gene expression atlas that was completed on a single assay platform (by default, a microarray platform is assumed). The platform specification is important because it provides a guideline by which the genes to be examined can be filtered; genes with no associated expression data cannot contribute informatively to any immediate depictions of gene family function or any reconstructions thereof.

Following the selection of a platform, the Bioconductor package associated with that platform is accessed and current, researcher-curated associations between probe sequences and genome targets are collected into an association table. This is typically done at the "normalized probe-set summary" stage.

3.2 Gathering sequence information

Microarray sample data is typically stored in a file format that contains only a bare minimum amount of annotation information. To associate each probe-set record with an associated host gene (and its many database identifiers), an additional annotation file or package is required. R-compatible microarray annotation files are routinely produced by the Bioconductor team (Gentleman, Carey et al. 2004). BranchOut uses these annotation files to establish a mapping

between probe-sets and single genome features. Specifically, Unigene accessions, which try to associate a single label with each individual gene (Benson, Cavanaugh et al. 2017), are used in an effort to maximize the number of one-to-one matchings between probe-sets and PFAM (Punta, Coggill et al. 2012) annotation records. In circumstances where multiple probe sets map to the same Unigene identifier, BranchOut arbitrarily selects one of the two probe sets. This decision is potentially controversial, but there is often little reason to trust one probe set over another. A mixed-measures approach, like averaging the expression summaries for the probe sets, might make better use of the available data but risks making the aggregate gene metric as (un)reliable as the most poorly designed probe set.

Once a table of gene-relatable expression measures has been collected, the next task is to assign genes to gene families. Although the idea of a gene family is well-founded, the task of assigning family membership is complicated by a number of factors. Just as this project is focused on finding genes with similar ancestry with different functions, there are many cases of genes with similar functions but different ancestry. The task of assigning a gene to a family involves making a hypothesis about its ancestry based on the sequence of the gene and this sequence's relatedness to other believed members of the family. These hypotheses are typically based on some form of sequence alignment. For example, a multiple sequence alignment of known family members could be compared to a second multiple sequence alignment (MSA) that includes the gene to be classified, and some measure of distance could be used to gauge whether the candidate gene appears to be a good fit.

There are a number of databases compiling these hypotheses of ancestry. The default used by BranchOut is PFAM, a database organized by gene family that reports all sequences that associate with a family irrespective of their species of origin. The entirety of the database can be downloaded as a plain text file (ftp://ftp.pir.georgetown.edu/databases/pirsf/).

3.3 Processing Details

To make things easier for BranchOut, I have prepared a helper script in R that reads the large database text file (>2GB) and returns only gene sequences that are related to an input species. This intermediate file is available for a small number of model organisms on PFAM's website, but there is no webtool for focusing on a specific organism.

The sequences (and families) present in this filtered text database are then compared to the list of sequences available on the analysis platform (microarray) being used. BranchOut will then generate one intermediate sequence file per gene family containing the sequence information for all members of the family that are present in both the gene family database and analysis platform.

Because BranchOut will want to reconstruct gene activity records on a phylogenetic tree, two further intermediate preprocessing steps are required. To obtain phylogenetic trees that reflect the gene- (and platform)-availability of gene family members, each family sequence input file is used to generate a multiple sequence alignment (MSA), and each of these is in turn used to generate a gene-family phylogenetic tree. By default, BranchOut invokes the standalone executable

form of the multiple sequence alignment program MUSCLE (Edgar 2004) to construct the multiple sequence alignments using default parameters. These multiple sequence alignments are then used as input (unedited and as-is) in the phylogenetic tree estimation software PhyML (version 3), which constructs hypothetical phylogenetic trees using the maximum likelihood reconstruction framework (Guindon, Dufayard et al. 2010, Guindon, Gascuel 2003). Again, for pragmatic reasons, the default program settings were used here as well.

Next, BranchOut internally creates an object for each gene family composed of its component sequences, multiple sequence alignment, phylogenetic tree, and associated expression values. If there are replicates for any tissues, a separate guide file can be used to direct BranchOut to take means of replicates from the same tissue to obtain a summary expression measure for each tissue/condition. Following log-transformation (base 2) and row-standardization, gene expression levels should be centered around zero. This means an expression value of 0 corresponds to the average expression level for the gene across all the conditions being examined. In order to make it somewhat easier to determine whether assigned expression states correspond to "activation" or "inactivation", median subtraction is used in place of mean subtraction. Median subtraction is less affected by outliers than mean subtraction. Additionally, for mean subtraction, if half the conditions had the gene "off" state and half had the gene "on" state, the mean expression level would lie in-between and a corrected value of 0 would correspond with an expression state that is not ever observed in practice.

Values different from the 0 reference value correspond to upregulation or downregulation relative to that specific gene's median across tissues. For purposes of exposition, let us introduce arbitrary positive and negative thresholds at +1 and -1 (so, keeping the logarithm base 2 transformation in mind, 2-fold up and 2-fold down from the median, respectively). These can correspond to new states: "upregulated relative to the average" and "downregulated relative to the average". Gene expression values that surpass these thresholds would be assigned "upregulated" and "downregulated" status, respectively, to be contrasted with the "neutral" state assigned to values sufficiently close to 0.

It is worth noting that, because expression values are typically log-transformed, positive or negative expression scores represent fold changes up or down over the average, respectively. Suppose a gene were being transcribed at a rate that generated 100 units of activity on an assay. If this activity level were shifted down by 100 units (to zero), the resulting log difference would be negative infinity. An increase of 100 units would result in a log (base 2) fold change of +1. In practice, assay values near zero are not encountered in microarray studies, but they are common in RNA sequencing platforms. In the latter case, expression values of 0 may need special care, such as special assignment of an alternative minimum value that is within the domain of the logarithm function.

Next, for each of set of expression records associated with a gene family, MCLUST (Scrucca, Fop et al. 2016) is invoked once per tissue/condition to assign the expression values into categories. There are a couple of parameters to the MCLUST software that can be toggled within BranchOut that affect the number of

modes MCLUST can assign (and how likely it is to do so). The ideal values for these two parameters are unlikely to be consistent across all gene families being studied, so an option to toggle these values is available. Furthermore, the software can be set to adaptively toggle these values itself if it detects a poor reconstruction.

Following the assignment of expression values to categories, the APE (Paradis, Claude et al. 2004b) package is used to invoke a general parsimony-based ancestral character estimation protocol to infer the "cluster membership" state of ancestral nodes in the tree (where each cluster membership transition is treated as an equally likely event). At this point the processing of the gene family is done and the program switches to output.

To provide a rudimentary indicator of the "surprise value", or deviation from expected values based on a random model, of each hypothesized state reconstruction, a simple score is calculated using a method based on reshuffling. Using the MCLUST-assigned expression category labels as input, BranchOut first performs a traditional reconstruction and counts the number of category transitions that occur for the true, observed state of labels on the tree. To then gauge how uncommon this number of state changes is (given the number of states and their distribution on the true tree), BranchOut shuffles the labels on the leaves and then repeats the reconstruction step using these labels as an alternative input. As before, the number of state transitions required is counted and recorded. This process is then repeated a large number of times, generating a distribution of reconstruction scores based on the reshuffled labels. BranchOut then produces a score by taking

the ratio of the average number of state changes in the reshuffled trees to the number of state changes in the true tree.

Since the objective of the BranchOut software is to locate gene families that appear to have members involved in new function(s) (against a background of a shared common ancestral function), reconstructions based on the actual expression categories should require comparatively fewer state transitions compared to reconstructions based on shuffled states. Thus, the BranchOut score should be high for families where the number of required state changes in lower than chance.

3.4 Output

For each gene family, a set of output files is produced in a folder named after the family identifier. This set is composed of four primary separate PDF files, as follows, where the gene family name appears in place of <family_name>:

<family_name>_cluster assignments heat map.pdf

This PDF is a single image showing the heatmap of expression values based on median-standardized expression values only. Each column corresponds to a gene in the gene family, and each row corresponds to a sample (condition/microarray/tissue/stimulus). Colors map the magnitude of the expression assay measurements relative to the average, with increasingly vibrant blue colors assigned to expression groupings below the average and increasingly vibrant yellow colors assigned to groups with above average expression. This perspective is useful for determining whether a gene family is active for the

conditions being examined, and it indicates sub-groups that show coordinated expression behavior. Figure 3.1 shows an example heatmap.

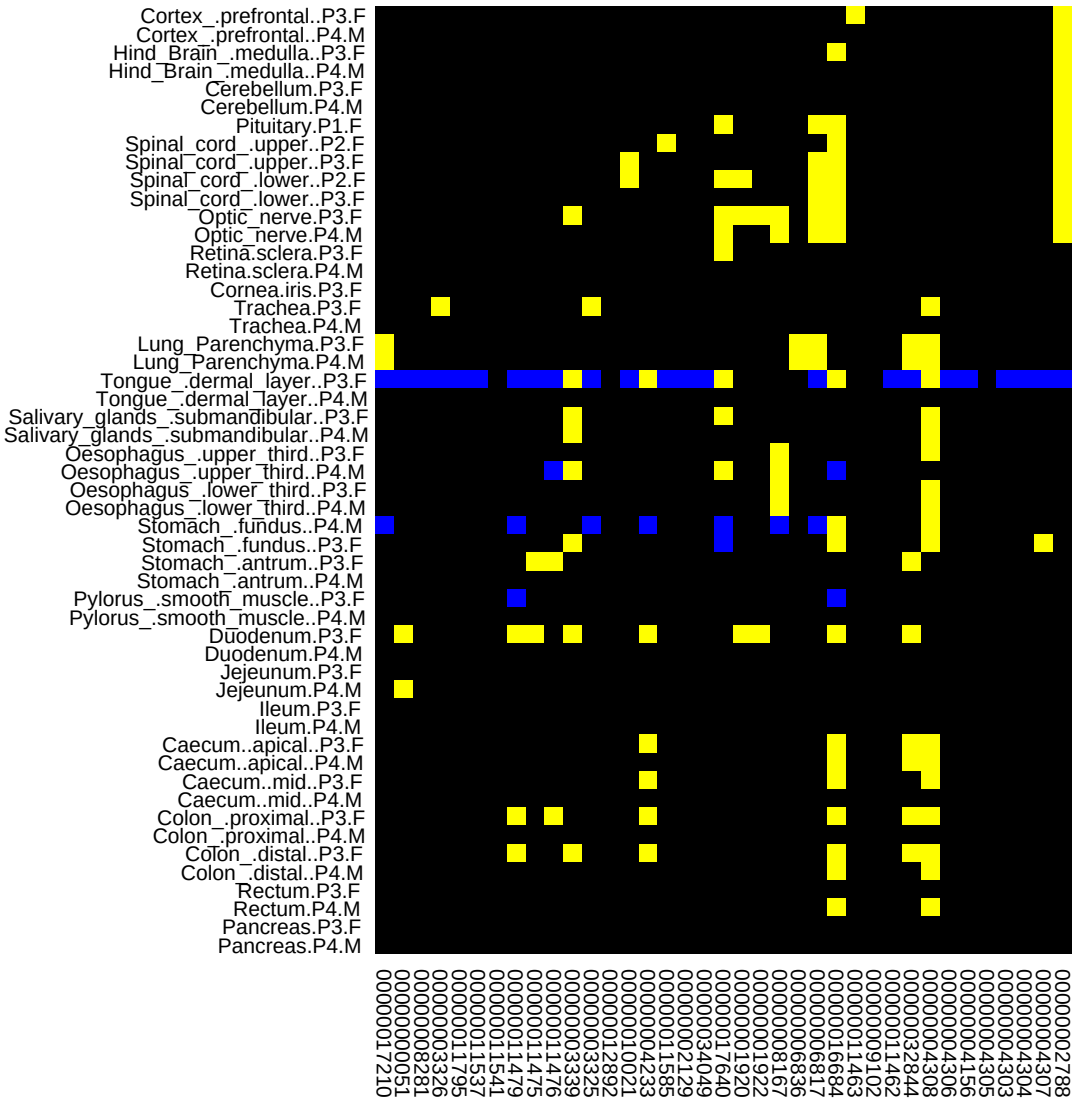


Figure 3.1: An example of an "all conditions" heatmap from BranchOut. The set of expression assays (corresponding to tissues) are displayed along the vertical axis. Transcripts from the selected gene family are displayed along the horizontal axis. Blue colors correspond to low/reduced expression cluster membership. Black (and dark colors) are used for near-average expression membership, and increasingly yellow colors are used for high expression group membership. Example taken from *Sus scrofa* results (see chapter 5). Transcript identifiers have been truncated to omit the prefix "ENSSSCT".

<family_name>_major_reconstruction_categories.pdf

For each set of hypothesized ancestral functions, the change in behavior occurring on a branch can be interpreted as a change in category (0 – no change, -1 – category dropped, +1 – category raised). Thus, the history of expression changes for a given expression category can be summarized as a vector of -1s, 0s, and 1s as long as it has at least two expression categories. For this output file, the number of times each pattern of expression changes appears is recorded. For each pattern, starting from the most frequently occurring to the least, the associated heatmap is shown (depicting the changes at the expression level), the tissues/conditions/samples for which this pattern was observed are reported, and the individual reconstructions are listed subsequently. The underlying hypothesis behind these graphs is that a user may be interested to see which reconstruction patterns apply to which categories. Figure 3.2 shows an example reconstruction block and the associated subset of the heatmap.

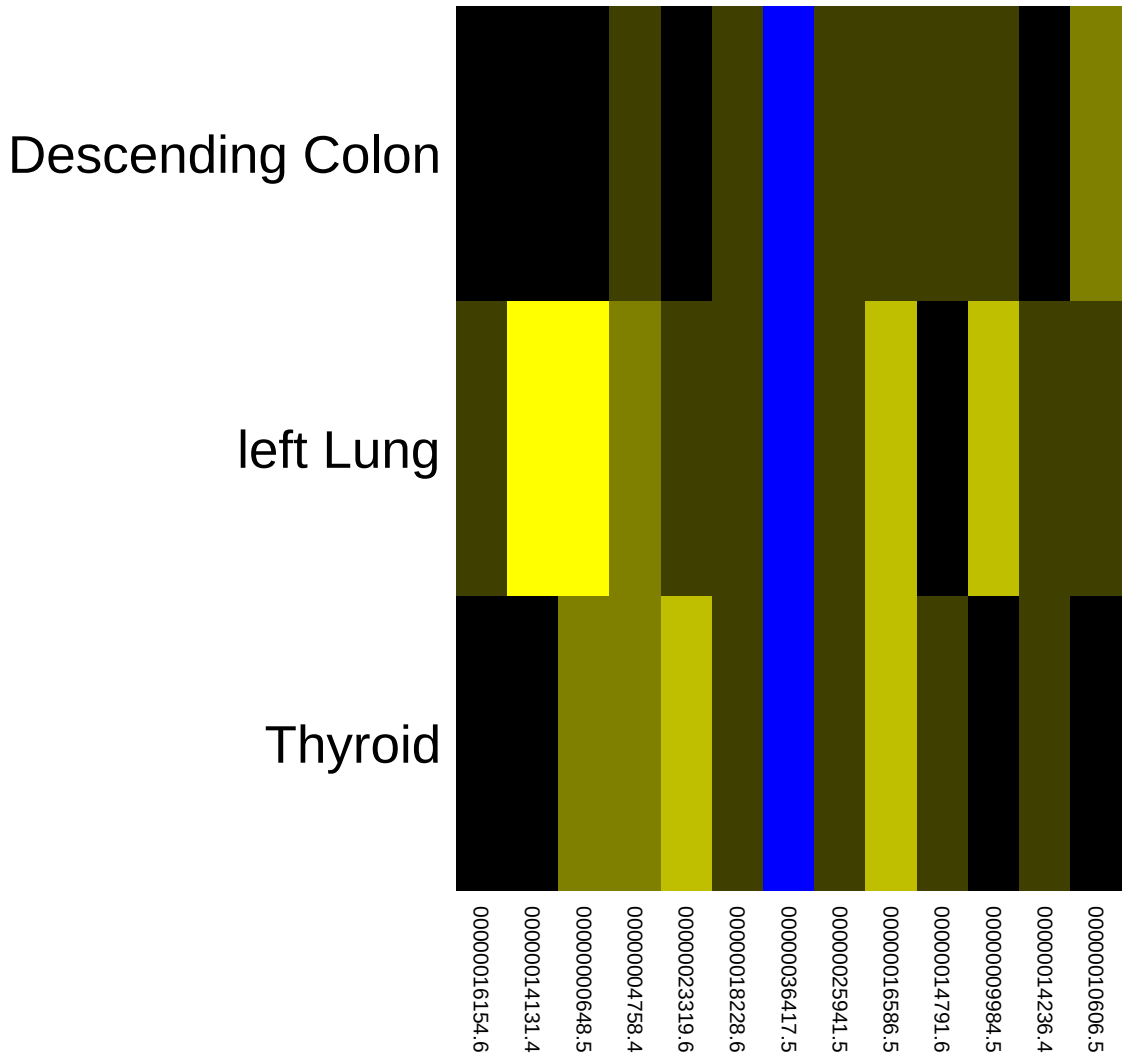


Figure 3.2: A reconstruction block produced by BranchOut. For the gene family in question, a subset of three tissues were found to share the same overall ancestral state reconstruction pattern. The subset of the heatmap for these tissues and genes is shown, colored as in Figure 3.1. Selected figure is taken from the *Bos taurus* analysis (ligand-binding domain of nuclear hormone receptor family, see chapter 6). Transcript identifiers have been truncated to omit the prefix "ENSBTAT".

<family_name>_MCLUST_results.pdf

All the MCLUST category assignments are depicted here against the backdrop of an empirical density function showing the distribution of MSFCs for the given tissue/category. This can be used to perform a quality check on the cluster assignment, and to identify gene families that could benefit from re-analysis using an alternative pair of MCLUST parameter settings. Figure 3.3 shows an example cluster assignment plot.

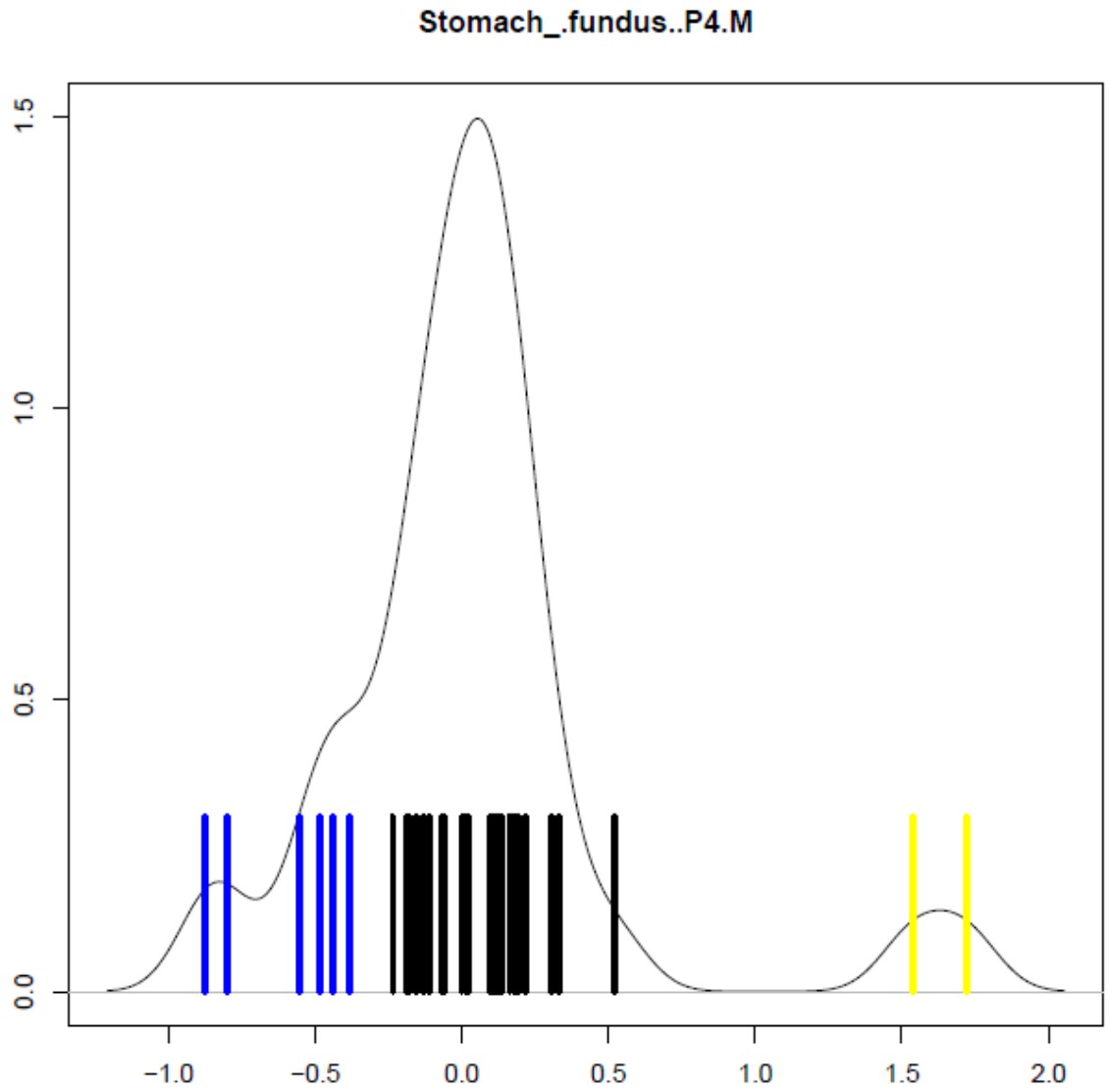


Figure 3.3: An example state assignment diagram. State assignments corresponding to the coloring of leaves on the BranchOut reconstruction trees illustrate how the clustering algorithm split observed expression indices into clusters. One such diagram is produced for each tissue/gene family combination. Selected diagram taken from stomach tissue as part of the *Sus scrofa* analysis (see Chapter 5).

<family_name>_individual_reconstructions.pdf

This document includes one phylogenetic tree for each expression condition included in the input experiment. These trees are annotated with BranchOut's hypothesized ancestral expression states, and optionally include the BranchOut score described above. An example is shown in Figure 3.4.

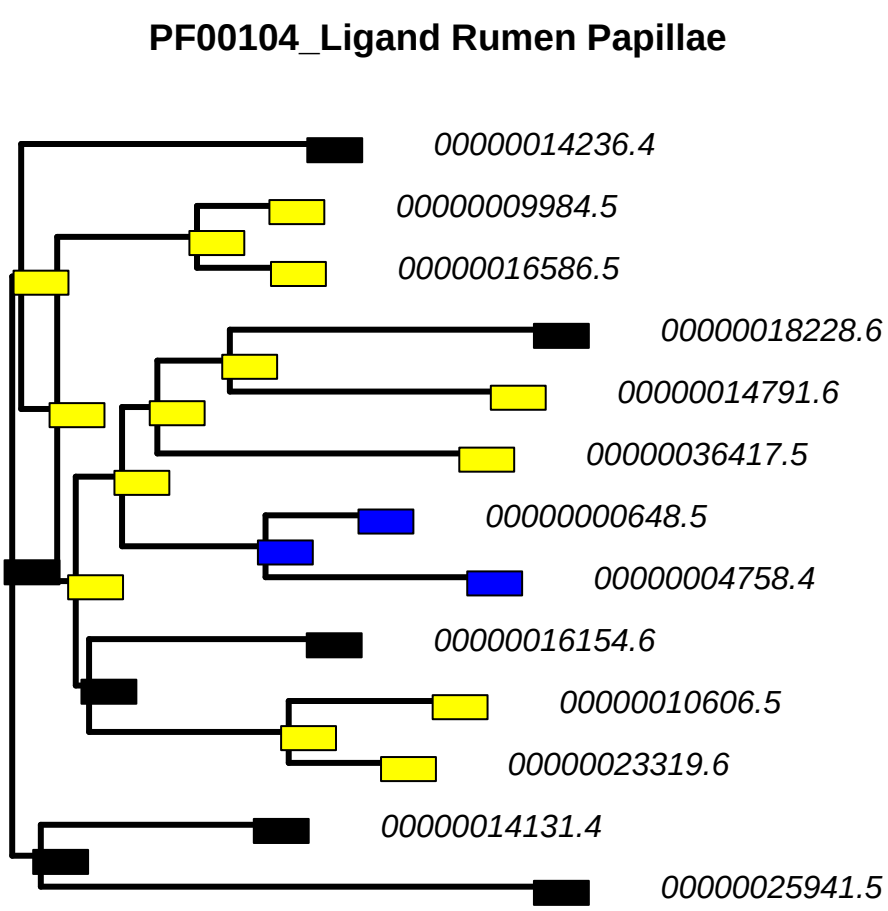


Figure 3.4: An example BranchOut single-condition reconstruction. Colors on the leaves of the tree indicate the expression categories assigned by MCLUST, with internal node colors reflecting the inferred ancestral states. Diagram selected from the output of the *Bos taurus* analysis, ligand binding domain of nuclear hormone receptor family in the rumen (see Chapter 6). Transcript identifiers have been truncated to omit the prefix "ENSBTAT".

3.5 Availability

BranchOut is available publicly on GitHub at the following URL:

https://github.com/owoody/BranchOut

3.6 Conclusions

The BranchOut software produces a folder of images for each included family. Each of these folders consists of a self-contained analysis, highlighting reconstructions that seem to have a strong phylogenetic signal for functional diversification. These folders can be examined as the user specifies, but several summary tables are also available that attempt to direct the user to tissue/family pairings of particular noteworthiness (see Chapters 5, 6, Appendices B, C). To help illustrate the nature of the analysis possible within each of these folders, the subsequent chapter will show an example case of a gene family analysis in *Apis mellifera* which was produced as a proof-of-concept using an early, single-family variant of the BranchOut software.

Chapter 4: Small-scale Application Involving the Honeybee, *Apis mellifera*

4.1 Introduction

Apis mellifera, the western honeybee, has some evolutionary characteristics that make it an interesting test case for the BranchOut software. For example, honeybees have a stinger fed by a venom gland. The venom complement includes an assortment of weaponized proteins. From an evolutionary perspective, all these venom proteins must have been derived from ancestors with presumably benign purposes; the existence of the venom gland is a derived characteristic unique to the aculeata subclade (Tang, Vogler 2017) including wasps, ants, and bees. Therefore, most venom genes must have undergone some form of functional adaptation, being at least examples of neofunctionalization through expression in a novel tissue. This could be the case were something like a digestive protein redirected to venom, for example.

The honeybee is a eusocial organism. In eusocial organisms, there is a pronounced division of labor, including cooperative care and rearing of offspring, with reproductive activities being restricted to a limited number of individuals, (Queller, Strassmann 2003). Though all members of a colony share a common genome, the phenotype this genome produces is pliable and can produce several distinct morphs known as castes. Eusociality is a complex evolutionary trait, and there is an ongoing effort to determine the adaptations underlying a transition to a

eusocial lifestyle at the genetic level (Quinones, Pen 2017). Hymenoptera, the order of insects to which *Apis mellifera* belongs, uses a sex determination system known as haplodiploidy, where sex is determined by ploidy level (rather than through sex chromosomes)(Foster, Wenseleers et al. 2006). In this system, unfertilized eggs typically develop into haploid males, whereas fertilized eggs develop into diploid females. This system affects the relatedness of siblings (based on proportion of shared genetic material) and has been suggested to be a strong supportive element for the evolution of a eusocial lifestyle. Adoption of a eusocial colony structure is believed to have multiple (>9) independent origins in the Hymenoptera (Foster, Wenseleers et al. 2006).

4.2 The honeybee as a eusocial organism

Despite sharing a common genome, the ultimate developmental fate of bee larvae is somewhat malleable, with several distinct phenotypical outcomes. Depending on the conditions under which the larva develops (fertilized vs unfertilized, and depending on nutritional complement -- see 4.3 below), bees can be born into one of several castes (Barchuk, Cristino et al. 2007). These castes determine the role taken by the bee in the hive. Males are haploid, carry out no work in the colony, and serve as a disperser class. Most fertilized eggs develop into diploid worker females, who carry out colony maintenance and gather food resources. If, however, a fertilized (diploid) egg is exposed to a specific complement of nutritional supplements during development, it will develop into a queen, a caste dedicated to reproduction that founds hives and carries out the vast majority of the

reproductive activity for the hive of individuals. This is in stark contrast to worker females, which only rarely attempt to rear eggs of their own and serve entire lives as caretakers (Johnson, B. R. 2010). The female workers are not uniform in their activities, with some having roles predominantly within the hive as maintainers and caretakers, and others serving outside the nest as scouts and gatherers. A worker's current role is also somewhat plastic even in adulthood, and physiological traits (such as fat reserves) can vary considerably depending on what role a worker is currently undertaking (Lattorff, Moritz 2013).

Queens have a "mating flight" where they mate with several males, and they then retain a stock of sperm that are kept viable and used to fertilize eggs through the queen's tenure as the reproductive font of the colony. As a result, most of the female workers in the colony will share a large proportion of their genetic material. Specifically, they inherit one of the two sets of chromosomes possessed by their mother (the queen), but (assuming the sperm are from the same male) they share 100% of the chromosomal complement from their father. One consequence of this is that worker females are more related to their sisters than they would be to their own daughters. The implications this has on kin selection and the development of eusociality have inspired an active field of research (Hughes, Oldroyd et al. 2008).

4.3 The yellow protein family

The yellow protein family is common to most insects and includes genes that typically play a role in early development (Ferguson, Green et al. 2011). The family has relatively ancient origins and is usually present in modest numbers in insect

genomes (~25 blastp hits in *Apis mellifera* and ~10 in *Anopheles gambiae*, based on blastp search, 2016).

The yellow family is an interesting test case for the evolution of novel function for one particularly salient reason: a subset of the yellow proteins have evolved to hold a central role in caste determination and have come to be known as the "major royal jelly proteins" based on how their role is carried out (Drapeau, Albert et al. 2006). In addition to this prominent novel role, yellow proteins have been implicated in taking roles in other tissues in adult *Apis mellifera* (Foret, Kucharski et al. 2009).

The major royal jelly proteins are produced and excreted from the mouthparts of colony workers inside the hive. The jelly serves a dual role, serving as both a foodstuff and as a type of hormone. The hormone effect is particularly pronounced on developing larva. Indeed, the ultimate developmental fate of the larva is determined in large part by the quantity of major royal jelly proteins (MRJPs) that the larva is fed, with heavy investment tilting the path of development towards that of a new queen (Buttstedt, Muresan et al. 2018).

To investigate the evolutionary origins of this subset of royal jelly proteins, BranchOut was applied to the yellow protein family. This application simplifies two aspects of the standard BranchOut pipeline. First, only a single gene family is examined in this pilot. Second, the gene atlas (Foret, Kucharski et al. 2009) is of a relatively small scale, including only 7 to 9 tissues (some samples do not correspond exactly to a single organ). The benefit to having a small-scale pilot is that the results

can be examined manually. Similarly, the single gene family is an established participant in insect development, and is thus likely to show changes reflecting the unique traits of the honeybee species.

4.4 Expression data

Microarray expression data was obtained from the supplementary information from a study on the role DNA methylation plays in regulating invertebrate genes (Foret, Kucharski et al. 2009). The expression atlas here was produced on custom-made cDNA arrays. The microarray design used 12,915 unique oligomer probes that matched sequences from the honeybee. They also included some repeat probes and a number of controls for testing the quality of the hybridization. The samples for hybridization mostly corresponded to tissues: the antennae, the hypopharyngeal gland, the brain, the thorax, the ovary, and the larvae. Most tissues had four unique samples (biological replicates), but some had only three. Each sample was hybridized to a separate microarray. The expression measures used for this study were taken from the publication itself without any modification or alternative analysis.

4.5 Methods

A modified, small-scale version of the BranchOut software was used to analyze the data. As this experiment is intended to illustrate the reasoning behind the approach, the following explanation will focus on the interpretation of the analysis steps. The steps described below were followed once per gene in the yellow gene family.

First, the normalized expression values for a gene were transformed using a logarithm (base 2) transform and then scaled by subtracting the median transformed expression value based on this gene's expression across all tissues to obtain a set of "median-scaled fold changes" (MSFCs). After this process, an MSFC value of 0 corresponds to a gene being expressed at a level exactly matching the median. Values of +1 would correspond to a doubling of this expression level, and -1 would correspond to having the gene be expressed at half the level of the median. A histogram of MSFCs is typically centered close to zero, though this is not strictly necessary.

A set of all gene-specific MSFCs for a given tissue were then used as input to MCLUST (Scrucca, Fop et al. 2016), the clustering software package used by BranchOut. MCLUST assumes the observed MSFCs are drawn from a set of one or more (normally distributed) expression distributions; the software can infer both the number of distributions and their properties (mean, variance) directly. For this study, the number of distributions was not specified in advance and MCLUST decided on the number to model for each tissue on a case-by-case basis. The number of categories varied from 2 (most common) to 7.

For each tissue, the number of distributions was likened to a set of possible expression states, and each state was assigned a color corresponding to the ordering of the mean intensity level of the category (lowest: red, through blue, yellow, violet, orange, green, to highest: white). These colors were assigned for display purposes in summary graphs.

To put these categories in a framework that depicts their evolutionary relatedness and history, a phylogenetic tree was built using the corresponding gene sequences. Following multiple sequence alignment with MUSCLE (Edgar 2004), The MrBayes software (Ronquist, Huelsenbeck 2003) was used to construct a Bayesian phylogenetic tree.

The colors assigned to the categories were used as labels for the leaves of the phylogenetic tree, generating one tree per tissue included in the study. Then, for each of these trees, the ancestral expression states were estimated using the ACE ("Ancestral Character Estimation") algorithm included in the 'ape' (Paradis, Claude et al. 2004a) package for the R statistical software (R Development Core Team 2010). ACE applies a parsimony approach to reconstruction, with transitions between states scored by the number of states separating them. For example, if MCLUST suggested that the MSFCs were drawn from three distinct distributions, these could be conceptually related to the states of "downregulated", "typical regulation", and "upregulated", and a switch from "downregulated" to "upregulated" would pass through one intermediate step and thus be scored with a distance of 2 units. The ancestral state estimation procedure sought to minimize the total number of units of change required to explain the extant expression states on that tree for each tissue. In cases where two rival reconstructions had an equivalent cost, one of the two possible reconstructions was selected at random.

4.6 Results

A heatmap showing the relative distribution of expression signals is shown in Figure 4.1. Note that in this traditional analysis it is clear that there are many genes with elevated expression in the hypopharyngeal gland, but not whether these genes share a common phylogenetic origin.

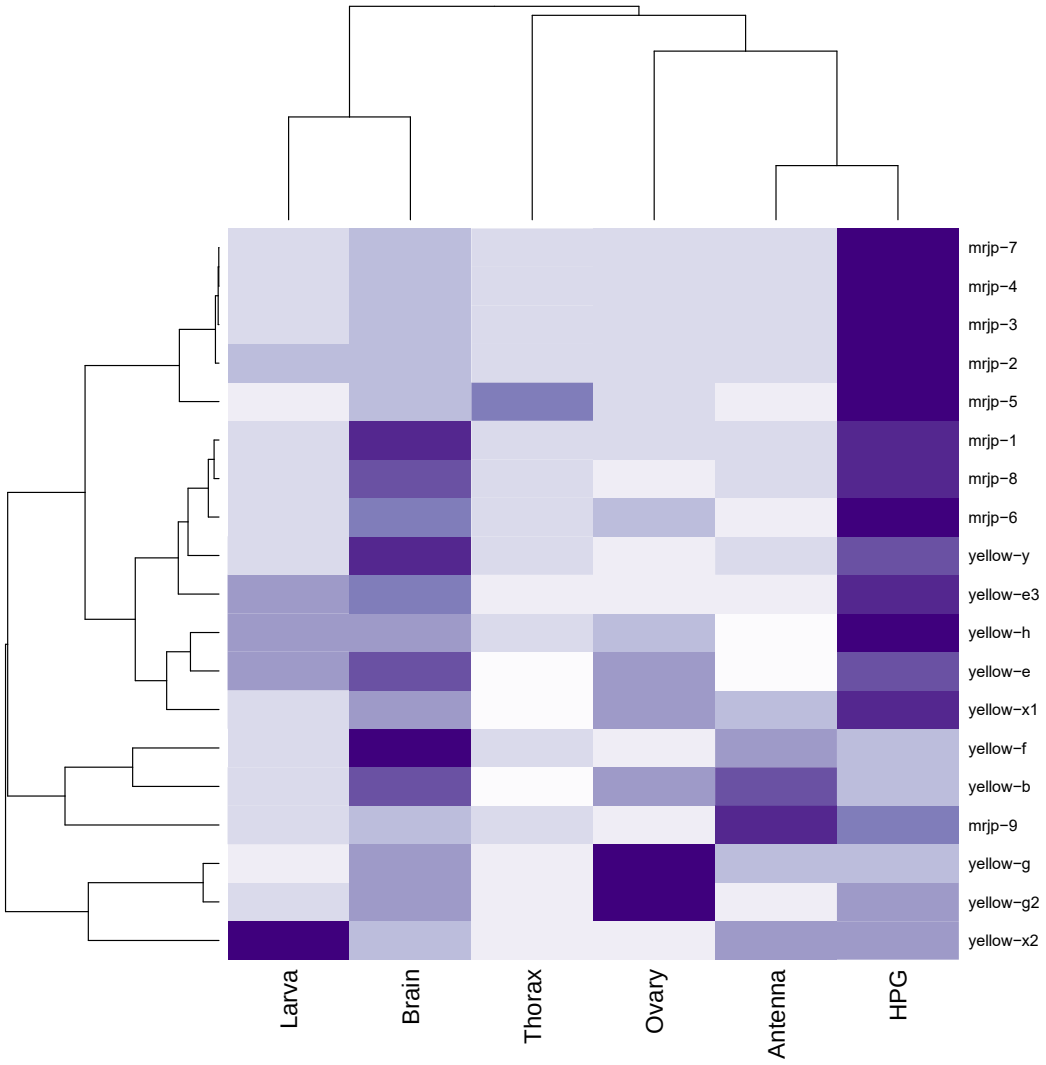


Figure 4.1: Heatmap showing gene expression behavior for the yellow gene family in *Apis mellifera*. Darker colors correspond to greater expression levels. HPG is an abbreviation for the hypopharyngeal gland; the secretory gland near the mouthparts. Genes are indexed using NCBI "Gene" identifiers; all genes from *Apis mellifera*. Alternative identifiers are listed in Appendix A.

Figure 4.2 shows an illustration of the MCLUST output for several tissues. The assignment of expression values to clusters was mostly well done, though the distribution of values in the ovary tissue caused the algorithm to determine that a large number of clusters were appropriate. This behavior could have been avoided

by adjusting the parameters for MCLUST, but a common MCLUST parameter setting was kept for all tissues instead (for consistency and repeatability).

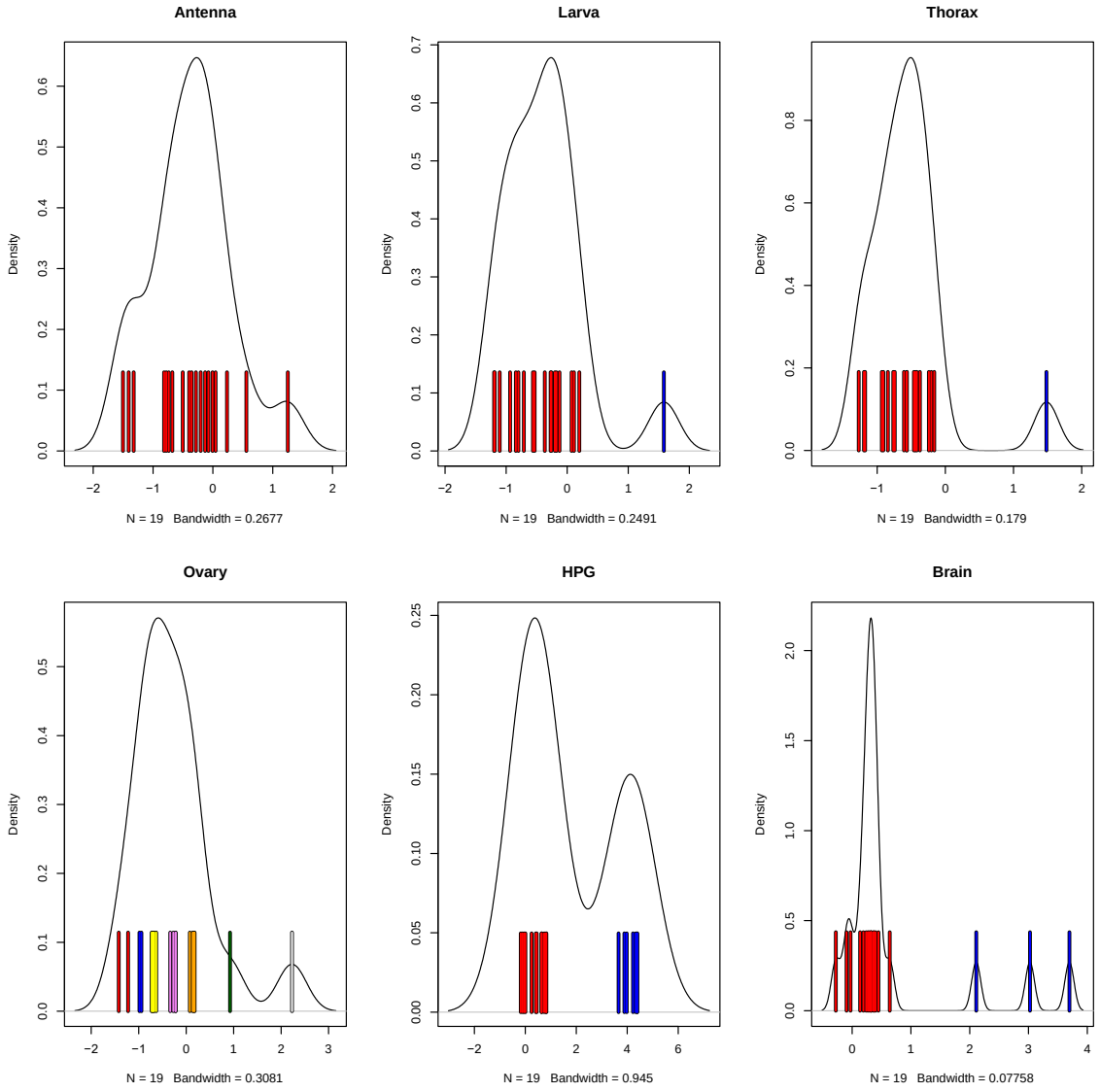


Figure 4.2: MCLUST results for the yellow protein family in *Apis mellifera*. Gene expression levels were standardized to the median (for each gene across tissues), and then MCLUST was used to classify the standardized values into expression categories (distinguished by color). The categories were then used as labels for reconstruction on a phylogenetic tree (Figures 4.4 & 4.5). The line above the colored bars shows the estimated density function, and can be used to infer where MCLUST thought to split the measures into different sample distributions.

The MrBayes phylogenetic tree, using the default parameters, is shown in Figure 4.3. This tree represents a hypothesis of the evolutionary path taken by the

yellow protein family, with splitting branches in the tree representing gene duplication events that led to the present day diversity of yellow proteins in the honeybee genome.

Figure 4.4 shows some example reconstructed trees for the tissues included in this study, contrasting one annotated BranchOut tree for a tissue with clear signal (HPG) to a tissue where the active genes were more distributed (brain).

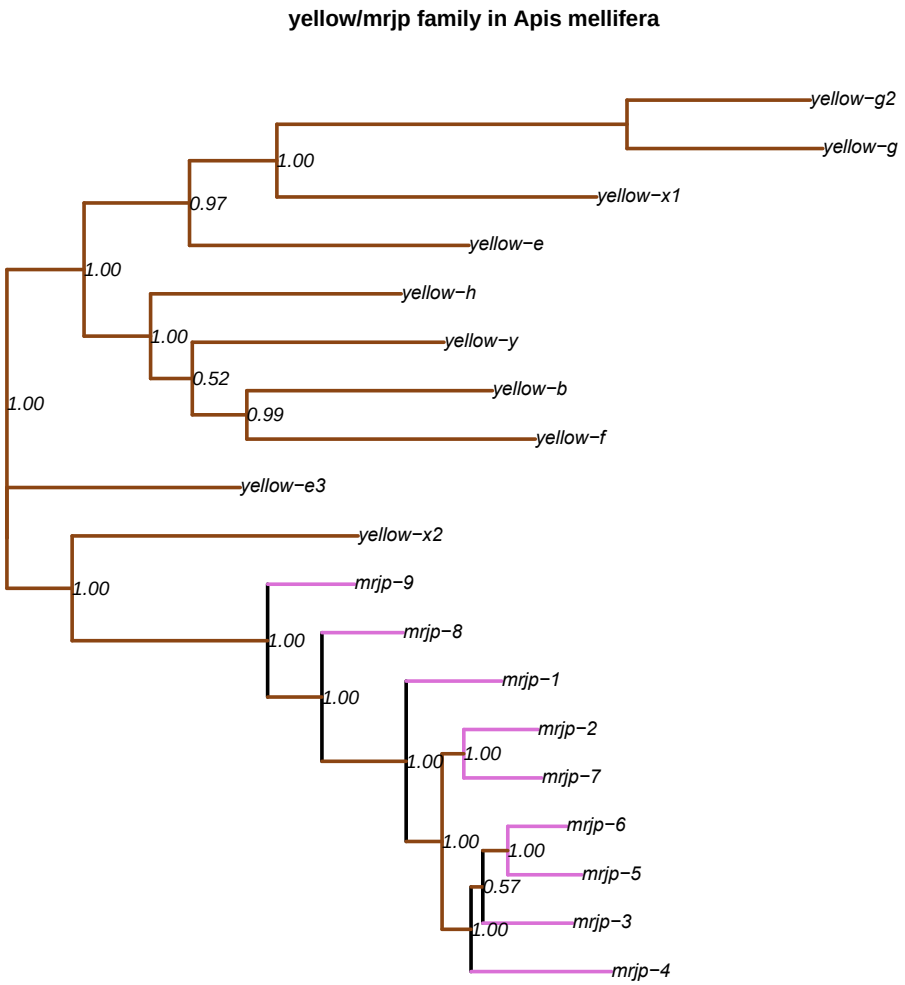


Figure 4.3: MrBayes tree for the yellow protein family in *Apis mellifera*. Branches are colored brown with the exception of the major royal jelly proteins, which are shown with pink branches. Posterior probabilities are shown at internal nodes throughout the tree.

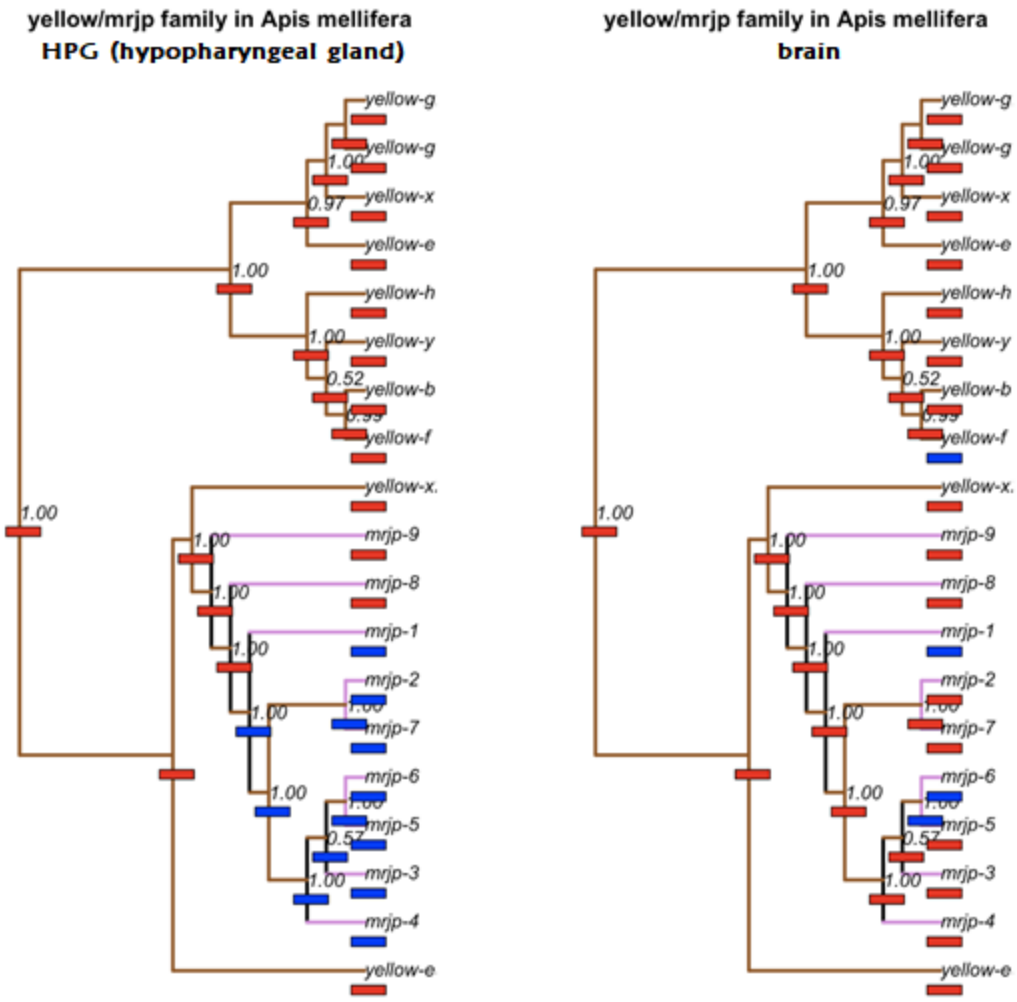


Figure 4.4: Hypothetical expression states of ancestral yellow gene family members from two *Apis mellifera* tissues, hypopharyngeal gland (left) and the brain (right)**.** Expression categories are depicted as colors and correspond to the clusters assigned in Figure 4.2. The hypopharyngeal gland shows a strong association between the upregulation category (blue) and the subtree associated with the major royal jelly proteins, whereas the brain does not.

The most interesting finding is in the hypopharyngeal gland. This gland is associated with the mouthparts of the honeybee. Here, the MCLUST algorithm has split the MSFCs into two categories roughly corresponding to "low" and "high" levels of expression. Many members of the yellow protein family show little-to-no activity in this tissue, but a subset show activation. Intriguingly, this subset of expression levels is clustered together in the phylogenetic tree, suggesting they share common ancestry. In this case, common ancestry is in the sequence-similarity sense – the subgroup of genes all share a common ancestor gene located midway up the tree, and together form a largely monophyletic group.

A further interesting result can be obtained by contrasting this subgroup with common function against the remainder of the gene family tree. In Figure 4, the "low/off" state has been selected for all the other members of the gene family, and the inferred ancestral expression patterns suggest that the ancestor of the entire family should also have been inactive in the hypopharyngeal gland. This is the exact sort of scenario that BranchOut was designed to highlight and present to the user: a gene family tree where a majority follow an ancestral expression pattern, but with a minority that seem to have adopted a novel expression pattern in at least one tissue.

4.7 Discussion

This small test case has shown potential BranchOut approach. In the software's favor, we have been able to identify a tissue where the yellow protein family seems to have undergone an event of functional innovation – that is to say, a

subset of the yellow protein family has been adapted to a new role in the mouthparts, where the remainder of the gene family (and all its ancestors) appear to have had no expression activity there.

This subset of the yellow protein family includes almost all of the members who have the particular additional designation of being "major royal jelly" proteins. These are the members of the yellow protein family that are known to be produced in the mouthparts of worker bees so that they can in turn be fed to developing larvae in the hive. As mentioned earlier, the ultimate developmental trajectory of honeybee young is determined by the quantity of MRJPs allocated to them during larval development. In this case, the BranchOut approach has revealed a key facet of yellow protein evolution that has been characterized previously (Foret, Kucharski et al. 2009) providing proof of concept that the BranchOut approach is effective at identifying function innovations.

Chapter 5: Application of BranchOut to *Sus scrofa* Microarray Expression Atlas

5.1 Introduction

Sus scrofa (the domesticated pig) has become established as a model organism for mammalian biology. In addition to its long history as a livestock animal and food source (Caliebe, Nebel et al. 2017), the pig is sufficiently similar physiologically to humans that several research groups are striving to establish a safe xenotransplantation protocol which would allow pigs to serve as organ donors to address medical demand (Ekser, Rigotti et al. 2009). Pigs are also used as models for a number of human diseases (Cullen, Lu et al. 2018, Bailey, Carlson 2019).

Due to its importance, the *Sus scrofa* genome was first sequenced relatively early (circa 2007) (Fan, Gorbach et al. 2011). This was followed up with a first effort at a full-tissue expression atlas project in 2012 (Freeman, Ivens et al. 2012). With both these primary streams of data publicly available it was possible to perform an analysis using BranchOut on the entire pig genome/transcriptome.

5.2 Methods

Gene expression (microarray) data were downloaded from the ArrayExpress (Kolesnikov, Hastings et al. 2015) web portal (https://www.ebi.ac.uk/arrayexpress/, Accession # E-MTAB-1183). Both the normalized expression values and the "snowball" Affymetrix array definition file (a custom-designed Affymetrix microarray that included a number of probes from the most recent cDNA sequence database, many of which were absent from previous generations of Affymetrix pig arrays) were downloaded from this resource. The

ADF also included each target's transcript, gene, and protein accessions (when available), which facilitated matching microarray probes to various coding sequences.

The expression atlas contains expression values from 105 tissue samples, with most tissues being present in duplicate with one sample from each sex (65 unique tissues). This custom-designed of the *Sus scrofa* Affymetrix array has 47,846 probe sets (including controls) (Freeman, Ivens et al. 2012).

Gene sequence information was downloaded from the Ensembl database (Hunt, McLaren et al. 2018). A complete list of Sus scrofa coding sequences was obtained as a fasta file from (ftp://ftp.ensembl.org/pub/release-75/fasta/sus_scrofa/cds/). Each probeset in the microarray dataset had a corresponding transcript identifier, and these transcript identifiers were matched to corresponding gene sequences in this fasta file. These sequences were used as the input for the phylogenetic component of the BranchOut analysis.

Gene Family affiliation was determined using an identifier mapping table provided on the "Protein Information Resource" (Nikolskaya, Arighi et al. 2007) website (https://proteininformationresource.org/). As discussed in methods, the PIRSF database was queried to identify all *Sus scrofa* records affiliated with known evolutionarily related structure families. In all instances where a probe identifier from the ADF could be associated with a PIRSF record, coding sequence information was obtained and paired with the associated protein family and gene expression records. In order to focus on families where novel functions could be readily distinguishable, very small (and a few very large) families were excluded from

further processing. Any gene families that contained at least 7 or at most 120 members with full sequence, expression, and family annotation were kept for subsequent analysis.

For each family, the cDNA sequences for all fully annotated sequences were exported from R and subjected to multiple sequence alignment using MUSCLE (Edgar 2004) with default parameters. Resulting sequence alignments were converted to PHYLIP format (Retief 2000) and input into PhyML for phylogenetic tree construction (using default parameters including an HKY85 sequence evolution model)(Guindon, Gascuel 2003). This tree was then collected and associated with its gene family record in R. The multiple sequence alignments and tree reconstructions took 8 hours and 24 hours, respectively, on a standard desktop computer.

BranchOut then performed an analysis on each family, clustering member gene expression on a tissue-by-tissue basis allowing up to nine possible clusters. The expression cluster with expression levels closest to the median was assigned a black color, and then clusters with higher/lower expression medians were colored progressively brighter shades of yellow/blue, respectively. The cluster memberships were then displayed on the leaves of the gene family phylogenetic tree, colored accordingly. Cluster reconstruction (as a discrete character) was then performed using the ace function in the ape package in R.

A total of 350 gene families were successfully processed by the BranchOut software. The expression clustering and reconstruction were completed within approximately 3 hours on a standard desktop computer.

BranchOut provides a basic index of phylogenetic signal for prioritizing gene families for inspection. This index is based on the number of expression cluster changes, or "color changes", required to reconstruct the evolutionary history of the gene family. One index is computed for each family/tissue combination. The index itself is the ratio of the estimated number of state changes required in the observed history and the average number required when the same distribution of leaf cluster assignments are randomly reassigned to leaves and subjected to reconstruction. For large subtrees that share a unique expression cluster assignment (corresponding to a sub-family that shares a unique function within the broader gene family tree), the true number of changes should be smaller than the number obtained by shuffling these labels across the entire gene family proper. Correspondingly, large differences should correspond to families with a potentially promising phylogenetic signal.

5.3 Results and discussion

The BranchOut software produces a collection of plots and images for each individual protein family. In addition to this low-level analysis, the software's output can be collated to make some high-level inferences about the tissues and proteins showing signs of functional specialization in the organism as a whole. Three high-level summaries are provided in the following tables.

The first summarization strategy is to identify the tissues with the highest-scoring family reconstructions and to rank-order the findings based on these most prominent results. This approach has the potential advantage of narrowing the search to those findings most likely to correspond to tissue-specific functional

evolution. As a preliminary ranking scheme, each tissue was ranked based on the average of the top three reconstruction scores for all protein families. An abbreviated subset of this list (highlighting the top 20 tissues) is shown in Table 5.1.

Alternatively, instead of focusing on the "interaction" between tissues and protein families, these two axes can be examined in isolation. Table 5.2 shows a subset of a list indicating the frequency at which each tissue category was observed in the set of most-significant BranchOut scores (top 1547 scores). Based on resampling statistics obtained by taking random samples of 1547 tissue labels from the complete output without replacement, a score of 22 or higher is rather uncommon (being the next whole number past the mean plus twice the standard deviation).

Table 5.1: *Sus scrofa* tissues with many high-scoring BranchOut signal scores and a summary of findings.

Tissue	Largest BranchOut Score	Second Largest Score	Third Largest Score
Fallopian tube P3 F	ABC transporter [40](4.625)	MORN repeat [9](4.6)	Cytochrome b5 like Heme Steroid binding domain [10](2.85)
Testis adult M	Disintegrin [24](4.083)	ADAM cysteine rich [22](3.867)	Eukaryotic aspartyl protease [7](3.35)
Abdominal aorta P3 F	X3 5 cyclic nucleotide phosphodiesterase [11](5.85)	ARID BRIGHT DNA binding domain [11](2.95)	Cysteine rich secretory protein family [10](2.025)
Hind Brain medulla P4 M	Disintegrin [24](3.95)	ADAM cysteine rich [22](3.85)	Reprolysin family propeptide [31](2.7)
Spinal cord lower P2 F	Disintegrin [24](3.7)	ADAM cysteine rich [22](3.6)	TIR domain [8](2.65)
Spinal cord upper P2 F	ADAM cysteine rich [22](3.9)	Disintegrin [24](3.75)	G protein alpha subunit [10](2.25)
Mesenteric lymph node P3 F	PCI domain [10](5.3)	Myosin N terminal SH3 like domain [7](2.35)	C1q domain [18](2.25)
Kidney medulla P4 M	Myosin N terminal SH3 like domain [7](3.75)	Frizzled Smoothened family membrane region [7](3)	NHL repeat [7](2.9)
Jejeunum P4 M	Aminotransferase class I and II [13](4.25)	Myb like DNA binding domain [24](2.7)	Class I Histocompatibility antigen domains alpha 1 and 2 [10](2.683)
Stomach fundus P3 F	Cyclin C terminal domain [8](3.85)	Amino acid permease [17](2.9)	Thyroglobulin type 1 repeat [12](2.8)
Trachea P4 M	Zinc carboxypeptidase [16](3.95)	Eukaryotic aspartyl protease [7](3.05)	NAD dependent epimerase dehydratase family [8](2.45)
Jejeunum P3 F	Cullin family [7](3.5)	MIR domain [7](3.05)	PAP2 superfamily [11](2.9)

Liver P3 F	Type I phosphodiesterase nucleotide pyrophosphatase [7](3.85)	C1q domain [18](2.8)	Serpin serine protease inhibitor [25](2.58)
Blood 1	Elongation factor Tu GTP binding domain [15](3.65)	Sodium ion transport associated [7](2.9)	PAP2 superfamily [11](2.6)
Cerebellum P4 M	G protein alpha subunit [10](3.25)	Cyclophilin type peptidyl prolyl cis trans isomerase CLD [15](3.15)	RecF RecN SMC N terminal domain [8](2.625)
Pancreas P4 M	PAP2 superfamily [11](3.2)	Eukaryotic aspartyl protease [7](3.15)	Myosin N terminal SH3 like domain [7](2.65)
Penis P4 M	Annexin [8](3.6)	Elongation factor Tu C terminal domain [9](2.775)	Gelsolin repeat [9](2.6)
Trachea P3 F	Annexin [8](4)	Cysteine rich secretory protein family [10](2.65)	uDENN domain [12](2.3)
Ileum P4 M	Methyltransferase domain [10](4.45)	MAM domain [9](2.6)	ABC transporter [40](1.875)

Note: Numbers in square brackets indicate the number of transcripts assigned to the corresponding family. Numbers in round brackets provide the BranchOut signal score (ratio of expected state transitions to estimated number of state transitions). M and F designations indicate whether a given tissue sample was from a male or female specimen.

Table 5.3 indicates the frequencies at which various protein families appeared in the top 1547 results. Based on resampling statistics obtained by taking random samples of 1547 protein family labels from the complete output without replacement, a score of 8 or higher is rather uncommon (being the next whole number past the mean plus twice the standard deviation).

Complete versions of tables 5.2 and 5.3 can be found in Appendix B.

Table 5.2: Rank-ordered list of *Sus scrofa* tissues that contained a large number of high-scoring BranchOut reconstruction signals.

Tissue Sample	Representation in High-Scoring Reconstructions
Thymus.P3.F	29
Jejeunum.P3.F	26
Cortex_.prefrontal..P3.F	25
Jejeunum.P4.M	25
Blood_1	23
Hind_Brain_.medulla..P4.M	23
Liver.P3.F	23
Cerebellum.P4.M	22
Cortex_.prefrontal..P4.M	22
Placenta.F	22
Tongue_.dermal_layer..P4.M	22
Caecum..apical..P4.M	21
Colon_.distal..P3.F	21
Spinal_cord_.lower..P2.F	21
Testis_.adult..M	21
Colon_.distal..P4.M	20
Colon_.proximal..P3.F	20
Fallopian_tube.P3.F	19
Gall_bladder.P3.F	19
Kidney_.medulla..P4.M	19
Lung_Parenchyma.P3.F	19
Rectum.P4.M	19
Skeletal_muscle_.leg..P3.F	19
Tongue_.dermal_layer..P3.F	19
Blood_2	18
Duodenum.P4.M	18
Hind_Brain_.medulla..P3.F	18
Pancreas.P4.M	18
Rectum.P3.F	18
Spinal_cord_.lower..P3.F	18

Note: The entry in the right column indicates the number of times the indicated tissue contained a high-scoring protein family reconstruction. Roughly 1547 high-scoring reconstructions were present across all tissue/family combinations; a representation of 22 is the next whole number after the expected mean count plus twice the standard deviation of counts (assuming random sampling of 1547 tissue labels). A complete list is shown in Appendix B.

Table 5.3: Rank-ordered list of *Sus scrofa* gene families that contained a large number of high-scoring BranchOut reconstruction signals.

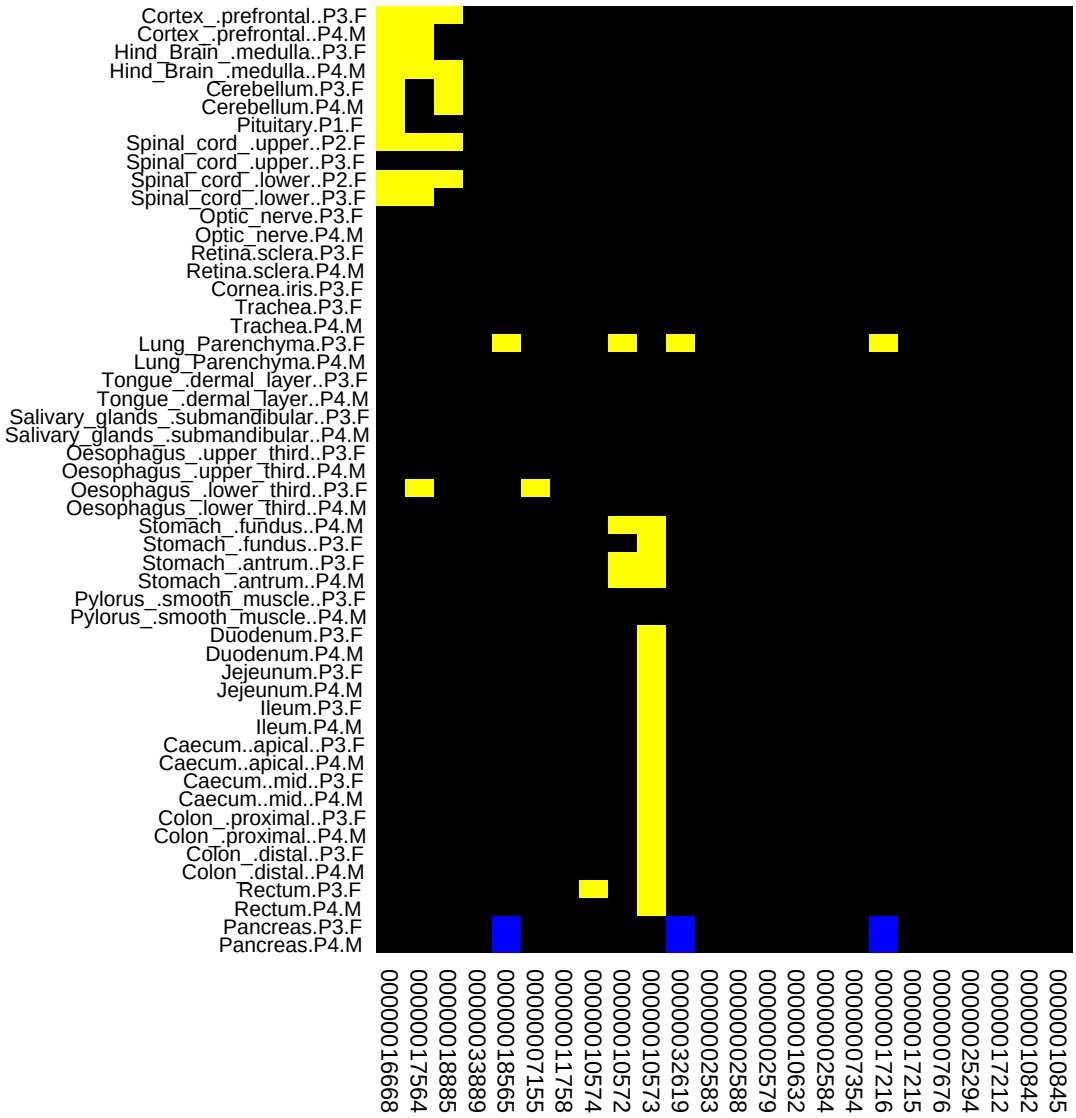
Protein Family Identifier and Description	Representation in High-Scoring Reconstructions
PF02736: Myosin.N.terminal.SH3.like.domain	34
PF00386: C1q.domain	32
PF00026: Eukaryotic.aspartyl.protease	26
PF00244: X14.3.3.protein	25
PF06512: Sodium.ion.transport.associated	21
PF00175: Oxidoreductase.NAD.binding.domain	19
PF00503: G.protein.alpha.subunit	18
PF00030: Beta.Gamma.crystallin	17
PF00160: Cyclophilin.type.peptidyl.prolyl.cis.trans.isomerase.CLD	17
PF00188: Cysteine.rich.secretory.protein.family	17
PF00010: Helix.loop.helix.DNA.binding.domain	15
PF01582: TIR.domain	15
PF02463: RecF.RecN.SMC.N.terminal.domain	15
PF00040: Fibronectin.type.II.domain	14
PF00191: Annexin	14
PF01569: PAP2.superfamily	14
PF01534: Frizzled.Smoothened.family.membrane.region	13
PF00022: Actin	12
PF00055: Laminin.N.terminal..Domain.VI.	12
PF00086: Thyroglobulin.type.1.repeat	12
PF00629: MAM.domain	12
PF00735: Septin	12
PF01266: FAD.dependent.oxidoreductase	12
PF01436: NHL.repeat	12
PF01462: Leucine.rich.repeat.N.terminal.domain	12
PF01576: Myosin.tail	12
PF03062: MBOAT.family	12
PF03114: BAR.domain	12

Note: The entry in the right column indicates the number of tissues in which that family had a high-scoring reconstruction. Roughly 1547 high-scoring reconstructions were present across all tissue/family combinations; a representation of 8 is the next whole number past the mean count plus twice the standard deviation of counts (assuming random sampling of 1547 family labels). A complete list is shown in Appendix B.

Although these high-level summaries suggest "low-hanging fruit" for analysis, examining the visual output of BranchOut can also be informative. Several examples are highlighted in the following sections.

5.3.1 PF00200 – The Disintegrins

The disintegrins are cell receptors. They largely serve as anti-coagulants and play a prominent role in venoms in other species (Giebeler, Zigrino 2016). There were 24 transcripts associated with the disintegrin protein family in *Sus scrofa*. According to BranchOut summary scores, the tissues with the strongest evolutionary signals were the testis, spinal cord, hind brain/medulla, and prefrontal cortex. A summary of the expression classification can be found in Figure 5.1a & b.



(Figure continues on next page)

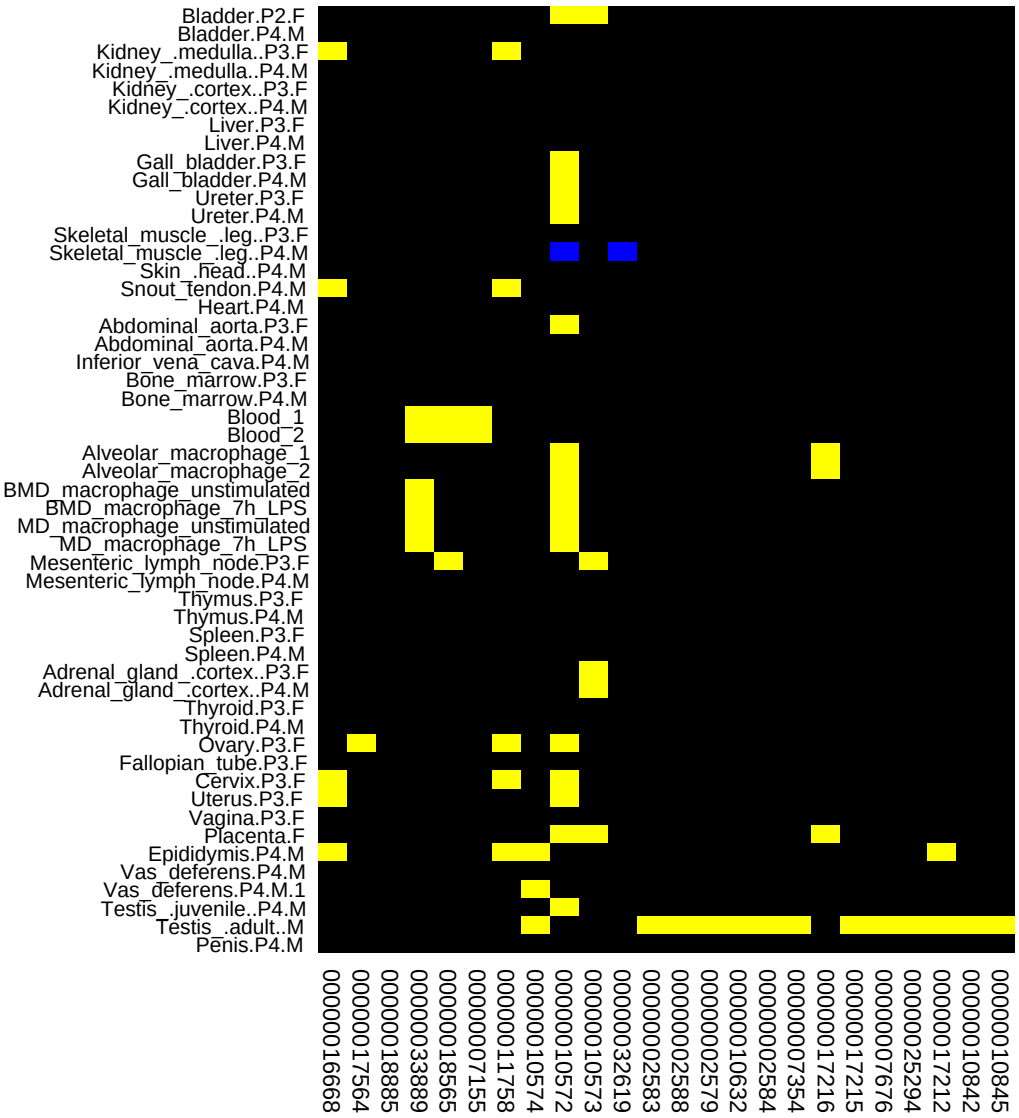


Figure 5.1a &b: Map of Expression Clustering Assignments for the Disintegrin protein family in *Sus scrofa*. Yellow coloring was used to indicate up-regulation of a transcript relative to the protein family-average in a given tissue. Blue indicates down-regulation, and black indicates membership to the cluster closest to the median. Transcript order (right-to-left) matches the ordering in subsequent tree figures (top-to-bottom). Transcript identifiers have been truncated to omit the prefix "ENSSSCT".

In the *Sus scrofa* expression atlas, we can see a monophyletic subgroup of genes that diverge from the expression of the remainder of the family via evidence of expression in the central nervous system. These three genes showed elevated expression in the cortex, cerebellum, medulla and spinal cord (prefrontal cortex shown in Figure 5.2). These three genes showed only limited expression in other tissues.

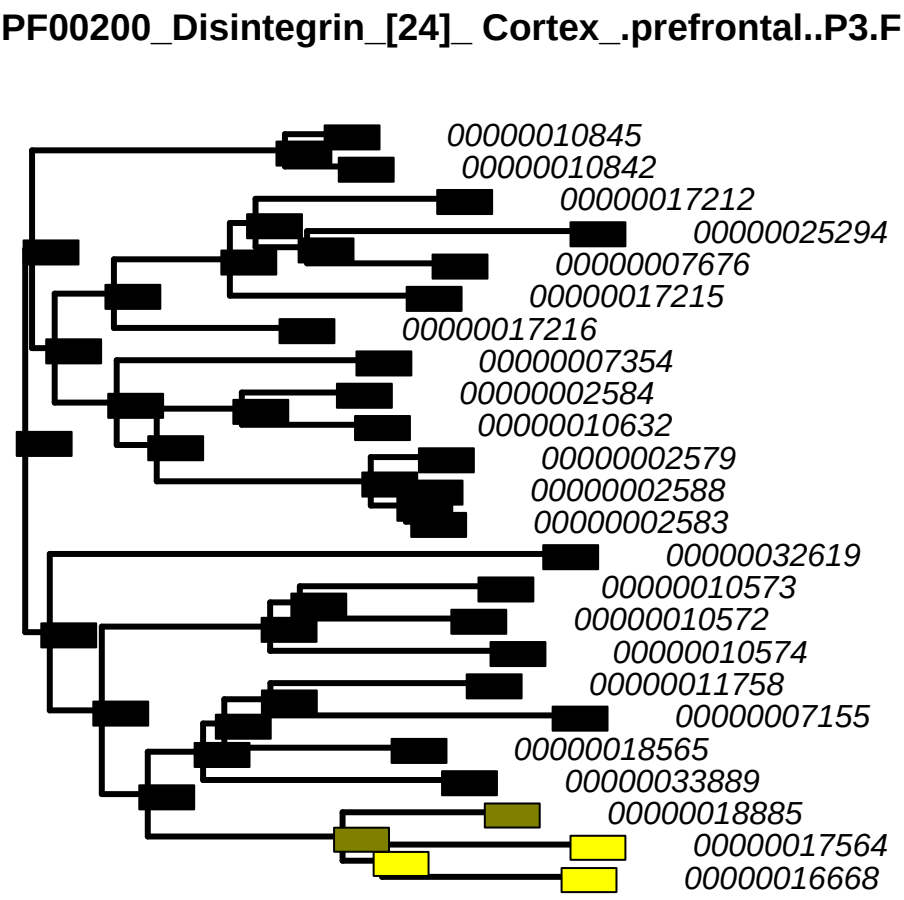


Figure 5.2: BranchOut reconstruction of the Disintegrin protein family in the prefrontal cortex. Dark and bright yellow colors correspond to "moderate-" and "highly-above average expression" classifications for nodes. Transcript identifiers have been truncated to omit the prefix "ENSSSCT".

Intriguingly, the diagram suggests that this subgroup is nestled within a set of genes that are otherwise expressed exclusively in blood cells. The ancestral character estimation proposes that this role in blood may be ancestral to the entire sub-group (see Figure 5.3).

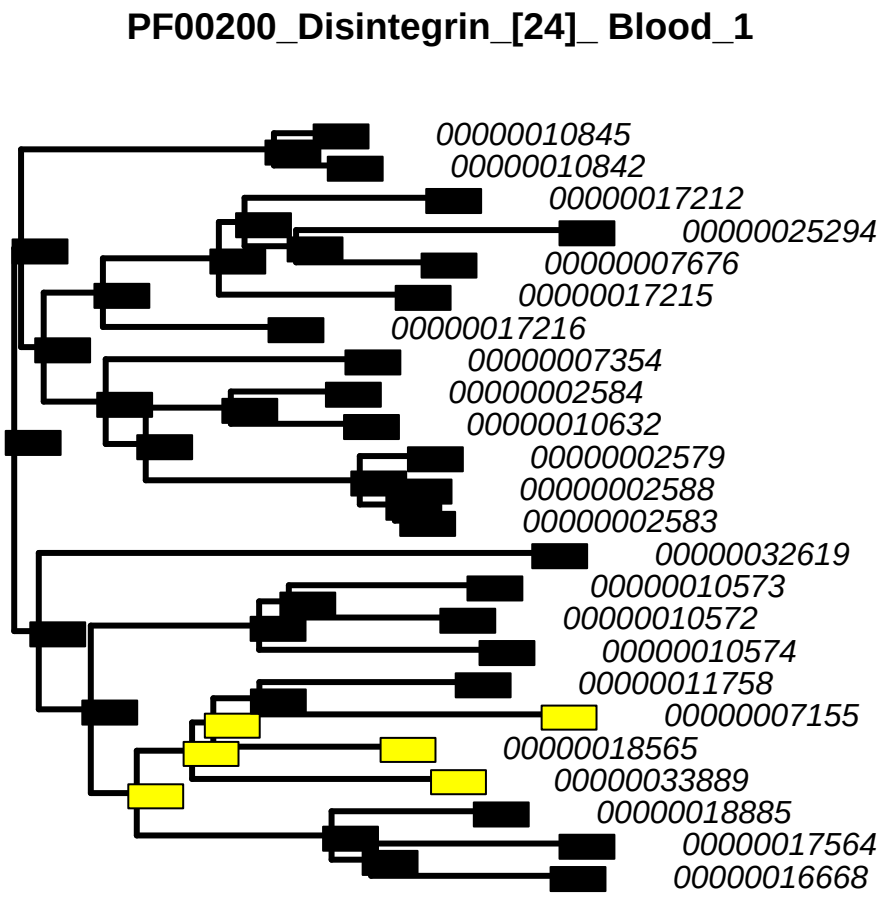


Figure 5.3: BranchOut reconstruction of the Disintegrin protein family in a blood sample. Bright yellow coloring corresponds to a "above average expression" classification for nodes. Transcript identifiers have been truncated to omit the prefix "ENSSSCT".

The majority of the remainder of this family is exclusive to a single tissue: the testes. The entire monophyletic group occupying the top half of the tree showed expression almost exclusively within this gender-specific tissue (see Figure 5.4).

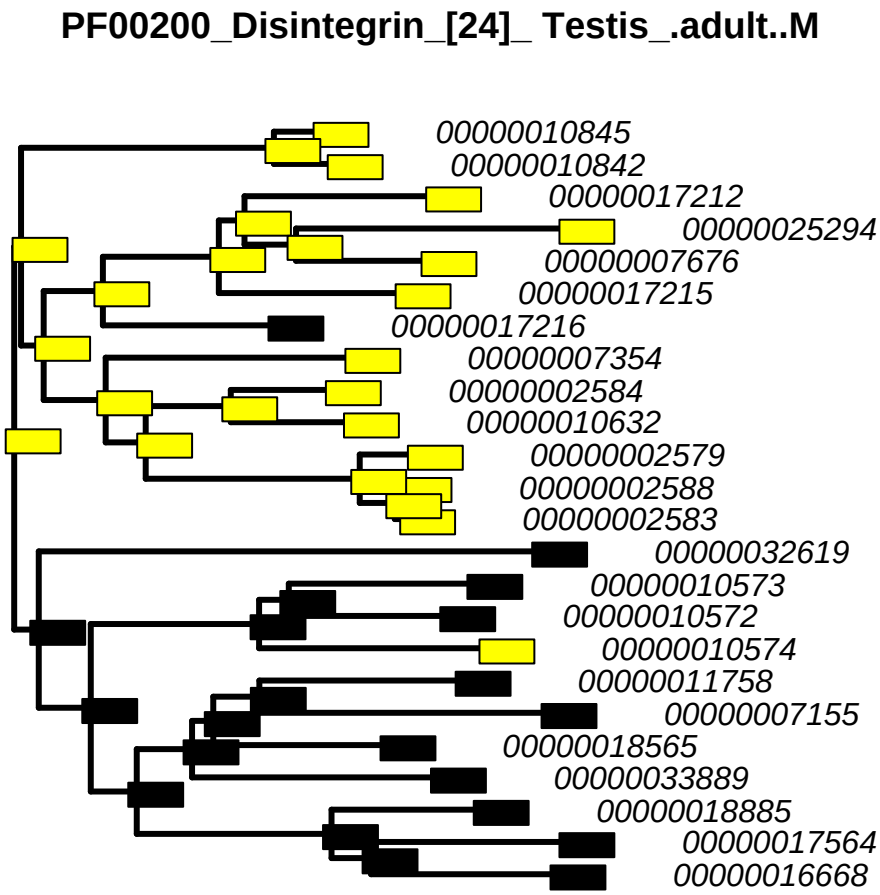


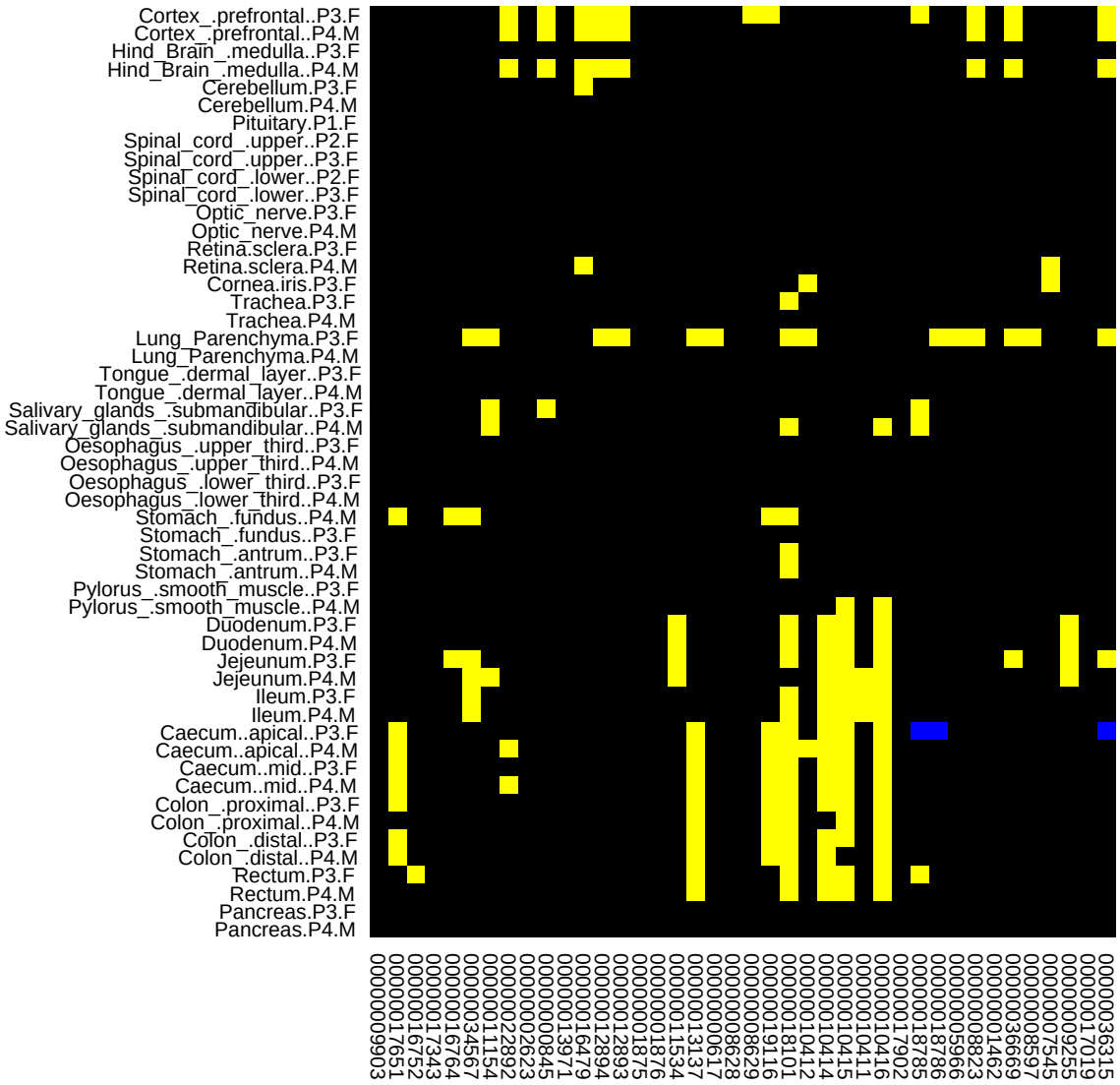
Figure 5.4: BranchOut reconstruction of the Disintegrin protein family in the testis. Bright yellow coloring corresponds to a "above average expression" classification for nodes. The subgroup shown at the top of the figure is composed of mostly yellow nodes, and is one of the most highly scoring BranchOut signals in the *Sus scrofa* data set. Transcript identifiers have been truncated to omit the prefix "ENSSSCT".

The remaining genes showed sporadic expression throughout the digestive system (not shown).

It is tempting to speculate about the potential roles these disintegrin-like members play in the pig. There appear to be some instances of disintegrin-like proteins playing a role in the myelin sheath of nervous tissue (Giebeler, Zigrino 2016); perhaps some of the brain-localized transcripts play a similar role here. The large group showing preferential expression to the testis may play a role in arming sperm with proteins that influence membrane integrity (Kim, Park et al. 2009). Through BranchOut's results, it is easy to note that these roles appear to be largely restricted to specific sub-groupings within the gene family phylogeny.

5.3.2 PF00005: ABC transporters

The ABC-transporters are trans-membrane ATP-driven transport proteins. They are known to play a role in the handling of foreign biological material in the gut (Mercado-Lubo, McCormick 2010). There were 40 members of this family included in this analysis. According to BranchOut summary scores, the tissues with the strongest evolutionary signals were the fallopian tubes, thymus, and ileum. A summary of the expression classification can be found in Figure 5.5a & b.



(Figure continues on next page)

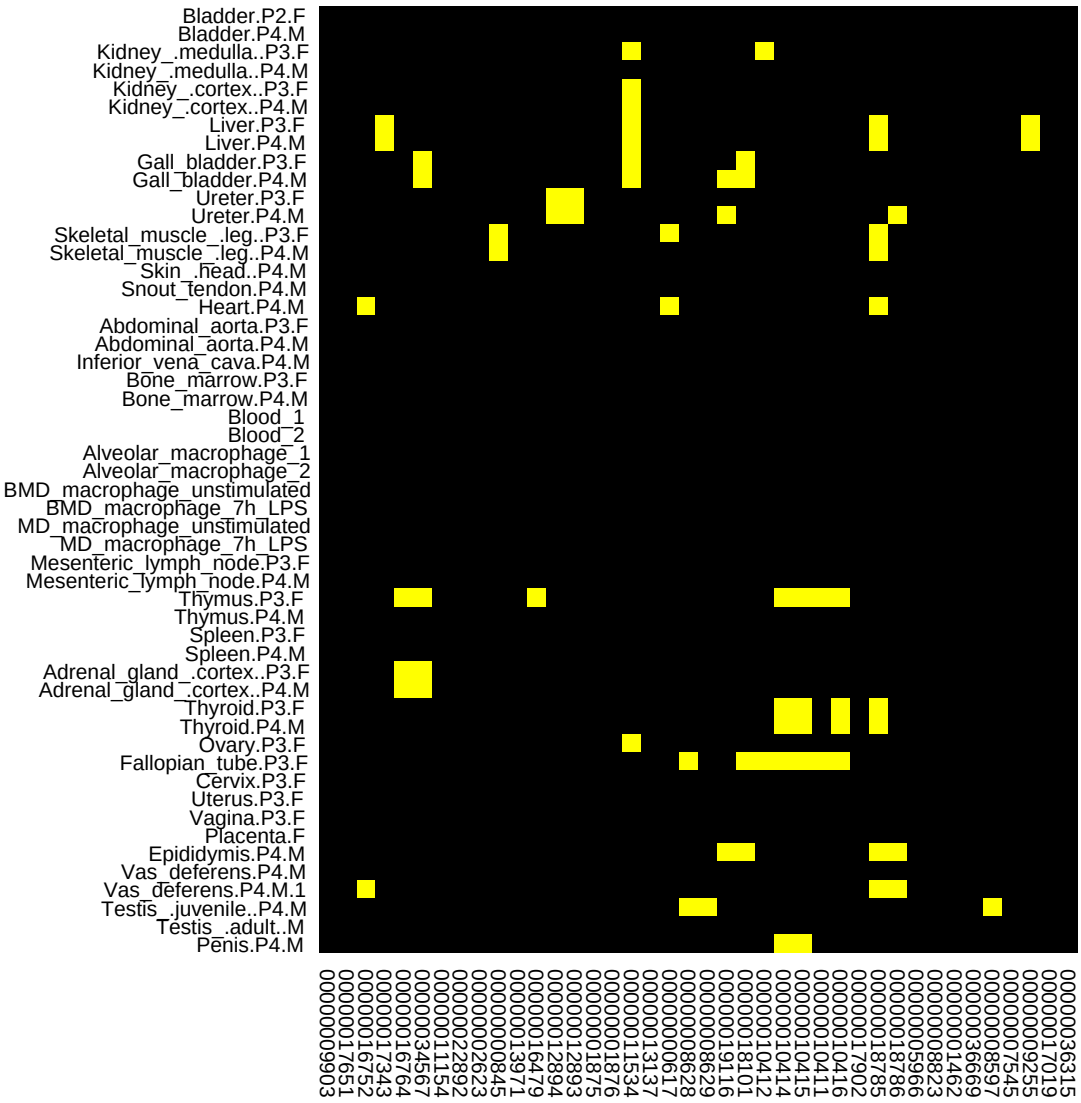


Figure 5.5a & b: Map of Expression Clustering Assignments for the ABC transporter protein family in *Sus scrofa*. Yellow coloring was used to indicate up-regulation of a transcript relative to the protein family-average in a given tissue. Blue indicates down-regulation, and black indicates membership to the cluster closest to the median. Transcript order (right-to-left) matches the ordering in subsequent tree figures (top-to-bottom). Transcript identifiers have been truncated to omit the prefix "ENSSSCT".

The subgroup consisting of transcripts 10416, 10411, 10415, 10414, and 10412 all show elevated expression throughout the duodenum, jejeunum, ilium, caecum, colon, and rectum (ileum shown in Figure 5.6).

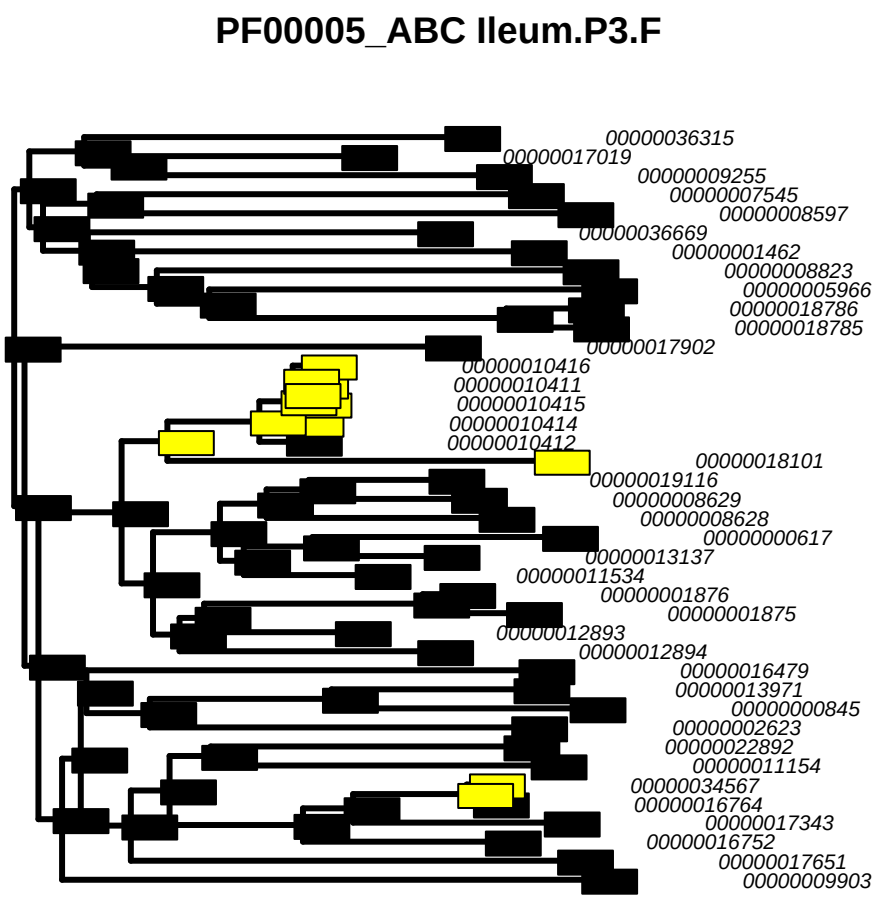


Figure 5.6: BranchOut reconstruction of the ABC transporter protein family in the ileum. Bright yellow coloring corresponds to a "above average expression" classification for nodes. The tight subgroup in the middle of the figure is composed of closely related paralogs that share some exons (and perhaps some regulatory control). Transcript identifiers have been truncated to omit the prefix "ENSSSCT".

The same genes also show elevated expression in the thymus and thyroid glands, in addition to the fallopian tubes in females. At least two of these transcripts represent alternative splicings of the same core genetic sequence; 10412 and 10414 share four exons but vary substantially over the remainder of their coding sequence (Figure 5.7). All five genes are in the same immediate region on the same chromosome.

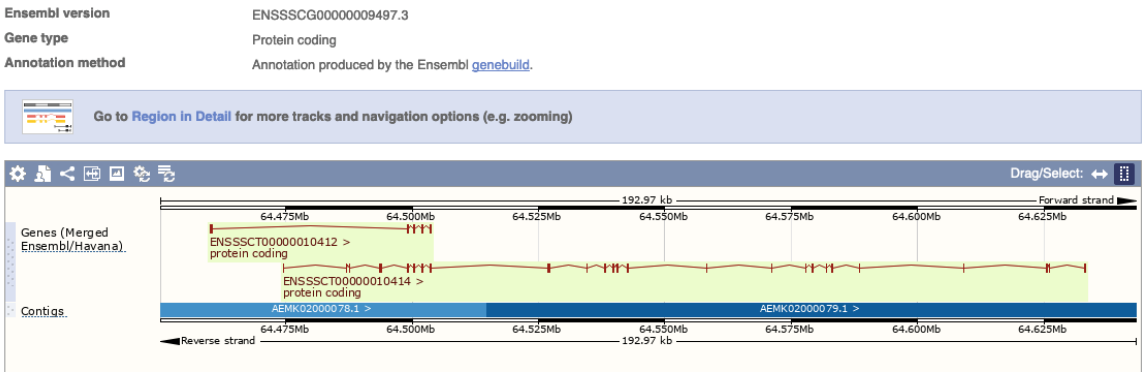


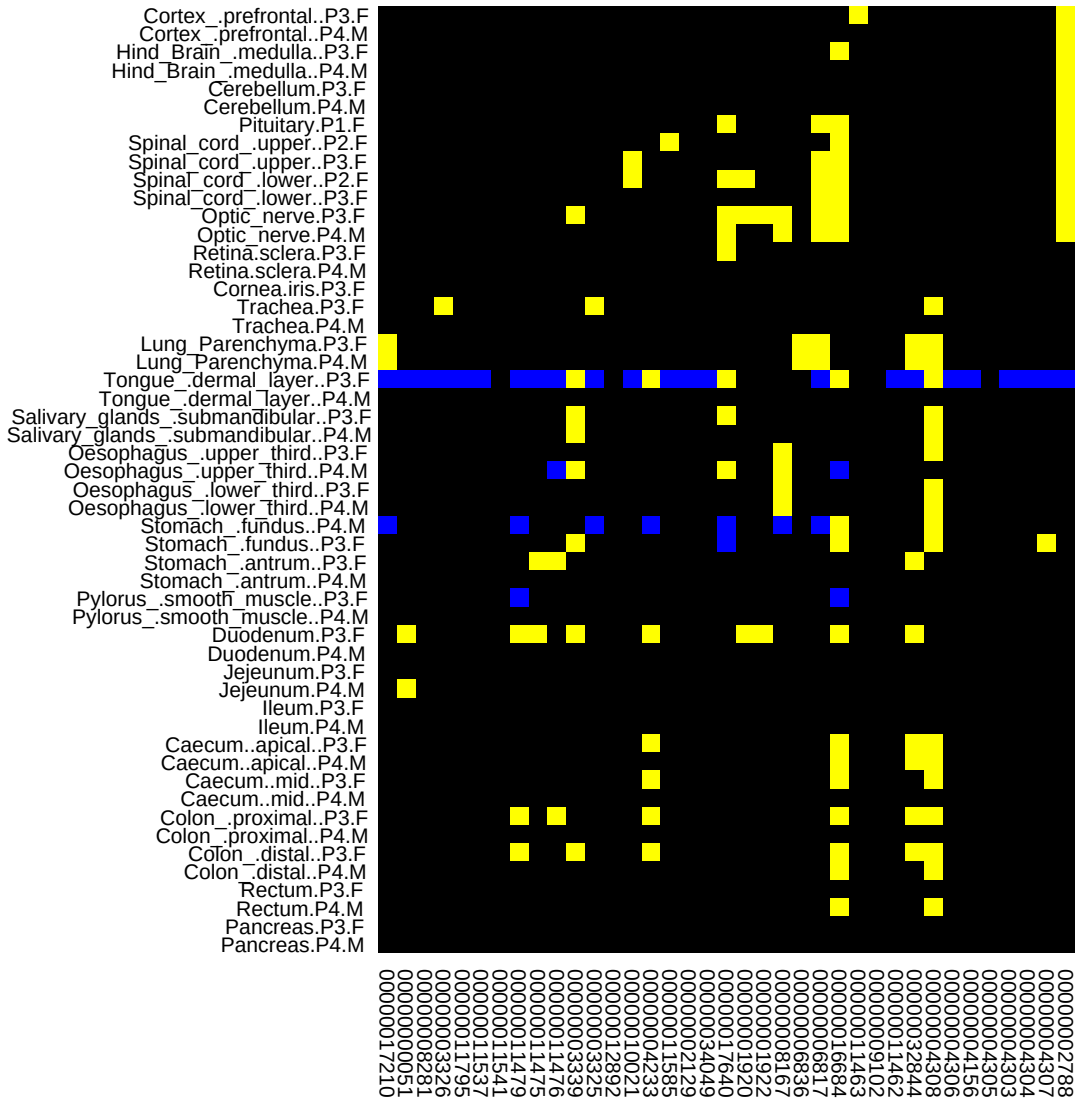
Figure 5.7: Screenshot of Ensembl exon/intron model for selected ABC-transporters. Two overlapping transcripts are shown here. Other transcripts in the immediate subgroup colored in yellow on the tree (see Figure 5.6) are located near this immediate sequence and show similar overlap.

In summary, a large number of closely related transporters are expressed throughout the entirety of the extensively sampled pig digestive system. Given the known role of these proteins in handling xenobiotics, it is tempting to suggest the proliferation of highly specialized forms of these genes could be a response to challenges imposed by domestication. Being restricted to close quarters with many conspecifics could put the domesticated pig at greater risk of contracting parasites. Similarly, the diets of domesticated pigs may include unique challenges when contrasted to what would be encountered foraging in the wild. BranchOut makes it easier to note that these transcripts are not only closely related paralogs, but seem

to be derived from variations on a duplicated region, to the point of even sharing exons. This suggests that splice variation may have been exploited to generate subtle variants that respond to different antagonists.

5.3.3 PF00067: Cytochrome P450

The cytochrome P450 protein family serves a variety of purposes throughout the body. In the liver and digestive tract, they play a role as detoxification enzymes (Ahalawat, Mondal 2018). In this analysis, 37 transcripts were associated with the cytochrome P450 protein family. According to BranchOut summary scores, the tissues with the strongest evolutionary signals were the surface of the tongue, fallopian tubes, and kidney cortex (from summary file in style of Table 5.1, not shown). A summary of the expression classification can be found in Figure 5.8a & b.



(Figure continues on next page)

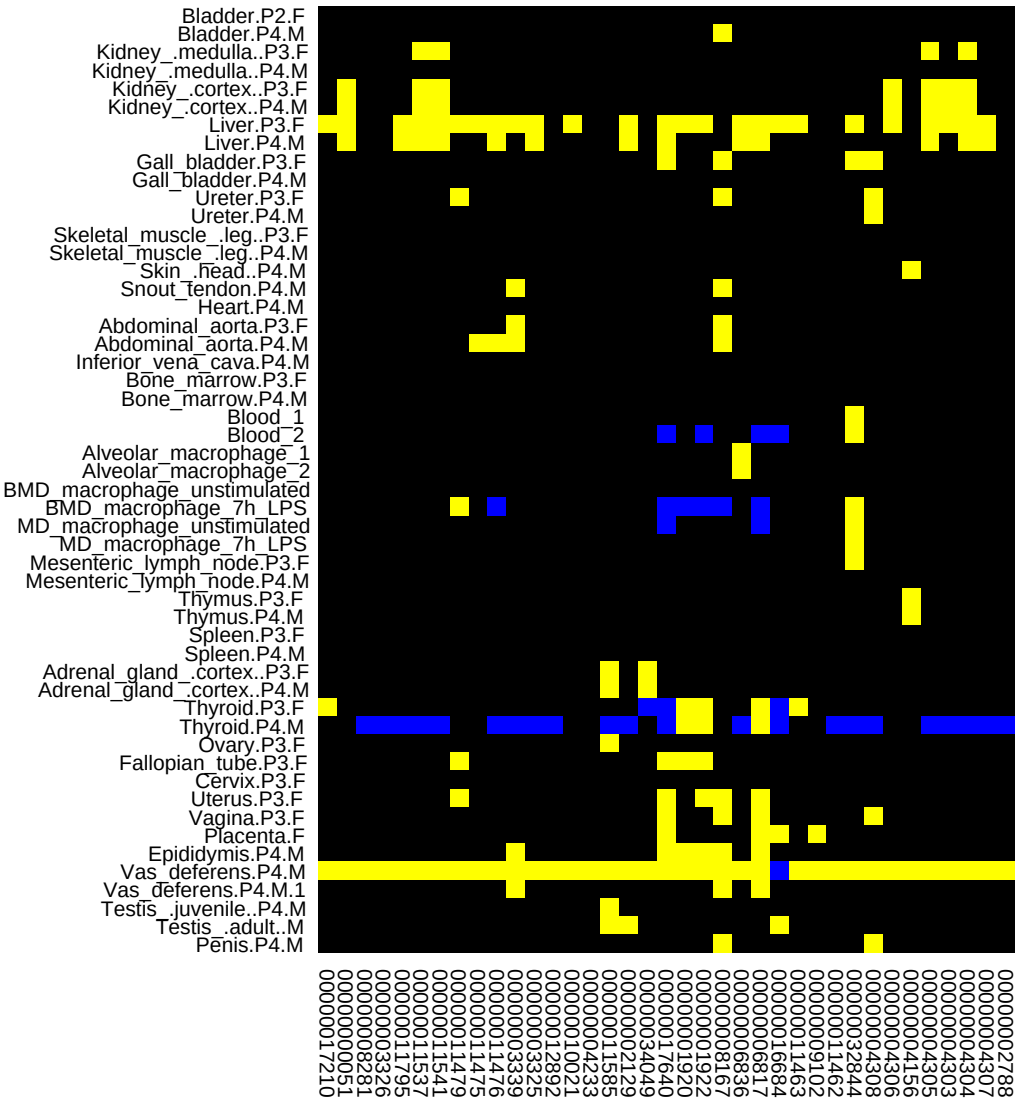


Figure 5.8a & b: Map of Expression Clustering Assignments for the Cytochrome P450 protein family in *Sus scrofa*. Yellow coloring was used to indicate up-regulation of a transcript relative to the protein family-average in a given tissue. Blue indicates down-regulation, and black indicates membership to the cluster closest to the median. Transcript order (right-to-left) matches the ordering in subsequent tree figures (top-to-bottom). Transcript identifiers have been truncated to omit the prefix "ENSSSCT".

In the output of BranchOut, a subgroup composed of highly similar sequences shows coordinated expression in both the liver (not shown) and kidney

(below). Interestingly, this subgroup clusters together with the cytochromes found in the optic nerve, despite having a presumably different biological function (indicative expression shown in Figure 5.9, for the kidney cortex tissue).

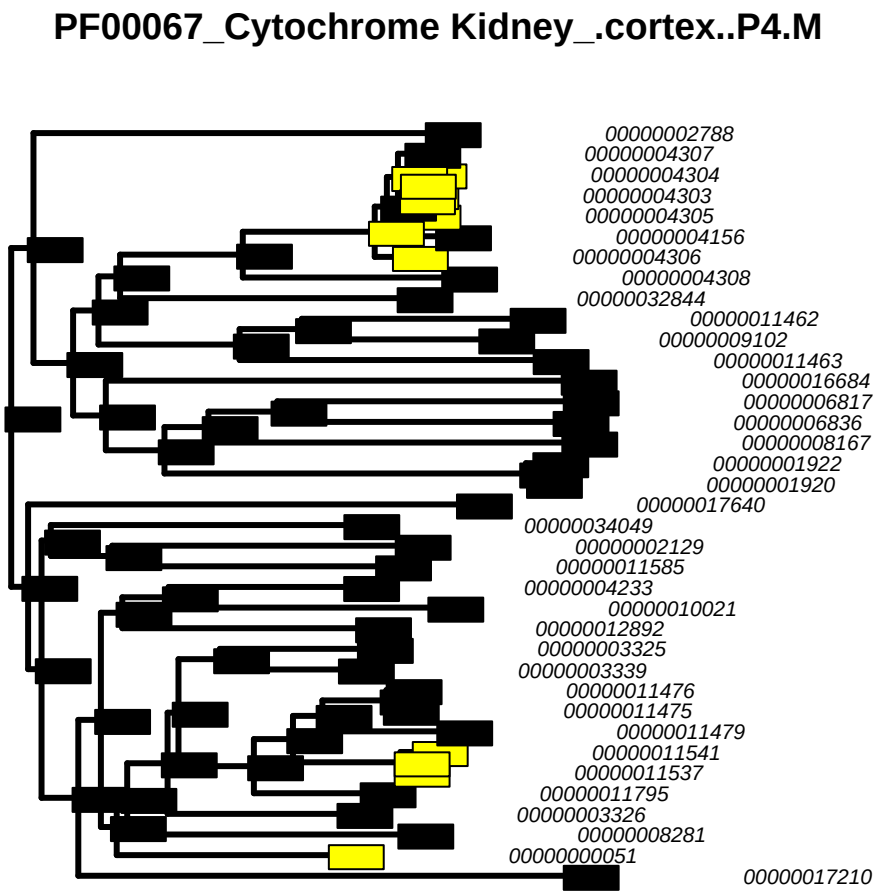


Figure 5.9: BranchOut reconstruction of the cytochrome P450 protein family in the cortex of the kidney. Bright yellow coloring corresponds to a "above average expression" classification for nodes. Transcript identifiers have been truncated to omit the prefix "ENSSSCT".

Several members from the family from the second broad subtree show elevated activity in kidney and liver together with the closely related group situated near the top of the tree. Interestingly, these cytochrome genes are nested within a subgroup that is otherwise active in the digestive tract, but in stomach, duodenum, and colon, all outside the specific vicinity of the kidney/liver.

It would be interesting to examine the extent to which members of subgroups of the cytochrome family are able to "trade roles" and migrate into other active tissues. Having multiple tissues where cytochrome P450 members are active allows diverse selective pressures to act on these genes, perhaps allowing some to pursue evolutionary paths that lend them to adoption in other roles. The BranchOut software helps identify this potential subfunctionalization role by making it clear that the genes active in the kidney and liver are not, on the whole, immediate paralogous siblings, but they may have functionality that is useful in multiple tissue types.

5.4 Summary of *Sus scrofa* BranchOut analysis

The *Sus scrofa* microarray data set provides expression for most tissues in pairs, with one sample taken from each sex. Although there are probably some sex-specific expression patterns, it seems likely that for the majority of shared tissues there should be relatively little sex-specific expression. If one is willing to make this assumption, the coincidence of both male and female tissues in ranking schemes may be suggestive of the reproducibility of BranchOut findings.

It is perhaps reassuring to see both upper and lower spinal cord together in the top scoring findings (Table 1), and to also see both male and female trachea

tissues showing up together in the list. They do not, however, have any high-scoring tissue reconstructions in common, which may suggest that BranchOut analyses must be treated with caution in isolation.

Certain protein families show up high in the ranking tables far more often than expected by chance. To some extent, this appears to be an interaction between features of some phylogenetic trees and the scoring scheme used in this analysis. Specifically, for families with a single outlying gene that often shows divergent expression behavior, reshuffled labels are far more likely to produce reconstructions that push this behavior up the tree than the actual estimated tree itself.

Nonetheless, these results provide a new perspective on the evolution of gene regulatory behavior within *Sus scrofa*. By comparing and contrasting expression profiles within a gene family, it is easier to determine whether the functions of genes within a family have become diversified, and in what ways these diversifications diverge from the expression behavior shown in other members of the family.

Chapter 6: Application of BranchOut to High-Throughput Sequencing: *Bos taurus* Data Set

6.1 Introduction

The domesticated cow (*Bos taurus*) is an economically important livestock animal. The initial bovine genome assembly was published in 2009 (Tellam, Lemay et al. 2009), and a preliminary expression atlas was published in 2010 (Harhay, Smith et al. 2010). The analysis in this chapter uses a more recent expression atlas made available on the Bovine Genome Database.

6.2 Methods

As a second trial run of the BranchOut suite, an RNA-seq expression atlas data set was collected from the Bovine Genome Database (Hagen, Unni et al. 2018). This expression data set does not appear to have a specific associated publication, but can be accessed in its entirety via the NCBI sequence read archive (Kodama, Shumway et al. 2012)(https://www.ncbi.nlm.nih.gov/sra?term=SRP049415). This dataset included 92 tissues, the vast majority of which were sampled from Dominette (the same animal used to produce a bovine genome assembly). The data was produced using an Illumina HiSeq 2000. The processed and normalized sequence read counts were queried from the Bovine Genome Database using its "Query Builder" tool, selecting all tissue expression records without any restrictions.

Normalized read counts (output of cuffnorm (Trapnell, Williams et al. 2010) as provided by Bovine Genome Database) were used as the primary expression data

input for BranchOut. One of the 92 tissues, spinal cord, had an expression profile that was vastly different from the remaining samples (including roughly only 10% the transcript diversity of other tissues) and was excluded from this analysis. To maximize ease-of-interpretation (but perhaps at some cost to biological richness and accuracy), the study was restricted to transcripts that showed evidence of expression in all of the remaining 91 tissues. A total of 25332 transcripts met this criterion for inclusion.

Gene sequence information was downloaded from the NCBI ftp server as suggested by the Bovine Genome Database (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/vertebrate_mammalian/Bos_taurus/latest_assembly_versions/GCF_002263795.1_ARS-UCD1.2/). The complete coding sequences (CDS) file was used to obtain all sequences.

Gene Family affiliation was determined using an identifier mapping table provided on the "Protein Information Resource" website (https://proteininformationresource.org/). The PIRSF database (Nikolskaya, Arighi et al. 2007) was queried to identify all *Bos taurus* records affiliated with known evolutionarily related structure families. This mapping table did not include the transcript labels used to annotate the sequencing data, so a tool made available by UniProt was used to map transcript records to uniprot IDs. A total of 58,077 transcript identifiers were mapped (many-to-many) to a total of 14,311 unique uniprot identifiers.

For each family, the cDNA sequences for all fully annotated sequences were exported from R and subjected to multiple sequence alignment using MUSCLE (Edgar 2004) with default parameters. Resulting sequence alignments were converted to PHYLIP format (Retief 2000, Felsenstein 1988, Felsenstein 1997) and input into PhyML (Guindon, Gascuel 2003) for phylogenetic tree construction (using default parameters including an HKY85 sequence evolution model). This tree was then collected and associated with its gene family record in R. The multiple sequence alignments and tree reconstructions took 2 hours and 9 hours, respectively, on a standard desktop computer.

Many sequences produced in this data set could not be easily matched to protein families. Often it was the case that a single transcript would map to multiple PFAM protein families, and conversely a single (protein) member of a PFAM group often mapped to multiple transcripts. An effort was made to keep as many unambiguous matches as possible, minimizing the number of arbitrary exclusions. In total, 154 PFAM protein families contained a sufficient number of transcripts (>=7) for BranchOut to constructively analyze the associated family.

The normalized read count measure included a number of expression levels that were very low, with some near machine precision. A histogram of normalized read counts was examined and it appeared as though most expression values were 1 or greater. Prior to a log-transformation for subsequent analysis, any expression values below 1 were set to a minimum value of 1 (log-value of zero).

6.3 Results and Discussion

As before (see chapter 5), three high-level summary tables were generated to provide some insight into active tissues and protein families within the *Bos taurus* data set. Table 6.1 shows the result of ranking each tissue based on the average of the top three reconstruction scores for all protein families. The 20 highest scoring tissues are shown. Table 6.2 shows a subset of a list indicating the frequency at which each tissue category was observed in the set of most-significant BranchOut scores (top 737 scores). Based on resampling statistics obtained by taking random samples of 737 tissue labels from the complete output without replacement, a score of 13 or higher is rather uncommon (being the next whole number past the mean plus twice the standard deviation).

Table 6.1: *Bos taurus* tissues with many high-scoring BranchOut signal scores and a summary of findings.

Tissue	Largest BranchOut Score	Second Largest Score	Third Largest Score
Mesenteric lymph node	Immunoglobulin C1 set domain [7](4.05)	Thioredoxin [7](3.5)	TPR repeat [28](2.45)
Super bull testis	Zinc finger C3HC4 type RING finger [7](3.65)	Acyltransferase [7](3.15)	Cyclin N terminal domain [9](3)
Gall bladder	Tubulin C terminal domain [13](3.45)	Tubulin FtsZ family GTPase domain [14](3.2)	Leucine Rich repeats 2 copies [13](2.8)
Internal tongue muscle	Zinc knuckle [8](3.85)	Mitochondrial carrier protein [21](3.1)	NHL repeat [7](2.35)
Jejunum	Immunoglobulin C1 set domain [7](3.55)	Adaptor complexes medium subunit family [8](2.9)	von Willebrand factor type A domain [8](2.8)

Adrenal	Immunoglobulin C1 set domain [7](3.9)	Thioredoxin [7](2.65)	EF hand [8](2.45)
Lymph nodes	Immunoglobulin C1 set domain [7](3.9)	EF hand [8](2.8)	von Willebrand factor type A domain [8](2.15)
Atrium	EGF like domain [12](4.5)	Histone like transcription factor CBF NF Y and archaeal histone [7](2.175)	LSM domain [12](2.05)
Left lung	Immunoglobulin C1 set domain [7](3.75)	Guanylate kinase [10](3.15)	KH domain [12](1.65)
Infraspinatus (top blade or flat iron from shoulder)	von Willebrand factor type A domain [8](4.05)	RNA recognition motif a.k.a. RRM RBD or RNP domain [8](2.075)	BTB And C terminal Kelch [17](2.067)
Vas deferens	Tubulin C terminal domain [13](3.4)	Tubulin FtsZ family GTPase domain [14](2.65)	X7 transmembrane receptor rhodopsin family [12](2)
Infundibulum (ipsilateral to CL)	UBX domain [7](3.95)	Zinc finger C4 type two domains [13](2.075)	Ligand binding domain of nuclear hormone receptor [13](2)
Ascending colon	Cofilin tropomyosin type actin binding protein [8](3.45)	EF hand domain pair [9](2.5)	Variant SH3 domain [19](2.05)
Rumen	Zinc finger C4 type two domains [13](2.75)	Ligand binding domain of nuclear hormone receptor [13](2.6)	Ets domain [9](2.575)
Caecum	Immunoglobulin C1 set domain [7](3.9)	Tubulin C terminal domain [13](2.05)	Tubulin FtsZ family GTPase domain [14](1.8)
Salivary gland	Immunoglobulin C1 set domain [7](3.95)	EF hand domain pair [26](2)	Snf7 [8](1.775)
Ampula (contralateral to CL)	Calcineurin like phosphoesterase [10](3.15)	Leucine Rich repeats 2 copies [13](2.65)	Aminotransferase class I and II [11](1.9)
Spleen	Immunoglobulin C1 set domain [7](3.8)	X7 transmembrane receptor rhodopsin family [12](2.125)	NUDIX domain [14](1.717)

Midbrain	Regulator of G protein signaling domain [7](3.2)	Leucine Rich repeats 2 copies [13](2.4)	Aminotransferase class I and II [11](1.867)
Anterior Eye	Sulfotransferase domain [7](2.75)	von Willebrand factor type A domain [8](2.55)	Regulator of G protein signaling domain [7](2.1)

Note: Numbers in square brackets indicate the number of transcripts assigned to the corresponding family. Numbers in round brackets provide the BranchOut signal score (ratio of expected state transitions to estimated number of state transitions).

Table 6.2: Rank-ordered list of *Bos taurus* tissues that contained a large number of high-scoring BranchOut reconstruction signals.

Tissue Sample	**Representation in High-Scoring Reconstructions**
ampula (contralateral to CL)	15
Pineal Gland	14
Ascending colon	13
Anterior Eye	12
Atrium	12
Cerebellum	12
infundibulum (ipsilateral to CL)	12
Internal Tongue Muscle	12
mesenteric lymph node	12
Posterior Pituitary	12
vas deferens	12
Cerebral cortex	11
follicle 2	11
Jejunum	11
Omasum	11
Super bull Testis	11
Thalamus	11
Ventricle	11
Gall Bladder	10
isthmus (contralateral to CL)	10
isthmus (ipsilateral to CL)	10
Liver	10
Midbrain	10
Spleen	10
Temporal Cortex	10

Bone Marrow	9
Caecum	9
Corpus Luteum (if present, estimate d of cycle)	9
Infraspinatus (top blade or flat iron from shoulder)	9
Infundibulum (contralateral to CL)	9
Longissimus dorsi (ribeye/loin)	9
Rumen	9
Sub-cutaneous Fat	9
uterine endometrium - caruncular (contralateral to CL)	9

Note: The entry in the right column indicates the number of times the indicated tissue contained a high-scoring protein family reconstruction. Roughly 737 high-scoring reconstructions were present across all tissue/family combinations; a representation of 13 is the next whole number past the mean count plus twice the standard deviation of counts (assuming random sampling of 737 tissue labels). A complete list is shown in Appendix C.

Table 6.3 indicates the frequencies at which various protein families appeared in the top 737 results. Based on resampling statistics obtained by taking random samples of 737 protein family labels from the complete output without replacement, a score of 9 or higher is rather uncommon (being the next whole number past the mean plus twice the standard deviation).

Complete versions of tables 6.2 and 6.3 can be found in Appendix C.

The visual output of BranchOut was again searched manually for results to highlight. Three such examples are highlighted in the following sections.

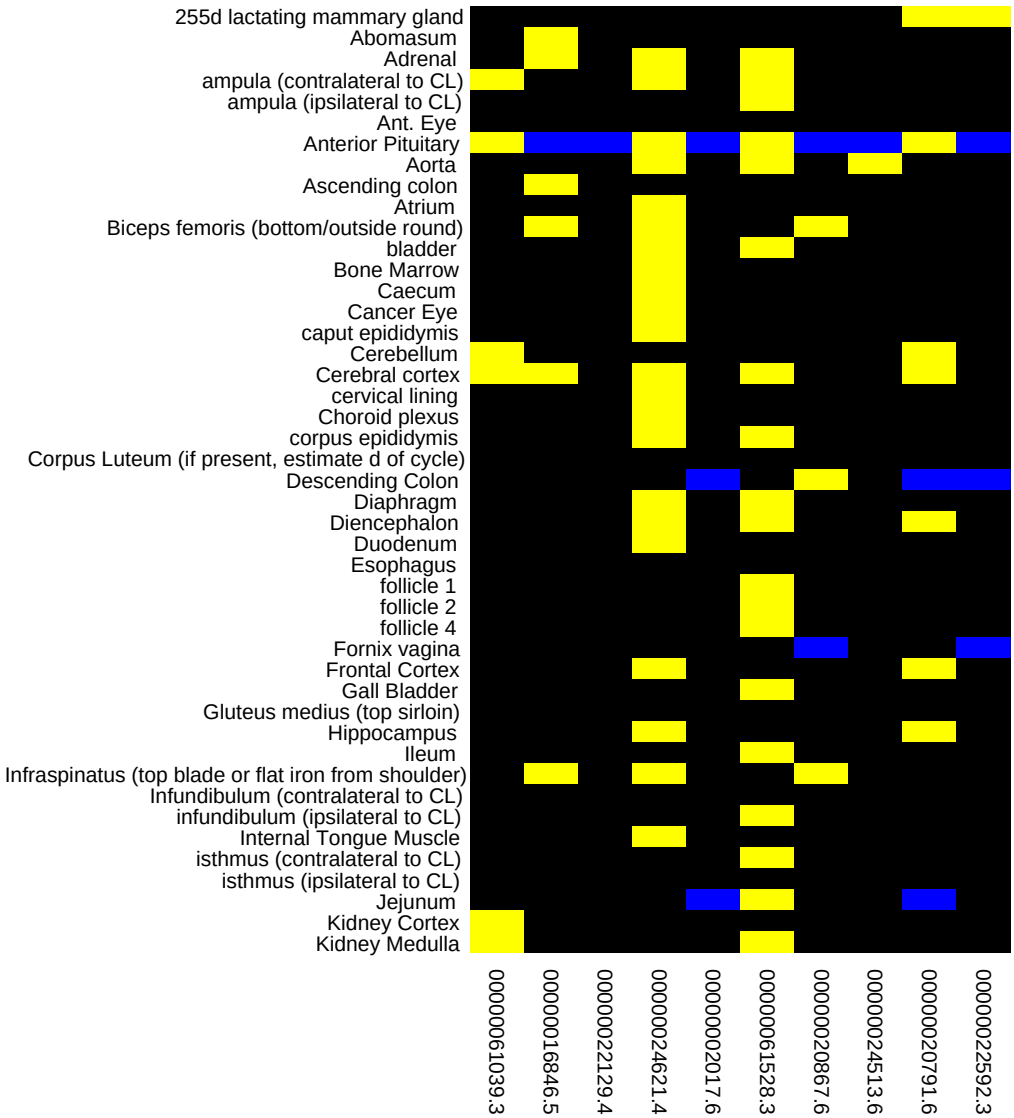
Table 6.3: Rank-ordered list of *Bos taurus* gene families that contained a large number of high-scoring BranchOut reconstruction signals.

Protein Family Identifier and Description	Representation in High-Scoring Reconstructions
PF00134: Cyclin..N.terminal.domain	28
PF00063: Myosin.head..motor.domain.	20
PF00102: Protein.tyrosine.phosphatase	19
PF00013: KH.domain	16
PF00615: Regulator.of.G.protein.signaling.domain	16
PF01553: Acyltransferase	16
PF07654: Immunoglobulin.C1.set.domain	16
PF00149: Calcineurin.like.phosphoesterase	15
PF00153: Mitochondrial.carrier.protein	14
PF00125: Core.histone.H2A.H2B.H3.H4	13
PF00105: Zinc.finger..C4.type..two.domains.	12
PF00226: DnaJ.domain	12
PF00501: AMP.binding.enzyme	12
PF07719: Tetratricopeptide.repeat	12
PF12937: F.box.like	12
PF00091: Tubulin.FtsZ.family..GTPase.domain	11
PF02214: BTB.POZ.domain	11
PF07645: Calcium.binding.EGF.domain	11
PF13516: Leucine.Rich.repeat	11
PF07525: SOCS.box	10
PF00097: Zinc.finger..C3HC4.type..RING.finger.	9
PF00104: Ligand.binding.domain.of.nuclear.hormone.receptor	9
PF03953: Tubulin.C.terminal.domain	9
PF05773: RWD.domain	9
PF08205: CD80.like.C2.set.immunoglobulin.domain	9
PF13637: Ankyrin.repeats..many.copies.	9

Note: The entry in the right column indicates the number of tissues in which that family had a high-scoring reconstruction. Roughly 737 high-scoring reconstructions were present across all tissue/family combinations; a representation of 9 is the next whole number past the mean count plus twice the standard deviation of counts (assuming random sampling of 737 family labels). A complete list is shown in Appendix C.

6.3.1 PF00397: WW-domain family

The WW-domain family, so-named based on the inclusion of a distinctive pair of consecutive tryptophan amino acids (WW), is known to mediate various protein ligand interactions (Dodson, Fishbain-Yoskovitz et al. 2015). Ten members of this family were included in this BranchOut analysis. According to BranchOut summary scores, the tissues with the strongest evolutionary signals were the omasum, duodenum and tongue. A summary of the expression classification can be found in Figure 6.1a & b.



(Figure continues on next page)

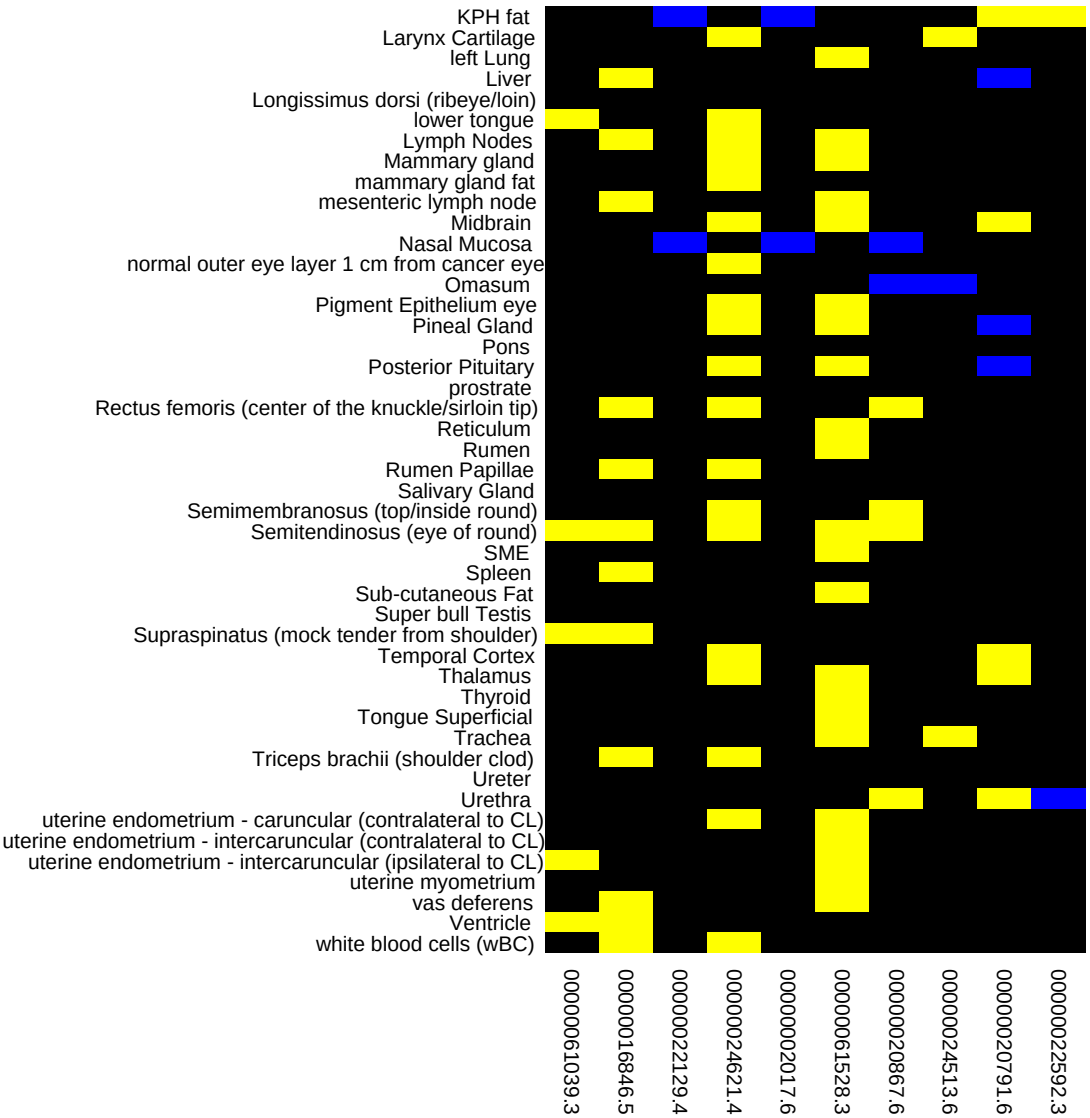


Figure 6.1a & b: Map of Expression Clustering Assignments for the WW-domain protein family in *Bos taurus*. Yellow coloring was used to indicate up-regulation of a transcript relative to the protein family-average in a given tissue. Blue indicates down-regulation, and black indicates membership to the cluster closest to the median. Transcript order (right-to-left) matches the ordering in subsequent tree figures (top-to-bottom). Transcript identifiers have been truncated to omit the prefix "ENSBTAT".

In cows, one WW-domain family member, PIN1, has been shown to play a role in ovarian follicles (Shimizu, Tetsuka et al. 2007). In contrast to expectation,

only one gene, MAGI3 (ENSBTAT00000061528, 6.2), showed consistent expression elevation exclusive to ovarian follicles (and did so across all follicle tissue samples). Moreover, this gene was not restricted to these tissues; it showed elevated expression across a number of various tissue types and samples.

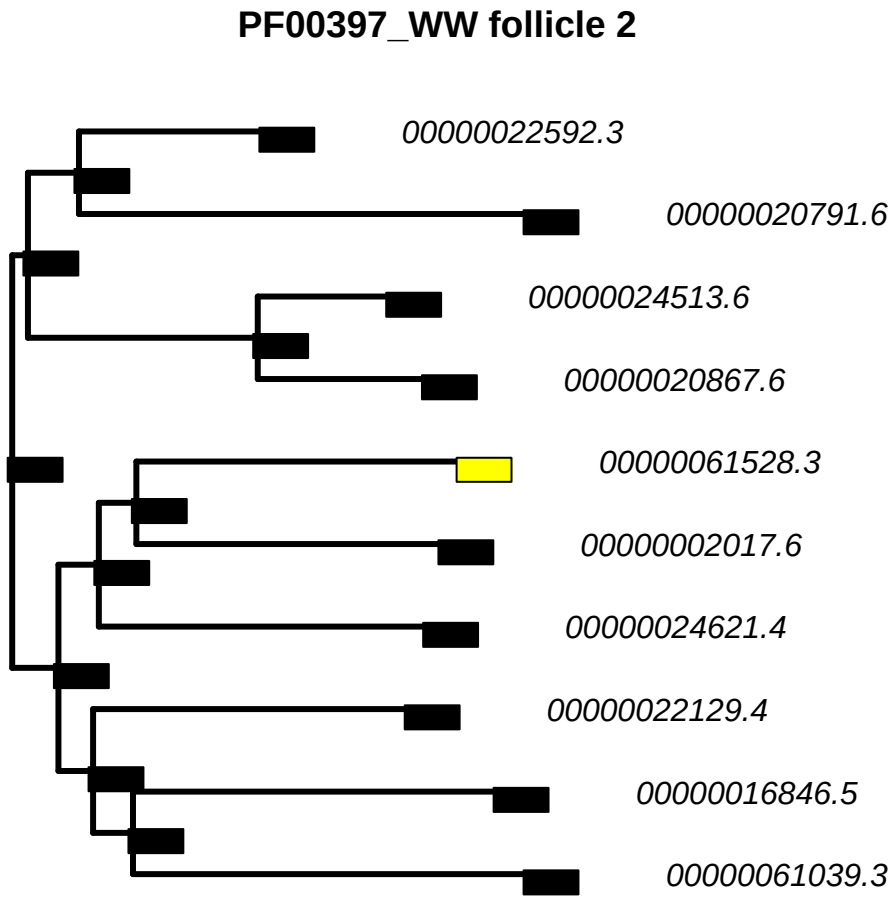


Figure 6.2: BranchOut reconstruction of the WW protein family in an ovarian follicle sample. Bright yellow coloring corresponds to an "above average expression" classification for a node. Transcript identifiers have been truncated to omit the prefix "ENSBTAT".

However, an alternative role for the WW gene family in reproduction is possible based on the behavior of transcript ENSBTAT00000022592 (Figure 6.3, at top). This is a transcript derived from the PIN1 gene (Shimizu, Tetsuka et al. 2007). This gene shows elevated expression in two expression theaters: lactating mammary glands and fat deposits. Its immediate neighbor in the tree, transcript 20791, is also expressed in these two tissues, but is also expressed throughout the brain.

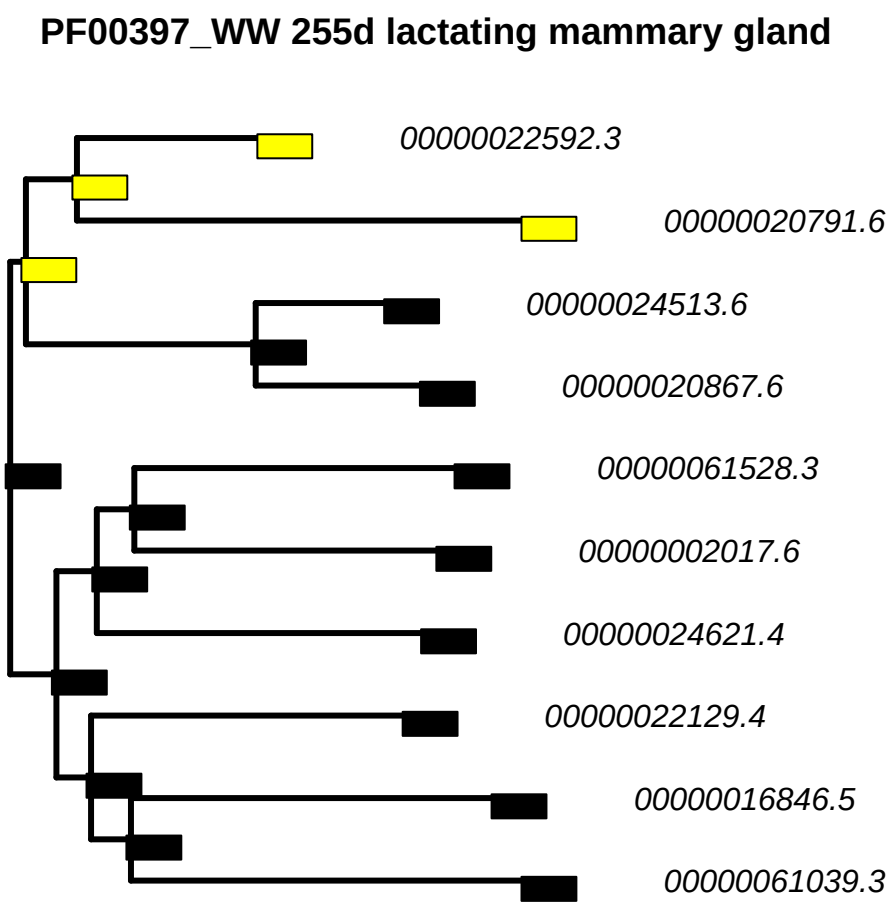


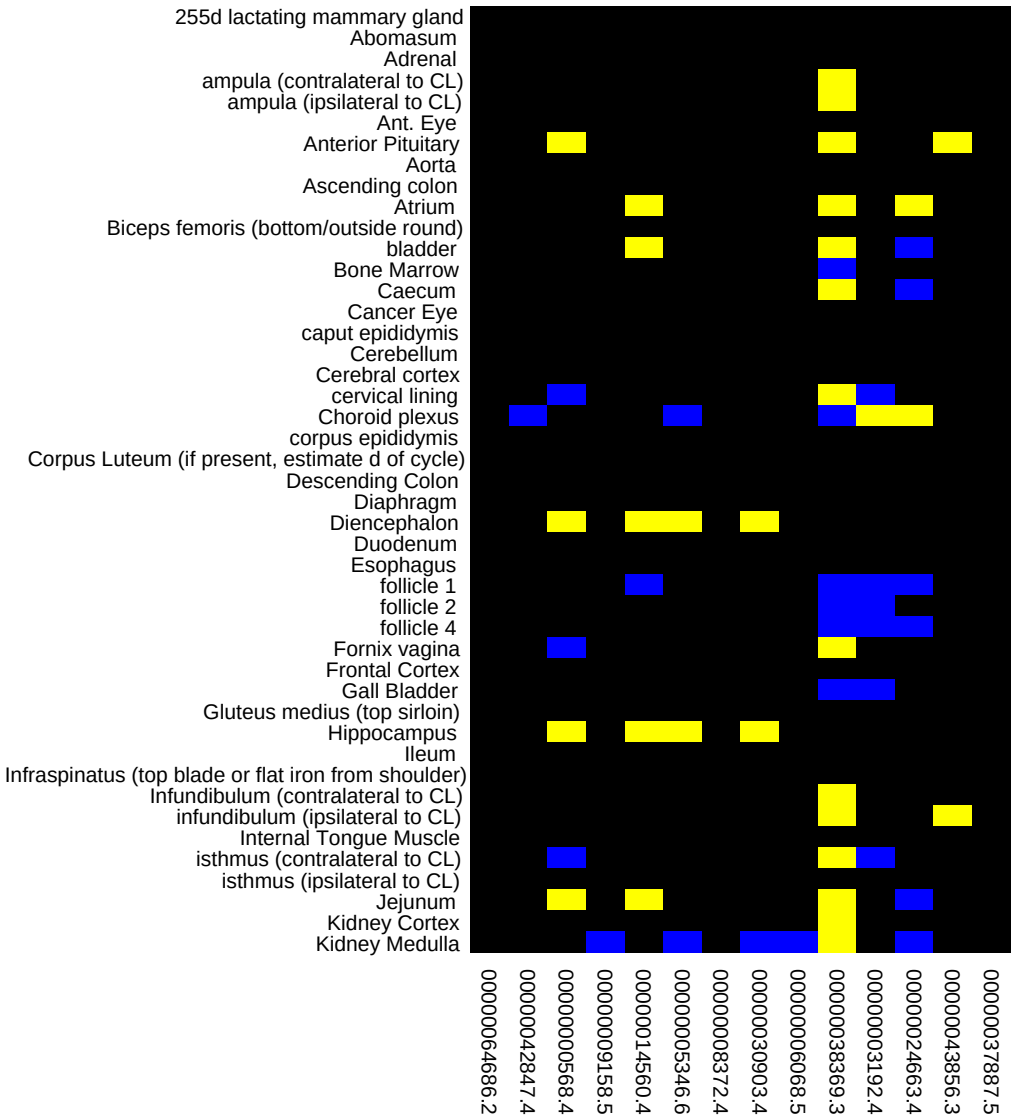
Figure 6.3: BranchOut reconstruction of the WW protein family in a lactating mammary gland. Bright yellow coloring corresponds to an "above average expression" classification for a node. Transcript identifiers have been truncated to omit the prefix "ENSBTAT".

Intriguingly, some WW-domain proteins have been implicated in breast tumors in mammals (Jamous, Salah 2018).

6.3.2 PF00091 Tubulin FtsZ family: GTPase.domain

The tubulin domain is known best for its role in the cytoskeleton. Orthologs of tubulins exist in bacteria, where the protein family is known as FtsZ. This family

is defined by presence of the GTPase domain in this overall structure. A total of 14 members in this family were included in this analysis. According to BranchOut summary scores, the tissues with the strongest evolutionary signals were the gall bladder, ovarian follicle and vas deferens. A summary of the expression classification can be found in Figure 6.4a & b:



(Figure continues on next page)

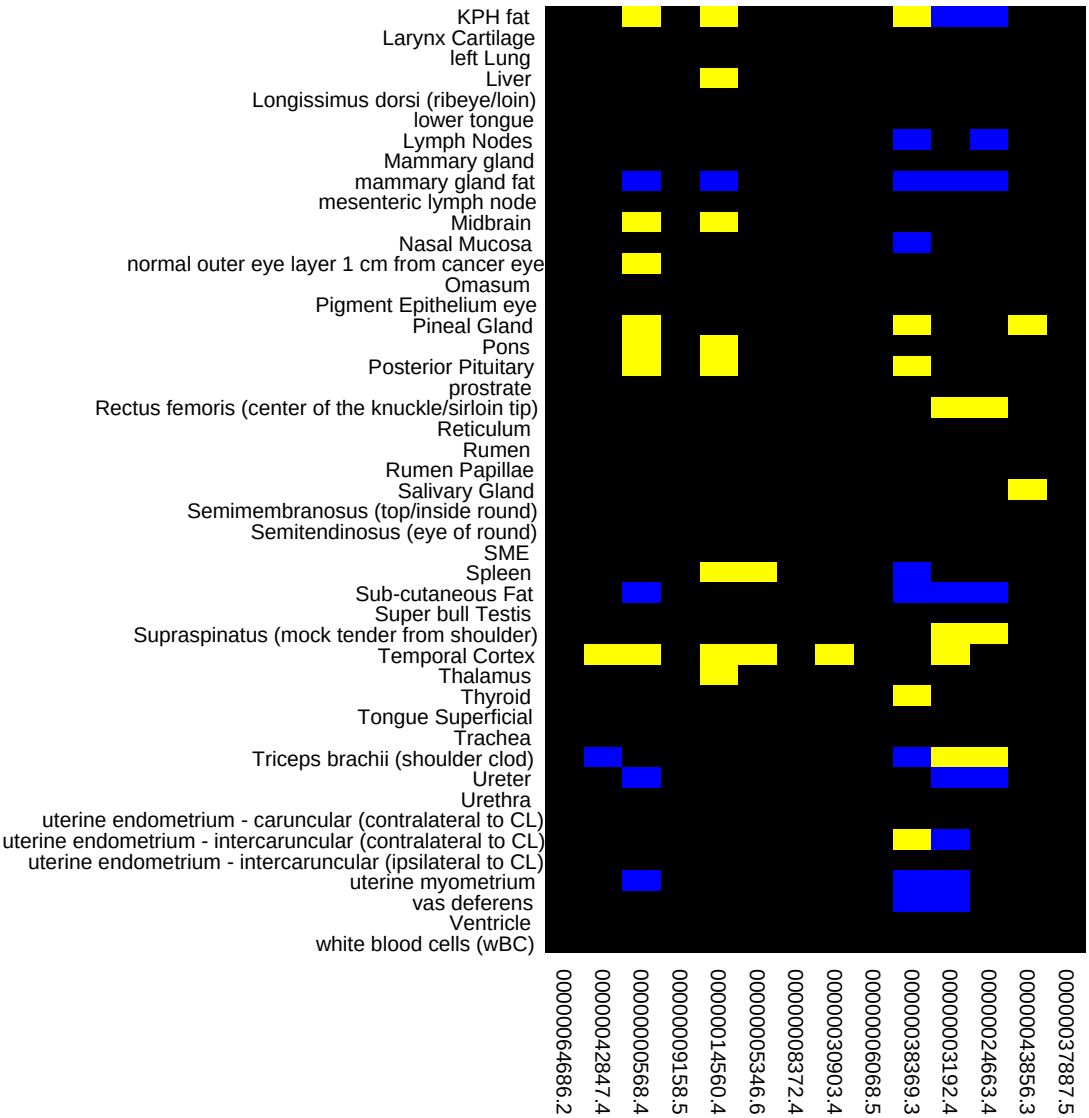


Figure 6.4a & b: Map of Expression Clustering Assignments for the Tubulin FtsZ family: GTPase domain protein family in *Bos taurus*. Yellow coloring was used to indicate up-regulation of a transcript relative to the protein family-average in a given tissue. Blue indicates down-regulation, and black indicates membership to the cluster closest to the median. Transcript order (right-to-left) matches the ordering in subsequent tree figures (top-to-bottom). Transcript identifiers have been truncated to omit the prefix "ENSBTAT".

A group located in the middle of the BranchOut tree diagram shows elevated activity in several regions of the brain, including the temporal lobe, hippocampus

and diencephalon. The temporal cortex specifically brings together elevated expression across all three prominent sub-groups in the phylogenetic tree (see Figure 6.5), and is the only tissue in which the bottom-most genes, based on the orientation in the figure, show differential regulation.

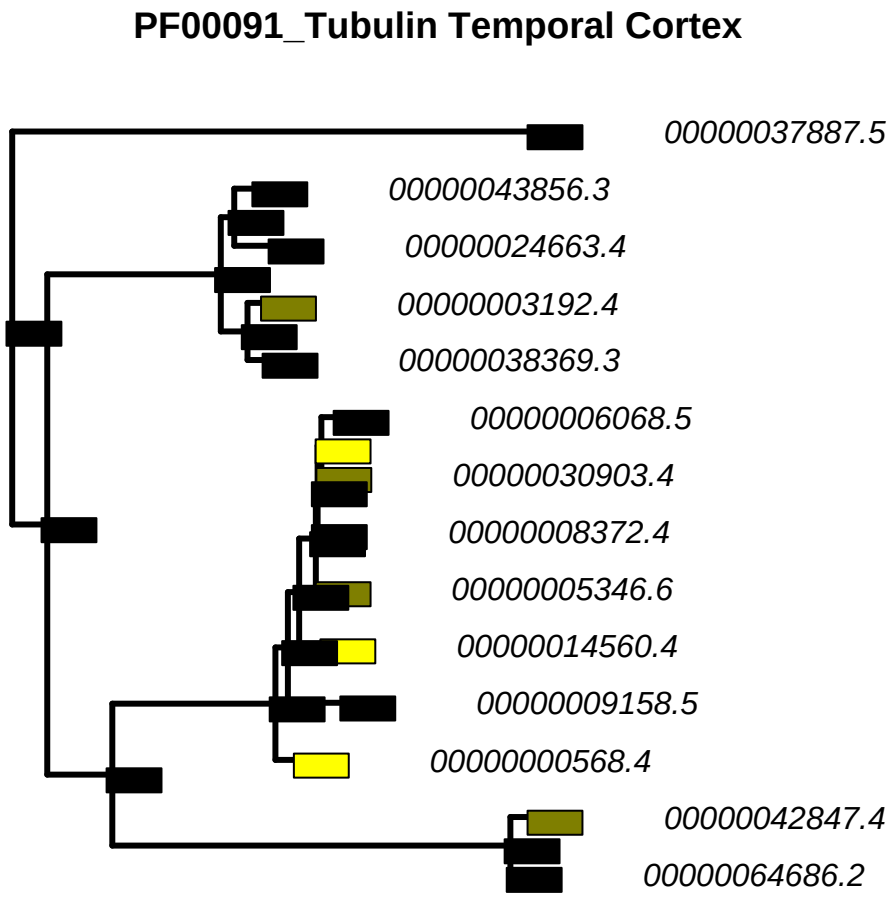


Figure 6.5: BranchOut reconstruction of the Tubulin protein family in the temporal cortex. Dark and bright yellow colors correspond to "moderate-" and "highly-above average expression" classifications for nodes, respectively. This distribution was typical of several related brain and nervous tissues. Transcript identifiers have been truncated to omit the prefix "ENSBTAT".

The transcripts in the top group often show elevated expression, together with transcripts near the base of the middle sub-group particularly in brain-proximate glands, like the pineal gland and the pituitary gland. However, a pair of transcripts in the top subgroup, ENSBTAT00000024663 and ENSBTAT00000003192, instead show elevated expression exclusively in muscle tissue (see Figure 6.6).

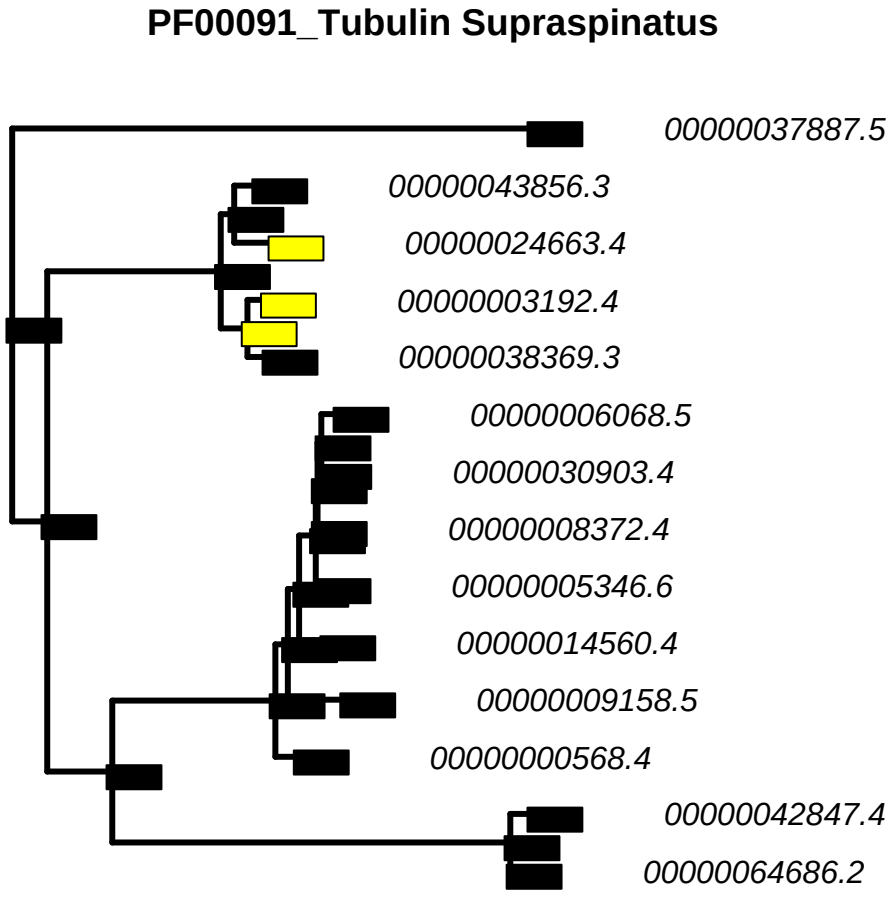
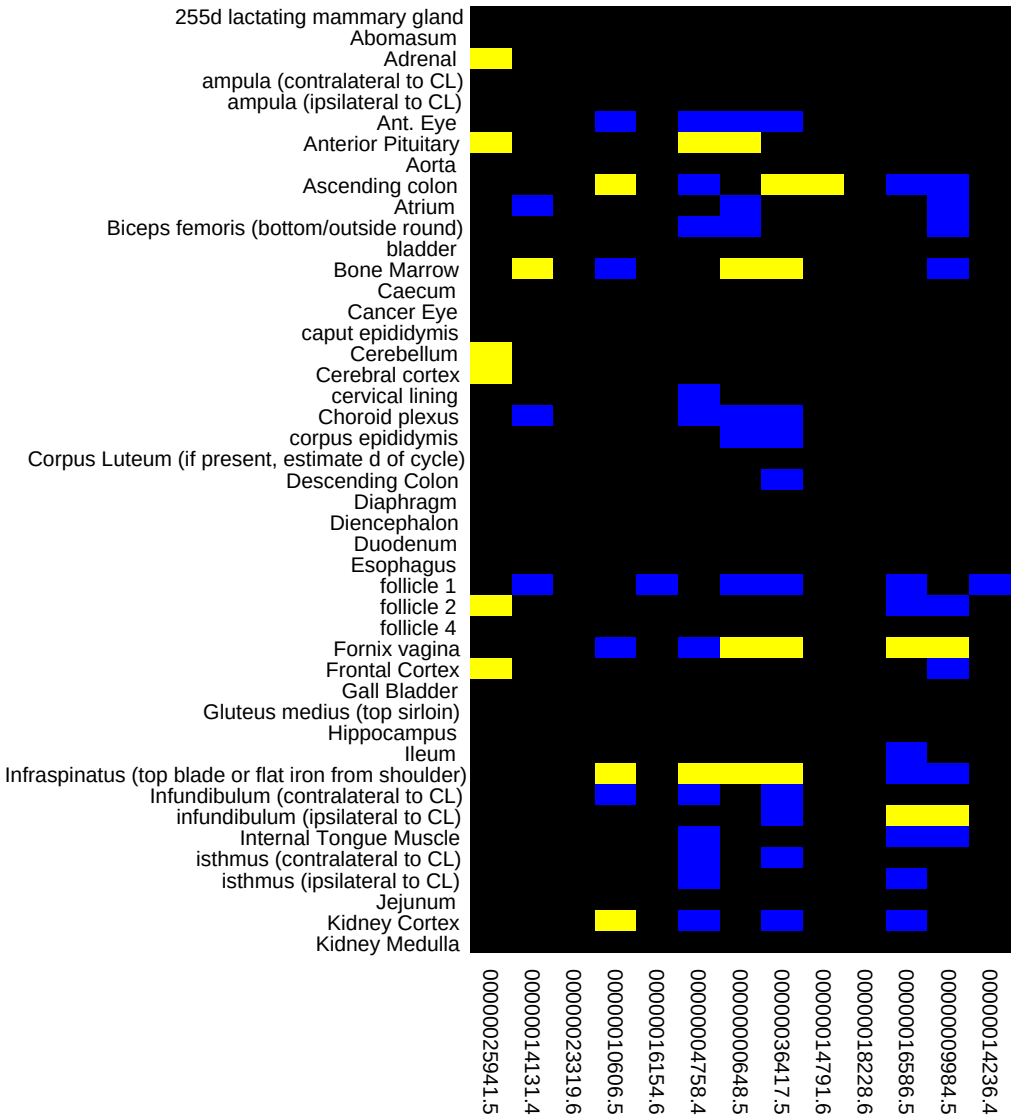


Figure 6.6: BranchOut reconstruction of the Tubulin protein family in the supraspinatus. Bright yellow coloring corresponds to an "above average expression" classification for a node. Transcript identifiers have been truncated to omit the prefix "ENSBTAT".

Transcripts including this region split into multiple phylogenetically distinct subgroups. Given that the unifying expression trend seems to be in endocrine-related, excretion-driven systems, it is somewhat surprising to see the structural role of tubulin (and its plausible relationship with musculature) relegated to a pair of genes nestled within a subgroup that seems to show no bias towards supportive tissue. For large mammals like cows, it may be possible that tubulin proteins play a role in maintaining very elongated (nerve) cells and tissues (Strocchi, Brown et al. 1981).

6.3.3 PF00104: Ligand binding domain of nuclear hormone receptor

The nuclear hormone receptor family is involved in differential gene regulation, playing a key role in development, homeostasis and metabolism. A total of 13 members of this family were included in the *Bos taurus* analysis. According to BranchOut summary scores, the tissues with the strongest evolutionary signals were the rumen, ovarian follicle, infundibulum, and pituitary gland. A summary of the expression classification can be found in Figure 6.7a & b:



(Figure continues on next page)

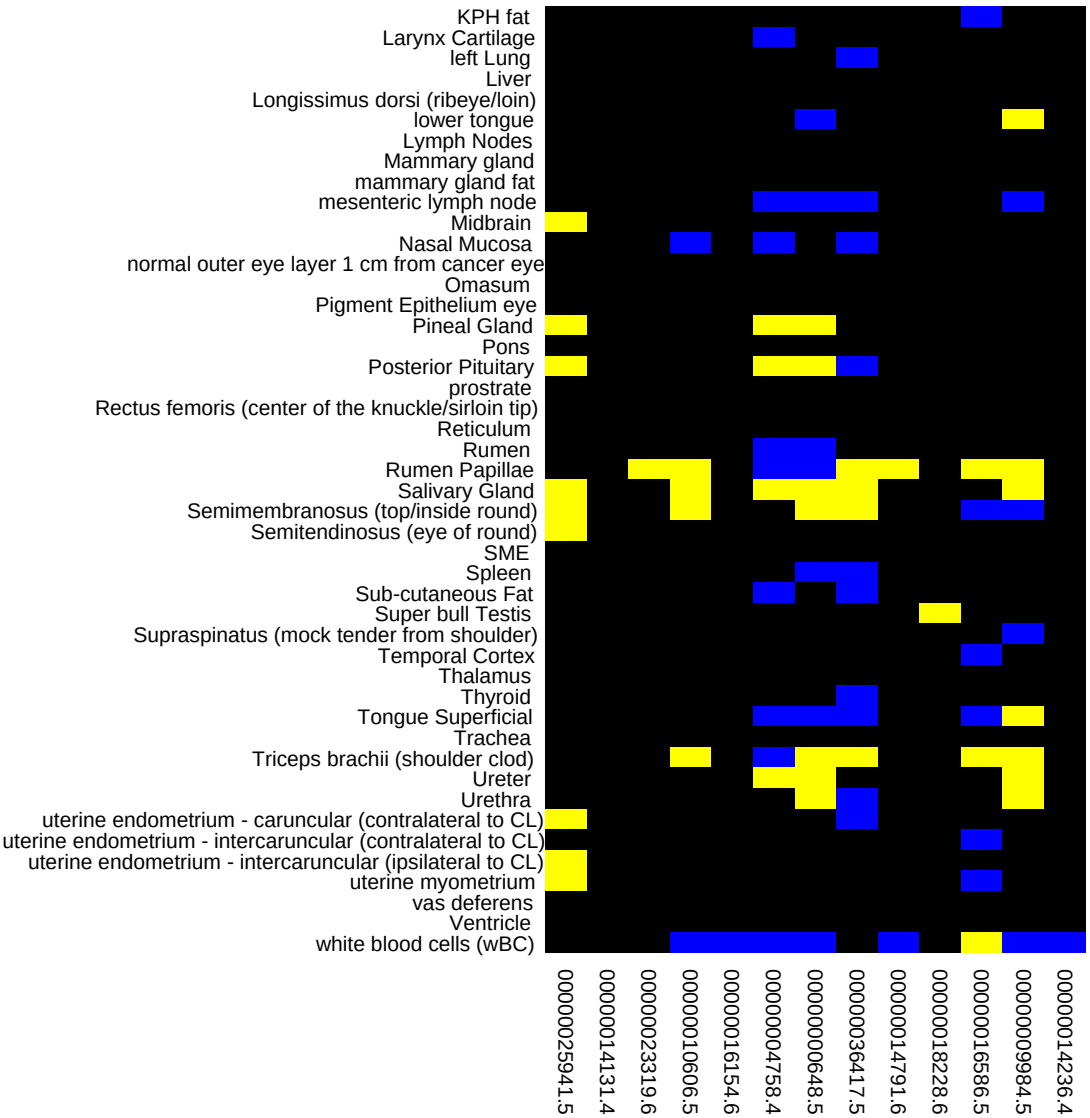


Figure 6.7a & b: Map of Expression Clustering Assignments for the Ligand binding domain of nuclear hormone receptor protein family in *Bos taurus*. Yellow coloring was used to indicate up-regulation of a transcript relative to the protein family-average in a given tissue. Blue indicates down-regulation, and black indicates membership to the cluster closest to the median. Transcript order (right-to-left) matches the ordering in subsequent tree figures (top-to-bottom). Transcript identifiers have been truncated to omit the prefix "ENSBTAT".

Elevated expression behavior in the rumen covers nearly the entire "central core" of the BranchOut tree (figure 6.8) (transcript ENSBTAT00000016154 did not show differential regulation in any of the tissues in this study).

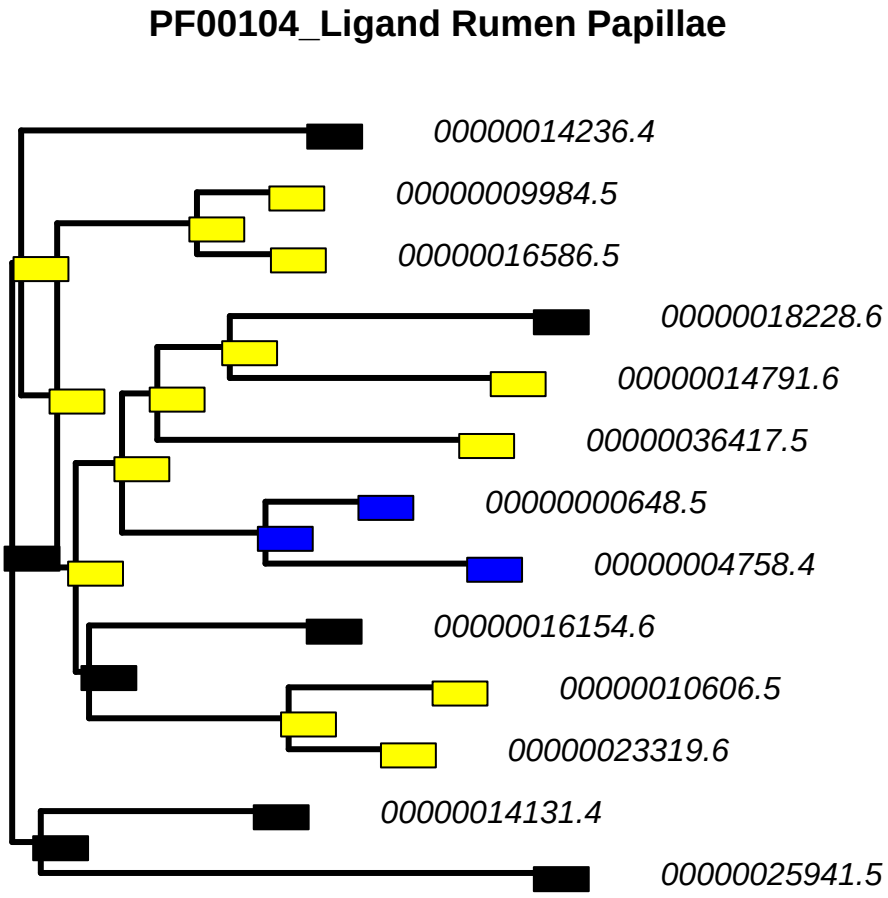


Figure 6.8: BranchOut reconstruction of the ligand binding domain of nuclear hormone receptor family in the rumen. Bright yellow coloring corresponds to an "above average expression" classification for a node. Blue indicates down-regulation, and black indicates membership to the cluster closest to the median. Transcript identifiers have been truncated to omit the prefix "ENSBTAT".

Two of the remaining exceptions, ENSBTAT00000000648 and ENSBTAT00000016154, showed specific evidence of transcript suppression in this otherwise active tissue. Further exploration revealed that these two transcripts instead seem to play a role in the endocrine system (see Figure 6.9).

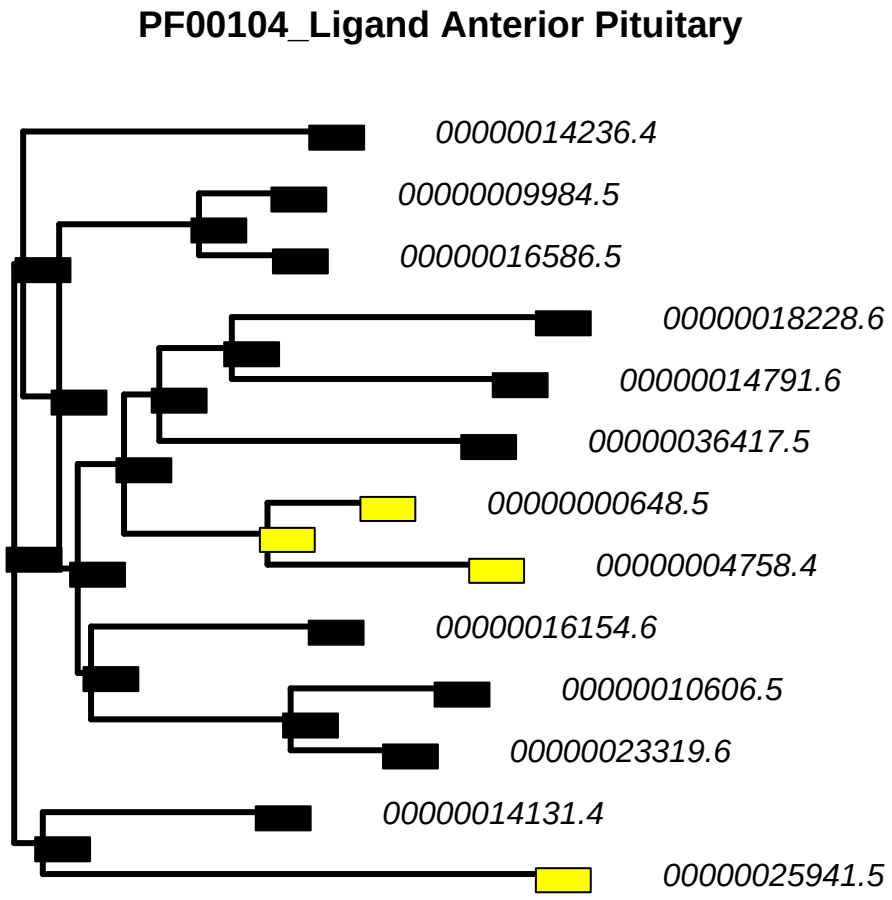


Figure 6.9: BranchOut reconstruction of the ligand binding domain of nuclear hormone receptor family in the pituitary gland. Bright yellow coloring corresponds to an "above average expression" classification for a node. Transcript identifiers have been truncated to omit the prefix "ENSBTAT".

This same split (pulling in the bottom-most transcript for elevated expression) was also present in the pineal gland and posterior pituitary gland. The bottom-most transcript is the only member of this family to show elevated expression in other brain regions as well.

Intermediate to these two extremes, we can see elevated expression across the transcript family in the salivary gland (Figure 6.10).

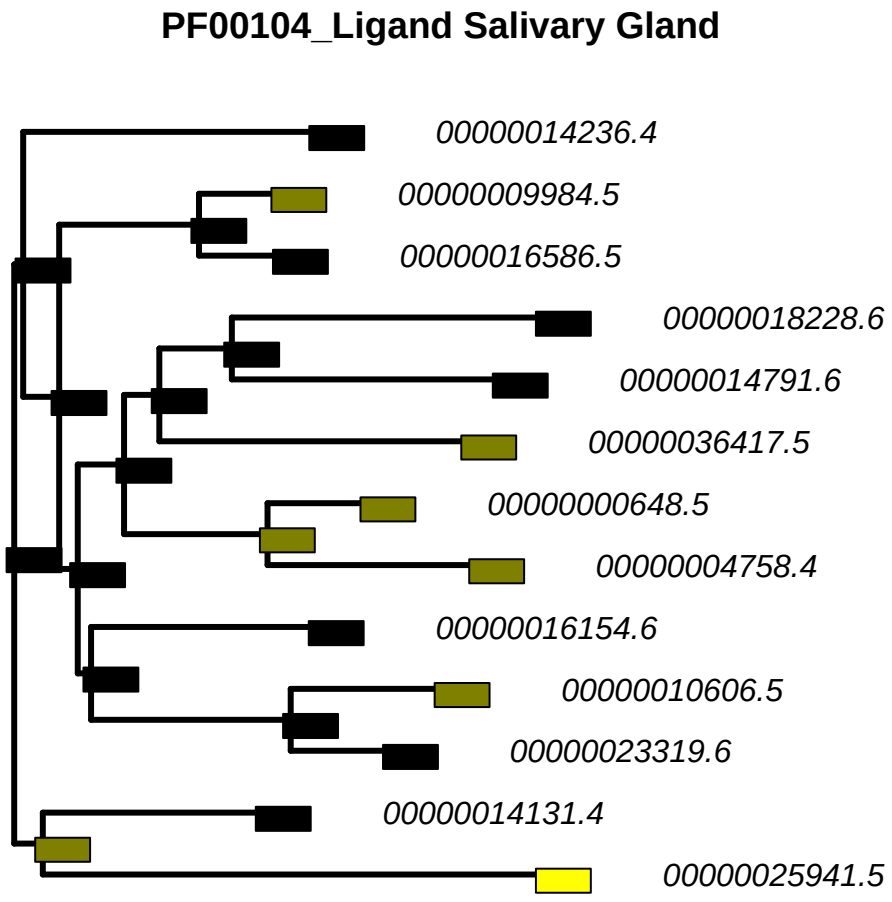


Figure 6.10: BranchOut reconstruction of the ligand binding domain of nuclear hormone receptor family in the salivary gland. Dark and bright yellow colors correspond to "moderate-" and "highly-above average expression" classifications for nodes. Transcript identifiers have been truncated to omit the prefix "ENSBTAT".

This tissue shows elevated activation both for the pair of transcripts that were silenced in the rumen, but also for another pair of transcripts that show elevated expression in the endocrine-associated infundibulum (Figure 6.11).

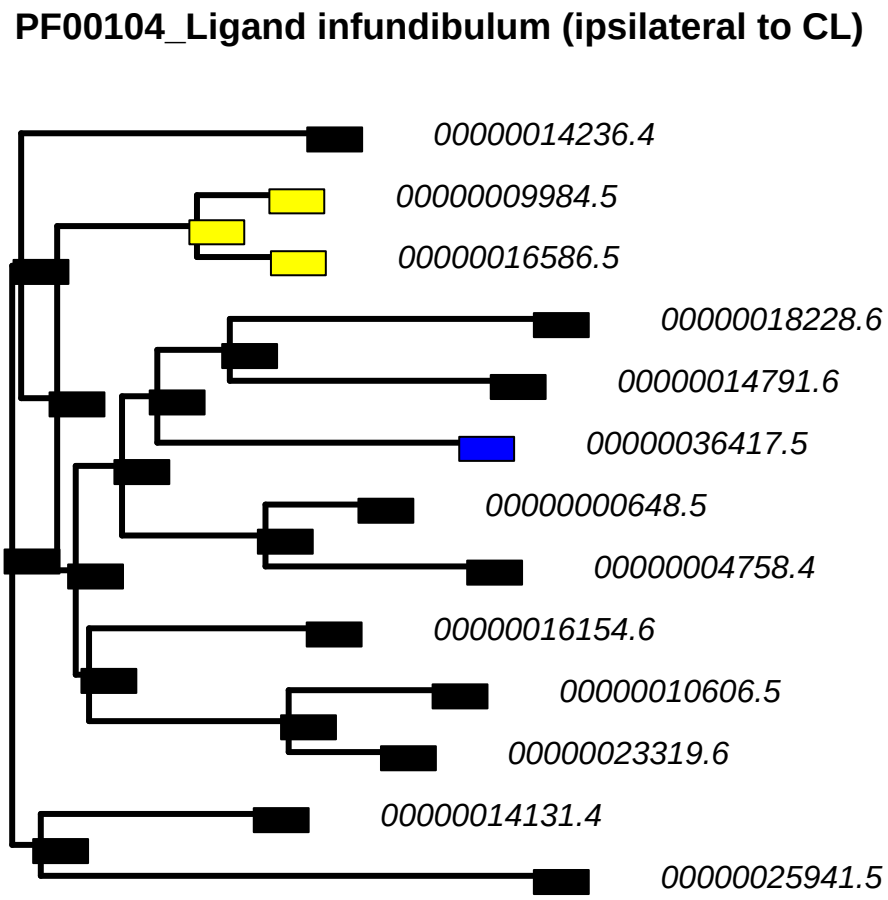


Figure 6.11: BranchOut reconstruction of the ligand binding domain of nuclear hormone receptor family in the infundibulum. Bright yellow coloring corresponds to an "above average expression" classification for a node. Bright blue coloring corresponds to "below average expression". Transcript identifiers have been truncated to omit the prefix "ENSBTAT".

Duplications within this family allowed the production of paralogous genes that, although sharing a common overall metabolism function, include some members that are highly specialized towards specific tissues (that also happen to be most closely related) while also often retaining a function in a standard milieu (rumen/nutrient detection).

6.4 Summary of *Bos taurus* BranchOut analysis

The large quantity of missing data inherent to high-throughput sequencing data presented a unique challenge for BranchOut in this analysis. Whereas microarrays will always produce a residual background signal for every probe included on the assay, absent transcripts generate no records of transcription whatsoever. In the data set examined, "missing" transcripts seemed to have been handled variably; sometimes absent transcripts were assigned a 0, sometimes they were assigned a value near machine precision (1.0x10^-30), and sometimes they were not included at all. In this analysis, BranchOut was coded to use a minimum expression value of 1 as a minimum value, and only a core set of transcripts present across most tissues were included in the analysis. This may have contributed to the smaller number of families and the lower BranchOut scores observed overall.

Closer inspection of the expression cluster assignment revealed that undesired behavior in families where a large number of values were imputed to the minimum value of 1. Because the cluster assignment is based on modeling a mixture of normal distributions, clusters of values at the floor seem to have zero apparent standard deviation, raising the threshold to an unachievable standard for

all expression data above the floor. As a result, in a small number of circumstances no productive reconstruction was possible.

Although there may be room for improvement in the handling of high-throughput sequencing data, the examples shown here suggest that BranchOut produces hypotheses about the evolutionary refinement of gene function in the cow. From Table 6.1 it is reassuring to see a relationship between lymph nodes and immunoglobulins. Similar to the pig analysis, the testes also ranks highly as a tissue subject to functional specialization. The relatively high ranking of tongue muscle in both Table 6.1 and Table 6.2 may suggest adaptations to an alternative diet through domestication, as may the high-ranking interaction between immunoglobulin and the salivary gland.

Chapter 7: Conclusions and Future Directions

7.1 The Future of the BranchOut Approach: Potential Refinements

When designing BranchOut, I was initially unsure about how many expression categories to expect for a given gene family/tissue combination. In deference to this uncertainty, I left the number of possible cluster categories largely open-ended, with a seldom-reached maximum of nine possible categories. In practice, the number of assigned categories rarely exceeded three, and in circumstances where more than three categories were assigned, the gain in precision seldom seemed informative. In the future, I might be inclined to experiment with a smaller number of possible clusterings (i.e. at most "UP", "DOWN", and "NEUTRAL"); this might be adequate given the exploratory nature of the software. The parameters of the clustering itself can be adjusted to make the software more- or less-inclined to split a set of expression values into separate clusters based on their relative proximity. It might be worth exploring the interplay between the number of possible clusters and the sensitivity of the software to clustering by focusing on a gene family for which a known history of sub- and/or neo-functionalization has been well-established.

BranchOut also assumes, by default, that all expression category changes should be treated as equally probable. In other words, the advent of a new expression profile (early) in a phylogenetic tree is treated as being as likely as the subsequent loss of that function, and transitions from "down-regulated relative to the median" to "up-regulated relative to the median" are regarded as equally

plausible as a shift from one up-regulation category to another. When considering these events at the genomic level, this assumption seems implausible – the acquisition of regulatory elements that would direct expression in a novel tissue seems like a far less probable event that the loss of such a characteristic, particularly when one considers that the loss could be considered a reversion to an expression paradigm (and biochemical role) that has a comparatively long evolutionary history. Moreover, the importance of distinguishing between being "slightly upregulated" and "highly upregulated" is not obvious, particularly in circumstances where the ultimate role of the encoded protein could be quite different from one tissue to the next. In most of the cases examined in this work, such a distinction does not appear to have been necessary (as most findings correspond to a mostly-monophyletic origin at some point in the tree), but this may reflect the settings of the analysis more than it does biological reality.

In this iteration of BranchOut, a maximum-likelihood algorithm was used to determine the ancestral expression states of each protein family. It may be interesting to explore other character reconstruction algorithms, including parsimony and variants thereof. Given that one of the primary goal of BranchOut is to elucidate innovation of gene function, an implementation using Dollo parsimony (where novel departures from the ancestral expression state are only allowed once, though reversion is allowed) might apply a stringency to the protein family analysis that could bring major functional changes to the foreground (Rogozin, I. Wolf et al. 2006).

The greatest challenge to using BranchOut comes from the preprocessing step. Depending on the platforms used to collect expression and sequence data, it is often not easy to compile a tissue expression profile that is uniquely and unambiguously mapped to a single protein product with a well-established genetic/genomic locus. This may become easier over time as bioinformatics continues to mature as a field, but for now these annotation efforts require a great deal of project-specific code that is not easily re-used. It is my hope that efforts like the work presented in this dissertation will encourage stewards of expression atlases to present expression information not just by gene (or by expression cluster), but also by family.

There is some uncertainty about the implications of treating all expressed transcripts as unique evolutionary entities. One alternative approach would be to include only one translated product per locus, in an effort to make the gene family phylogeny a construct where all leaves were, in some sense, "comparable". For purposes where the phylogeny/phylogenetic tree must be as exact as possible, more triage could be done to ensure all "leaves" have unique genomic origins. This would require considerable manual effort at present, however, so I chose instead to focus on the ability to process a whole transcriptome with minimal assumptions.

There is also room for improvement in how BranchOut handles missing data, particularly in the case of high-throughput sequencing data sets. The cow expression atlas was somewhat inconsistent in how it reported the absence of transcription, with effectively-off transcripts variably getting scores of zero,

machine precision (i.e. 1.0x10^-60) or being excluded from the results entirely. In this analysis, the results were restricted to the set of transcripts that were present in all tissue samples because a) this set was still quite large at 25000+ transcripts and b) because it is not immediately obvious how to impute a "off" score for the missing data. In particular, there is some risk in the expression clustering step: if a large number of transcripts are all assigned a floor/off expression state, these transcripts will encourage the construction of a normal distribution with mean equal to the floor value and with zero apparent standard deviation. This in turn raises the bar for all subsequent cluster definition assignments, resulting in undesired behavior (e.g. all transcripts to one cluster regardless of distribution, or each transcript being its own cluster). One possible workaround would be to apply the clustering only to those transcripts that are clearly above the signal detection threshold (1 in this analysis), and then to assign all noise/absent transcripts to the lowest generated expression category (corresponding to the best available "off" state). This would ensure that any fitted normal distributions would be based on the signal variation present in the reliable data with minimal impact from floored, identical values.

BranchOut also operates under the assumption that the multiple sequence alignment and associated phylogenetic trees are trustworthy. This may not always be the case, particularly when the entire transcriptome is subjected to a largely unsupervised analysis (here by MUSCLE and then PhyML) using only program defaults. Moreover, even with these settings, PhyML can provide indices of branch support which could be used to annotate the tree, making it potentially easier to identify which BranchOut findings are likely to be the consequence of arbitrary

decisions made in tree construction. Moreover, it may be possible to build the branch support values into the character reconstruction process itself, either by preferring some state assignments over others (based on the trustworthiness of a subtree to its parent) or by having the scoring scheme used to identify analysis highlights apply a punitive measure to reconstructions that cross very poorly supported branches.

In the analyses included in this dissertation, I went to great lengths to include as many gene families and tissues as possible while constraining the analyses to use a common set of preprocessing programs and analysis parameters. Realistically, any users of the software who detect a potentially interesting finding will later revisit these decisions with much more care. For example, it is common practice to examine the output of multiple sequence alignment software with the intent of pruning out regions that are likely to be uninformative. This level of care is not feasible when a transcriptome is being considered at scale, but could be warranted as a first step toward following up on a BranchOut finding. It is difficult to predict, on a case-by-case basis, how modifying the preprocessing decisions made by BranchOut will impact the results of the analysis overall. It is possible, however, to examine the effects when toggling these preprocessing decisions on the entire transcriptome.

To explore the potential for multiple sequence alignment and phylogenetic tree algorithmic decisions to impact BranchOut performance, several additional runs of the BranchOut software were conducted on a slightly restricted set of 249

protein families from *Sus scrofa* (selected for small size, see Appendix D for a complete listing of families). Two additional algorithms were introduced as candidates: Clustal Omega (Sievers, Higgins 2014) v1.2.2 (for Windows) as an alternative multiple sequence alignment program, and BAli-Phy (Suchard, Redelings 2006) v3.4.1 (for Windows) as an alternative phylogenetic tree algorithm. For each of the 249 families included in this supplementary analyses, each possible combination of multiple sequence alignment and phylogenetic algorithms was carried out. MUSCLE, PhyML and Clustal Omega were run with default parameters. BAli-Phy was directed to trust its multiple sequence alignment input (as it would otherwise iteratively modify both the alignment and the tree), to use a "gtr+Rates.gamma[4]+inv" sequence model, to run 1200 iterations, and then build a greedy consensus tree based on the last 1000 iterations (discarding the first 200 as "burn-in"). It was assumed that 1200 iterations was sufficient to achieve convergence in tree topology for each family.

The BranchOut scores from each run were collected and compared pairwise. Pearson correlation coefficients varied from a high of 0.834 (Clustal Omega sequence alignment, PhyML vs BAli-Phy trees) to a low of 0.407 (MUSCLE + PhyML vs Clustal Omega + BAli-Phy). The correlation scores are high (.834, .670) when the comparison keeps the multiple sequence alignment constant but varies the phylogenetic analysis. The scores are more moderate (.407-.426) when different multiple sequence alignment programs are compared. Based on these results, the choice of multiple sequence alignment program seems to have a large influence on the robustness of the results, but with the caveat that families that achieve high

BranchOut scores tend to do so irrespective of pre-processing decisions. The impact of phylogenetic tree software, by contrast, is relatively low, despite PhyML using a maximum likelihood framework and BAli-Phy using a Bayesian approach. Pairwise scatterplots annotated with Pearson correlation scores are shown in Figure 7.1.

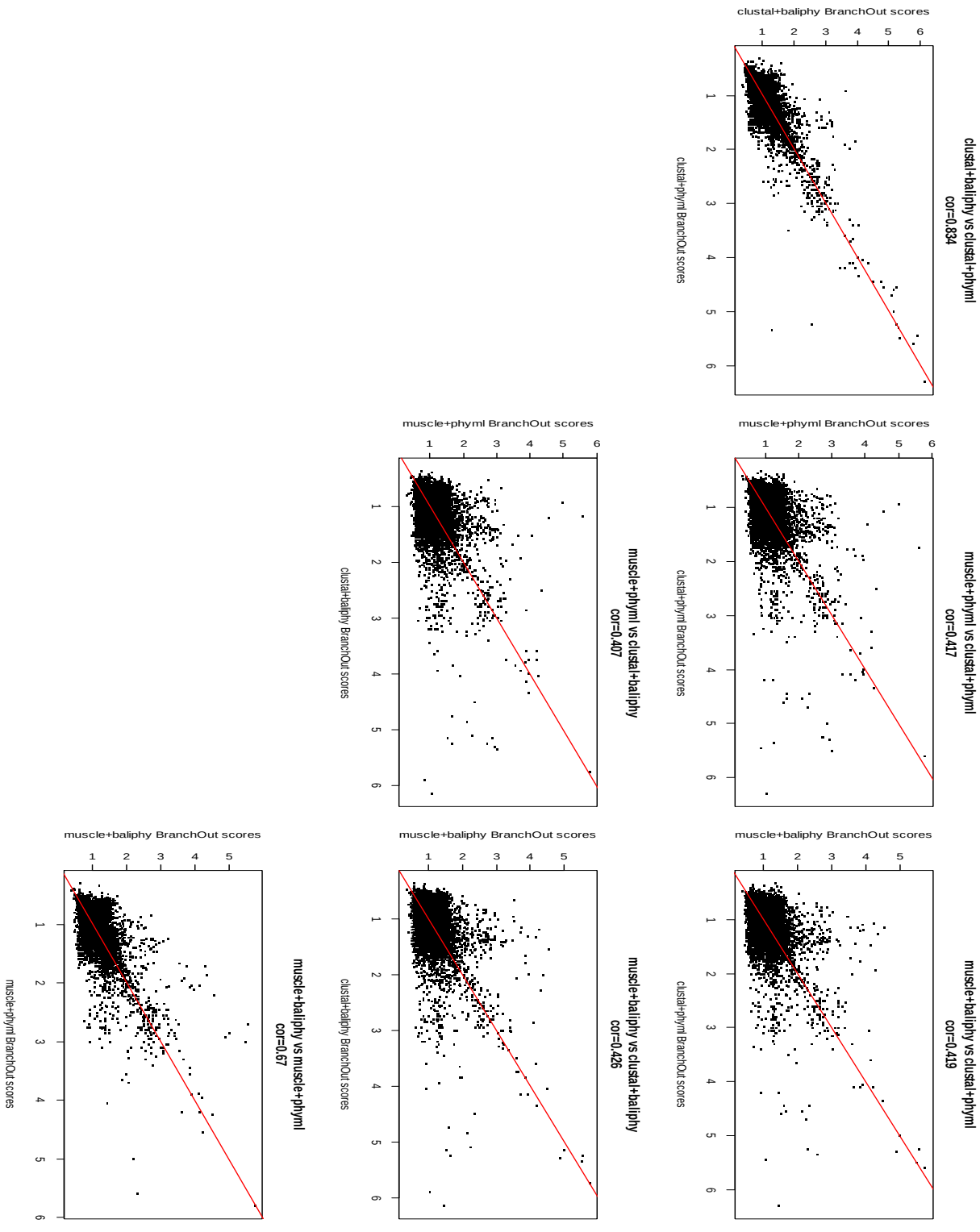


Figure 7.1: Pairwise comparisons of BranchOut scores with varying input sources. Unity line is shown in red. Pearson correlations of BranchOut scores are reported in the subtitles.

The results of these algorithmic comparisons seem to indicate that users who wish to evaluate the robustness of their BranchOut findings should focus more on the multiple sequence alignment step than on the subsequent development of the phylogenetic tree. Consider, for example, that the BAli-Phy software encourages its users to monitor the output of the program on a tree-by-tree basis, as it is not possible to determine a-priori how many runs will be required before iterations of the program converge on a common topology. Instead, this follow-up study simply assigned a common, fixed number of iterations to each tree estimation task. In spite of this, the results were comparable to the output of PhyML, a program that requires no user interaction, allowed to run to completion. Even under these conditions, the choice of tree estimation software seemed to have a comparatively minor impact, suggesting that a user attempting to validate a BranchOut finding would get more out of their time by scrutinizing the multiple sequence alignment first and foremost.

The robustness of BranchOut scores to preprocessing decisions provides an early indication of how the trustworthiness of BranchOut-derived hypotheses could be determined. It is, as of yet, unclear whether this additional work would be a valuable first step, or whether BranchOut findings are sufficiently trustworthy to be followed up directly (e.g. with wet-lab studies, or analyses involving ancestral sequence reconstruction to determine a hypothetical evolutionary history to the overall function of the protein scaffold in question). It would be very interesting to

see the results of several such confirmatory studies, but the feasibility and protocol for these follow-up studies remains to be determined.

Lastly, the signal score produced by BranchOut, which is based on the ratio of state changes in "real" versus "scrambled" transcript cluster assignments, has some potential room for improvement. While a score of 1.5 was generally found to exceed the distribution of "null" scores, it was possible to generate a score of 1.5 in gene family/tissue pairings that included only a single off-mode signaling label on a long branch. Many reconstructions would place the outlying label on the shorter of two sibling branches, leading BranchOut to infer that the altered state was ancestral. This situation always generated a state change score of "2" versus the true assignment's score of "1". Even though the scrambled score is an average of many repeated trials, certain tree shapes were highly susceptible to this score inflating pattern.

7.2 Closing Thoughts on BranchOut as a tool for Data Mining and Visualization

The BranchOut approach has a lot of merit as an exploratory data analysis tool. Being able to explore gene function and sequence evolution in tandem opens up a new avenue for hypothesis generation. Moreover, the manner in which BranchOut explores these hypotheses on a tissue-by-tissue basis avoids some of the common pitfalls of cluster-based functional analysis. Specifically, clustering the expression table as a whole unites genes based on common expression, often at the expense of noting unique deviations from common patterns restricted to a limited

number of tissues. BranchOut brings these unique patterns to the forefront, making it a valuable analytical tool for this purpose.

Bibliography

AHALAWAT, N. and MONDAL, J., 2018. Mapping the Substrate Recognition Pathway in Cytochrome P450. *Journal of the American Chemical Society,* **140**(50), pp. 17743-17752.

AKITAYA, T., TSUMOTO, K., YAMADA, A., MAKITA, N., KUBO, K. and YOSHIKAWA, K., 2003. NTP concentration switches transcriptional activity by changing the large-scale structure of DNA. *Biomacromolecules,* **4**(5), pp. 1121-1125.

ANISIMOVA, M., GIL, M., DUFAYARD, J.F., DESSIMOZ, C. and GASCUEL, O., 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Systematic Biology,* **60**(5), pp. 685-699.

ARNAIZ, O., GOUT, J.F., BETERMIER, M., BOUHOUCHE, K., COHEN, J., DURET, L., KAPUSTA, A., MEYER, E. and SPERLING, L., 2010. Gene expression in a paleopolyploid: a transcriptome resource for the ciliate Paramecium tetraurelia. *BMC genomics,* **11**, pp. 547.

ASHBURNER, M., BALL, C.A., BLAKE, J.A., BOTSTEIN, D., BUTLER, H., CHERRY, J.M., DAVIS, A.P., DOLINSKI, K., DWIGHT, S.S., EPPIG, J.T., HARRIS, M.A., HILL, D.P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J.C., RICHARDSON, J.E., RINGWALD, M., RUBIN, G.M. and SHERLOCK, G., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics,* **25**(1), pp. 25-29.

BAILEY, K.L. and CARLSON, M.A., 2019. Porcine Models of Pancreatic Cancer. *Frontiers in oncology,* **9**, pp. 144.

BARCHUK, A.R., CRISTINO, A.S., KUCHARSKI, R., COSTA, L.F., SIMOES, Z.L. and MALESZKA, R., 2007. Molecular determinants of caste differentiation in the highly eusocial honeybee Apis mellifera. *BMC developmental biology,* **7**, pp. 70-213X-7-70.

BARKMAN, T. and ZHANG, J., 2009. Evidence for escape from adaptive conflict? *Nature,* **462**(7274), pp. E1; discussion E2-3.

BEER, M.A. and TAVAZOIE, S., 2004. Predicting gene expression from sequence. *Cell,* **117**(2), pp. 185-198.

BENSON, D.A., CAVANAUGH, M., CLARK, K., KARSCH-MIZRACHI, I., LIPMAN, D.J., OSTELL, J. and SAYERS, E.W., 2017. GenBank. *Nucleic acids research,* **45**(D1), pp. D37-D42.

BIRCHLER, J.A. and VEITIA, R.A., 2019. Genomic Balance and Speciation. *Epigenetics insights,* **12**, pp. 2516865719840291.

BLANC, G. and WOLFE, K.H., 2004. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *The Plant cell,* **16**(7), pp. 1679-1691.

BOLSTAD, B.M., IRIZARRY, R.A., ASTRAND, M. and SPEED, T.P., 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England),* **19**(2), pp. 185-193.

BRENNER, S., JOHNSON, M., BRIDGHAM, J., GOLDA, G., LLOYD, D.H., JOHNSON, D., LUO, S., MCCURDY, S., FOY, M., EWAN, M., ROTH, R., GEORGE, D., ELETR, S., ALBRECHT, G., VERMAAS, E., WILLIAMS, S.R., MOON, K., BURCHAM, T., PALLAS, M., DUBRIDGE, R.B., KIRCHNER, J., FEARON, K., MAO, J. and CORCORAN, K., 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature biotechnology,* **18**(6), pp. 630-634.

BUTTSTEDT, A., MURESAN, C.I., LILIE, H., HAUSE, G., IHLING, C.H., SCHULZE, S.H., PIETZSCH, M. and MORITZ, R.F.A., 2018. How Honeybees Defy Gravity with Royal Jelly to Raise Queens. *Current biology : CB,* **28**(7), pp. 1095-1100.e3.

CALIEBE, A., NEBEL, A., MAKAREWICZ, C., KRAWCZAK, M. and KRAUSE-KYORA, B., 2017. Insights into early pig domestication provided by ancient DNA analysis. *Scientific reports,* **7**, pp. 44550.

CANNON, S.B., MITRA, A., BAUMGARTEN, A., YOUNG, N.D. and MAY, G., 2004. The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana. *BMC plant biology,* **4**, pp. 10.

CASNEUF, T., DE BODT, S., RAES, J., MAERE, S. and VAN DE PEER, Y., 2006a. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant Arabidopsis thaliana. *Genome biology,* **7**(2), pp. R13.

CASNEUF, T., DE BODT, S., RAES, J., MAERE, S. and VAN DE PEER, Y., 2006b. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant Arabidopsis thaliana. *Genome biology,* **7**(2), pp. R13.

CHAIN, F.J. and EVANS, B.J., 2006. Multiple mechanisms promote the retained expression of gene duplicates in the tetraploid frog Xenopus laevis. *PLoS genetics,* **2**(4), pp. e56.

COATE, J.E., SONG, M.J., BOMBARELY, A. and DOYLE, J.J., 2016. Expression-level support for gene dosage sensitivity in three Glycine subgenus Glycine polyploids and their diploid progenitors. *The New phytologist,* **212**(4), pp. 1083-1093.

COMELLI, R.N. and GONZALEZ, D.H., 2009. Divergent regulatory mechanisms in the response of respiratory chain component genes to carbohydrates suggests a model

for gene evolution after duplication. *Plant signaling & behavior,* **4**(12), pp. 1179-1181.

CULLEN, J.M., LU, G., SHANNON, A.H., SU, G., SHARMA, A., SALMON, M., FASHANDI, A.Z., SPINOSA, M.D., MONTGOMERY, W.G., JOHNSTON, W.F., AILAWADI, G. and UPCHURCH, G.R.,Jr, 2018. A novel swine model of abdominal aortic aneurysm. *Journal of vascular surgery,* .

CVIJOVIC, I., GOOD, B.H. and DESAI, M.M., 2018. The Effect of Strong Purifying Selection on Genetic Diversity. *Genetics,* **209**(4), pp. 1235-1278.

DAUGAARD, M., ROHDE, M. and JAATTELA, M., 2007. The heat shock protein 70 family: Highly homologous proteins with overlapping and distinct functions. *FEBS letters,* .

DES MARAIS, D.L. and RAUSHER, M.D., 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature,* **454**(7205), pp. 762-765.

DODSON, E.J., FISHBAIN-YOSKOVITZ, V., ROTEM-BAMBERGER, S. and SCHUELER-FURMAN, O., 2015. Versatile communication strategies among tandem WW domain repeats. *Experimental biology and medicine (Maywood, N.J.),* **240**(3), pp. 351-360.

DOXEY, A.C., YAISH, M.W., MOFFATT, B.A., GRIFFITH, M. and MCCONKEY, B.J., 2007a. Functional divergence in the Arabidopsis beta-1,3-glucanase gene family inferred by phylogenetic reconstruction of expression states. *Molecular biology and evolution,* **24**(4), pp. 1045-1055.

DOXEY, A.C., YAISH, M.W., MOFFATT, B.A., GRIFFITH, M. and MCCONKEY, B.J., 2007b. Functional divergence in the Arabidopsis beta-1,3-glucanase gene family inferred by phylogenetic reconstruction of expression states. *Molecular biology and evolution,* **24**(4), pp. 1045-1055.

DRAPEAU, M.D., ALBERT, S., KUCHARSKI, R., PRUSKO, C. and MALESZKA, R., 2006. Evolution of the Yellow/Major Royal Jelly Protein family and the emergence of social behavior in honey bees. *Genome research,* **16**(11), pp. 1385-1394.

DRUMMOND, D.A., BLOOM, J.D., ADAMI, C., WILKE, C.O. and ARNOLD, F.H., 2005. Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America,* **102**(40), pp. 14338-14343.

EDGAR, R.C., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics,* **5**, pp. 113.

EKSER, B., RIGOTTI, P., GRIDELLI, B. and COOPER, D.K., 2009. Xenotransplantation of solid organs in the pig-to-primate model. *Transplant immunology,* **21**(2), pp. 87-92.

FAN, B., GORBACH, D.M. and ROTHSCHILD, M.F., 2011. The pig genome project has plenty to squeal about. *Cytogenetic and genome research,* **134**(1), pp. 9-18.

FELSENSTEIN, J., 1997. An alternating least squares approach to inferring phylogenies from pairwise distances. *Systematic Biology,* **46**(1), pp. 101-111.

FELSENSTEIN, J., 1988. Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics,* **22**, pp. 521-565.

FERGUSON, L.C., GREEN, J., SURRIDGE, A. and JIGGINS, C.D., 2011. Evolution of the insect yellow gene family. *Molecular biology and evolution,* **28**(1), pp. 257-272.

FORET, S., KUCHARSKI, R., PITTELKOW, Y., LOCKETT, G.A. and MALESZKA, R., 2009. Epigenetic regulation of the honey bee transcriptome: unravelling the nature of methylated genes. *BMC genomics,* **10**, pp. 472-2164-10-472.

FOSTER, K.R., WENSELEERS, T. and RATNIEKS, F.L., 2006. Kin selection is the key to altruism. *Trends in ecology & evolution,* **21**(2), pp. 57-60.

FREEMAN, T.C., IVENS, A., BAILLIE, J.K., BERALDI, D., BARNETT, M.W., DORWARD, D., DOWNING, A., FAIRBAIRN, L., KAPETANOVIC, R., RAZA, S., TOMOIU, A., ALBERIO, R., WU, C., SU, A.I., SUMMERS, K.M., TUGGLE, C.K., ARCHIBALD, A.L. and HUME, D.A., 2012. A gene expression atlas of the domestic pig. *BMC biology,* **10**, pp. 90-7007-10-90.

FREILICH, S., MASSINGHAM, T., BLANC, E., GOLDOVSKY, L. and THORNTON, J.M., 2006. Relating tissue specialization to the differentiation of expression of singleton and duplicate mouse proteins. *Genome biology,* **7**(10), pp. R89.

GENTLEMAN, R.C., CAREY, V.J., BATES, D.M., BOLSTAD, B., DETTLING, M., DUDOIT, S., ELLIS, B., GAUTIER, L., GE, Y., GENTRY, J., HORNIK, K., HOTHORN, T., HUBER, W., IACUS, S., IRIZARRY, R., LEISCH, F., LI, C., MAECHLER, M., ROSSINI, A.J., SAWITZKI, G., SMITH, C., SMYTH, G., TIERNEY, L., YANG, J.Y. and ZHANG, J., 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology,* **5**(10), pp. R80.

GIEBELER, N. and ZIGRINO, P., 2016. A Disintegrin and Metalloprotease (ADAM): Historical Overview of Their Functions. *Toxins,* **8**(4), pp. 122.

GOETTEL, W. and MESSING, J., 2010. Divergence of gene regulation through chromosomal rearrangements. *BMC genomics,* **11**, pp. 678.

GU, X., 2004. Statistical framework for phylogenomic analysis of gene family expression profiles. *Genetics,* **167**(1), pp. 531-542.

GU, X. and SU, Z., 2007. Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. *Proceedings of the National Academy of Sciences of the United States of America,* **104**(8), pp. 2779-2784.

GU, X., ZHANG, Z. and HUANG, W., 2005. Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proceedings of the National Academy of Sciences of the United States of America,* **102**(3), pp. 707-712.

GU, Z., NICOLAE, D., LU, H.H. and LI, W.H., 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends in genetics : TIG,* **18**(12), pp. 609-613.

GUAN, Y., DUNHAM, M.J. and TROYANSKAYA, O.G., 2007. Functional analysis of gene duplications in Saccharomyces cerevisiae. *Genetics,* **175**(2), pp. 933-943.

GUINDON, S., DUFAYARD, J.F., LEFORT, V., ANISIMOVA, M., HORDIJK, W. and GASCUEL, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology,* **59**(3), pp. 307-321.

GUINDON, S. and GASCUEL, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology,* **52**(5), pp. 696-704.

HA, M., LI, W.H. and CHEN, Z.J., 2007. External factors accelerate expression divergence between duplicate genes. *Trends in genetics : TIG,* **23**(4), pp. 162-166.

HAGEN, D.E., UNNI, D.R., TAYAL, A., BURNS, G.W. and ELSIK, C.G., 2018. Bovine Genome Database: Tools for Mining the Bos taurus Genome. *Methods in molecular biology (Clifton, N.J.),* **1757**, pp. 211-249.

HARHAY, G.P., SMITH, T.P., ALEXANDER, L.J., HAUDENSCHILD, C.D., KEELE, J.W., MATUKUMALLI, L.K., SCHROEDER, S.G., VAN TASSELL, C.P., GRESHAM, C.R., BRIDGES, S.M., BURGESS, S.C. and SONSTEGARD, T.S., 2010. An atlas of bovine gene expression reveals novel distinctive tissue characteristics and evidence for improving genome annotation. *Genome biology,* **11**(10), pp. R102.

HUANG, H.Y., CHIEN, C.H., JEN, K.H. and HUANG, H.D., 2006. RegRNA: an integrated web server for identifying regulatory RNA motifs and elements. *Nucleic acids research,* **34**(Web Server issue), pp. W429-34.

HUANG, P., KELLER, C.A., GIARDINE, B., GREVET, J.D., DAVIES, J.O.J., HUGHES, J.R., KURITA, R., NAKAMURA, Y., HARDISON, R.C. and BLOBEL, G.A., 2017. Comparative analysis of three-dimensional chromosomal architecture identifies a novel fetal hemoglobin regulatory element. *Genes & development,* **31**(16), pp. 1704-1713.

HUGHES, W.O., OLDROYD, B.P., BEEKMAN, M. and RATNIEKS, F.L., 2008. Ancestral monogamy shows kin selection is key to the evolution of eusociality. *Science (New York, N.Y.),* **320**(5880), pp. 1213-1216.

HUMINIECKI, L. and WOLFE, K.H., 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome research,* **14**(10A), pp. 1870-1879.

HUNT, S.E., MCLAREN, W., GIL, L., THORMANN, A., SCHUILENBURG, H., SHEPPARD, D., PARTON, A., ARMEAN, I.M., TREVANION, S.J., FLICEK, P. and CUNNINGHAM, F., 2018. Ensembl variation resources. *Database : the journal of biological databases and curation,* **2018**, pp. 10.1093/database/bay119.

IRIZARRY, R.A., WU, Z. and JAFFEE, H.A., 2006. Comparison of Affymetrix GeneChip expression measures. *Bioinformatics (Oxford, England),* **22**(7), pp. 789-794.

JAMOUS, A. and SALAH, Z., 2018. WW-Domain Containing Protein Roles in Breast Tumorigenesis. *Frontiers in oncology,* **8**, pp. 580.

JARINOVA, O., HATCH, G., POITRAS, L., PRUDHOMME, C., GRZYB, M., AUBIN, J., BERUBE-SIMARD, F.A., JEANNOTTE, L. and EKKER, M., 2008. Functional resolution of duplicated hoxb5 genes in teleosts. *Development (Cambridge, England),* **135**(21), pp. 3543-3553.

JOHNSON, B.R., 2010. Division of labor in honeybees: form, function, and proximate mechanisms. *Behavioral Ecology and Sociobiology,* **64**(3), pp. 305-316.

JOHNSON, D.A. and THOMAS, M.A., 2007. The monosaccharide transporter gene family in Arabidopsis and rice: a history of duplications, adaptive evolution, and functional divergence. *Molecular biology and evolution,* **24**(11), pp. 2412-2423.

KHAITOVICH, P., WEISS, G., LACHMANN, M., HELLMANN, I., ENARD, W., MUETZEL, B., WIRKNER, U., ANSORGE, W. and PAABO, S., 2004. A neutral model of transcriptome evolution. *PLoS biology,* **2**(5), pp. E132.

KIM, E., PARK, K.E., KIM, J.S., BAEK, D.C., LEE, J.W., LEE, S.R., KIM, M.S., KIM, S.H., KIM, C.S., KOO, D.B., KANG, H.S., RYOO, Z.Y. and CHANG, K.T., 2009. Importance of the porcine ADAM3 disintegrin domain in sperm-egg interaction. *The Journal of reproduction and development,* **55**(2), pp. 156-162.

KODAMA, Y., SHUMWAY, M., LEINONEN, R. and INTERNATIONAL NUCLEOTIDE SEQUENCE DATABASE COLLABORATION, 2012. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic acids research,* **40**(Database issue), pp. D54-6.

KOLESNIKOV, N., HASTINGS, E., KEAYS, M., MELNICHUK, O., TANG, Y.A., WILLIAMS, E., DYLAG, M., KURBATOVA, N., BRANDIZI, M., BURDETT, T., MEGY, K., PILICHEVA, E., RUSTICI, G., TIKHONOV, A., PARKINSON, H., PETRYSZAK, R., SARKANS, U. and BRAZMA, A., 2015. ArrayExpress update--simplifying data submissions. *Nucleic acids research,* **43**(Database issue), pp. D1113-6.

LATTORFF, H.M. and MORITZ, R.F., 2013. Genetic underpinnings of division of labor in the honeybee (Apis mellifera). *Trends in genetics : TIG,* **29**(11), pp. 641-648.

LOCKTON, S. and GAUT, B.S., 2005. Plant conserved non-coding sequences and paralogue evolution. *Trends in genetics : TIG,* **21**(1), pp. 60-65.

MACCARTHY, T. and BERGMAN, A., 2007. The limits of subfunctionalization. *BMC evolutionary biology,* **7**, pp. 213.

MACKINTOSH, C. and FERRIER, D.E.K., 2017. Recent advances in understanding the roles of whole genome duplications in evolution. *F1000Research,* **6**, pp. 1623.

MAIER, T., GUELL, M. and SERRANO, L., 2009. Correlation of mRNA and protein in complex biological samples. *FEBS letters,* **583**(24), pp. 3966-3973.

MCCLINTOCK, J.M., KHEIRBEK, M.A. and PRINCE, V.E., 2002. Knockdown of duplicated zebrafish hoxb1 genes reveals distinct roles in hindbrain patterning and a novel mechanism of duplicate gene retention. *Development (Cambridge, England),* **129**(10), pp. 2339-2354.

MEHDI, A.M., PATRICK, R., BAILEY, T.L. and BODEN, M., 2014. Predicting the dynamics of protein abundance. *Molecular & cellular proteomics : MCP,* **13**(5), pp. 1330-1340.

MERCADO-LUBO, R. and MCCORMICK, B.A., 2010. The interaction of gut microbes with host ABC transporters. *Gut microbes,* **1**(5), pp. 301-306.

NIE, L., WU, G. and ZHANG, W., 2006. Correlation of mRNA expression and protein abundance affected by multiple sequence features related to translational efficiency in Desulfovibrio vulgaris: a quantitative analysis. *Genetics,* **174**(4), pp. 2229-2243.

NIKOLSKAYA, A.N., ARIGHI, C.N., HUANG, H., BARKER, W.C. and WU, C.H., 2007. PIRSF family classification system for protein functional and evolutionary analysis. *Evolutionary bioinformatics online,* **2**, pp. 197-209.

NUZHDIN, S.V., WAYNE, M.L., HARMON, K.L. and MCINTYRE, L.M., 2004. Common pattern of evolution of gene expression level and protein sequence in Drosophila. *Molecular biology and evolution,* **21**(7), pp. 1308-1317.

OAKLEY, T.H., GU, Z., ABOUHEIF, E., PATEL, N.H. and LI, W.H., 2005. Comparative methods for the analysis of gene-expression evolution: an example using yeast functional genomic data. *Molecular biology and evolution,* **22**(1), pp. 40-50.

OHSHIMA, K., 2013. RNA-Mediated Gene Duplication and Retroposons: Retrogenes, LINEs, SINEs, and Sequence Specificity. *International journal of evolutionary biology,* **2013**, pp. 424726.

OKAMURA, Y., OBAYASHI, T. and KINOSHITA, K., 2015. Comparison of Gene Coexpression Profiles and Construction of Conserved Gene Networks to Find Functional Modules. *PloS one,* **10**(7), pp. e0132039.

PANCHIN, A.Y., GELFAND, M.S., RAMENSKY, V.E. and ARTAMONOVA, I.I., 2010. Asymmetric and non-uniform evolution of recently duplicated human genes. *Biology direct,* **5**, pp. 54.

PARADIS, E., CLAUDE, J. and STRIMMER, K., 2004a. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics (Oxford, England),* **20**(2), pp. 289-290.

PARADIS, E., CLAUDE, J. and STRIMMER, K., 2004b. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics (Oxford, England),* **20**(2), pp. 289-290.

POTTER, S.C., LUCIANI, A., EDDY, S.R., PARK, Y., LOPEZ, R. and FINN, R.D., 2018. HMMER web server: 2018 update. *Nucleic acids research,* **46**(W1), pp. W200-W204.

PRELIC, A., BLEULER, S., ZIMMERMANN, P., WILLE, A., BUHLMANN, P., GRUISSEM, W., HENNIG, L., THIELE, L. and ZITZLER, E., 2006. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics (Oxford, England),* **22**(9), pp. 1122-1129.

PUNTA, M., COGGILL, P.C., EBERHARDT, R.Y., MISTRY, J., TATE, J., BOURSNELL, C., PANG, N., FORSLUND, K., CERIC, G., CLEMENTS, J., HEGER, A., HOLM, L., SONNHAMMER, E.L., EDDY, S.R., BATEMAN, A. and FINN, R.D., 2012. The Pfam protein families database. *Nucleic acids research,* **40**(Database issue), pp. D290-301.

QIAN, W., LIAO, B.Y., CHANG, A.Y. and ZHANG, J., 2010. Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends in genetics : TIG,* **26**(10), pp. 425-430.

QIAO, X., LI, Q., YIN, H., QI, K., LI, L., WANG, R., ZHANG, S. and PATERSON, A.H., 2019. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome biology,* **20**(1), pp. 38-019-1650-2.

QUELLER, D.C. and STRASSMANN, J.E., 2003. Eusociality. *Current biology : CB,* **13**(22), pp. R861-3.

QUINONES, A.E. and PEN, I., 2017. A unified model of Hymenopteran preadaptations that trigger the evolutionary transition to eusociality. *Nature communications,* **8**, pp. 15920.

R DEVELOPMENT CORE TEAM, 2010. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria,* .

RAJASHEKAR, B., SAMSON, P., JOHANSSON, T. and TUNLID, A., 2007. Evolution of nucleotide sequences and expression patterns of hydrophobin genes in the ectomycorrhizal fungus Paxillus involutus. *The New phytologist,* **174**(2), pp. 399-411.

RANZ, J.M. and MACHADO, C.A., 2006. Uncovering evolutionary patterns of gene expression using microarrays. *Trends in ecology & evolution (Personal edition),* **21**(1), pp. 29-37.

REDON, R., ISHIKAWA, S., FITCH, K.R., FEUK, L., PERRY, G.H., ANDREWS, T.D., FIEGLER, H., SHAPERO, M.H., CARSON, A.R., CHEN, W., CHO, E.K., DALLAIRE, S., FREEMAN, J.L., GONZALEZ, J.R., GRATACOS, M., HUANG, J., KALAITZOPOULOS, D., KOMURA, D., MACDONALD, J.R., MARSHALL, C.R., MEI, R., MONTGOMERY, L., NISHIMURA, K., OKAMURA, K., SHEN, F., SOMERVILLE, M.J., TCHINDA, J., VALSESIA, A., WOODWARK, C., YANG, F., ZHANG, J., ZERJAL, T., ZHANG, J., ARMENGOL, L., CONRAD, D.F., ESTIVILL, X., TYLER-SMITH, C., CARTER, N.P., ABURATANI, H., LEE, C., JONES, K.W., SCHERER, S.W. and HURLES, M.E., 2006. Global variation in copy number in the human genome. *Nature,* **444**(7118), pp. 444-454.

REN, X.Y., FIERS, M.W., STIEKEMA, W.J. and NAP, J.P., 2005. Local coexpression domains of two to four genes in the genome of Arabidopsis. *Plant physiology,* **138**(2), pp. 923-934.

RETIEF, J.D., 2000. Phylogenetic analysis using PHYLIP. *Methods in molecular biology (Clifton, N.J.),* **132**, pp. 243-258.

RICHARDSON, S.R., SALVADOR-PALOMEQUE, C. and FAULKNER, G.J., 2014. Diversity through duplication: whole-genome sequencing reveals novel gene retrocopies in the human population. *BioEssays : news and reviews in molecular, cellular and developmental biology,* **36**(5), pp. 475-481.

ROGOZIN, I., I. WOLF, Y., BABENKO, V. and KOONIN, E., 2006. Dollo parsimony and the reconstruction of genome evolution. pp. 190-200.

RONQUIST, F. and HUELSENBECK, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics (Oxford, England),* **19**(12), pp. 1572-1574.

ROSSNES, R., EIDHAMMER, I. and LIBERLES, D.A., 2005. Phylogenetic reconstruction of ancestral character states for gene expression and mRNA splicing data. *BMC bioinformatics,* **6**, pp. 127.

RUIZ-ORERA, J., HERNANDEZ-RODRIGUEZ, J., CHIVA, C., SABIDO, E., KONDOVA, I., BONTROP, R., MARQUES-BONET, T. and ALBA, M.M., 2015. Origins of De Novo Genes in Human and Chimpanzee. *PLoS genetics,* **11**(12), pp. e1005721.

SAHOO, D., DILL, D.L., TIBSHIRANI, R. and PLEVRITIS, S.K., 2007. Extracting binary signals from microarray time-course data. *Nucleic acids research,* **35**(11), pp. 3705-3712.

SARIPELLA, G.V., SONNHAMMER, E.L. and FORSLUND, K., 2016. Benchmarking the next generation of homology inference tools. *Bioinformatics (Oxford, England),* **32**(17), pp. 2636-2641.

SATO, A., 2018. Chaperones, Canalization, and Evolution of Animal Forms. *International journal of molecular sciences,* **19**(10), pp. 10.3390/ijms19103029.

SCHACHERER, J., TOURRETTE, Y., SOUCIET, J.L., POTIER, S. and DE MONTIGNY, J., 2004. Recovery of a function involving gene duplication by retroposition in Saccharomyces cerevisiae. *Genome research,* **14**(7), pp. 1291-1297.

SCHOENFELDER, S. and FRASER, P., 2019. Long-range enhancer-promoter contacts in gene expression control. *Nature reviews.Genetics,* .

SCRUCCA, L., FOP, M., MURPHY, T.B. and RAFTERY, A.E., 2016. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R journal,* **8**(1), pp. 289-317.

SEMON, M. and WOLFE, K.H., 2008. Preferential subfunctionalization of slow-evolving genes after allopolyploidization in Xenopus laevis. *Proceedings of the National Academy of Sciences of the United States of America,* **105**(24), pp. 8333-8338.

SEO, J. and HOFFMAN, E.P., 2006. Probe set algorithms: is there a rational best bet? *BMC bioinformatics,* **7**, pp. 395-2105-7-395.

SHIMIZU, T., TETSUKA, M., MIYAMOTO, A. and UCHIDA, T., 2007. Follicle-stimulating hormone (FSH) stimulates the expression of Pin1, a peptidyl-prolyl isomerase, in the bovine granulosa cells. *Domestic animal endocrinology,* **32**(3), pp. 226-234.

SIEVERS, F. and HIGGINS, D.G., 2014. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods in molecular biology (Clifton, N.J.),* **1079**, pp. 105-116.

SJOLANDER, K., 2004. Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics (Oxford, England),* **20**(2), pp. 170-179.

SKAMNIOTI, P., FURLONG, R.F. and GURR, S.J., 2008. Evolutionary history of the ancient cutinase family in five filamentous Ascomycetes reveals differential gene duplications and losses and in Magnaporthe grisea shows evidence of sub- and neo-functionalization. *The New phytologist,* **180**(3), pp. 711-721.

SPRACKLIN, G., FIELDS, B., WAN, G., BECKER, D., WALLIG, A., SHUKLA, A. and KENNEDY, S., 2017. The RNAi Inheritance Machinery of Caenorhabditis elegans. *Genetics,* **206**(3), pp. 1403-1416.

STORZ, J.F., OPAZO, J.C. and HOFFMANN, F.G., 2011. Phylogenetic diversification of the globin gene superfamily in chordates. *IUBMB life,* **63**(5), pp. 313-322.

STROCCHI, P., BROWN, B.A., YOUNG, J.D., BONVENTRE, J.A. and GILBERT, J.M., 1981. The characterization of tubulin in CNS membrane fractions. *Journal of neurochemistry,* **37**(5), pp. 1295-1307.

SUCHARD, M.A. and REDELINGS, B.D., 2006. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics (Oxford, England),* **22**(16), pp. 2047-2048.

TANG, P. and VOGLER, A.P., 2017. Evolution: Taking the Sting out of Wasp Phylogenetics. *Current biology : CB,* **27**(9), pp. R358-R360.

TELLAM, R.L., LEMAY, D.G., VAN TASSELL, C.P., LEWIN, H.A., WORLEY, K.C. and ELSIK, C.G., 2009. Unlocking the bovine genome. *BMC genomics,* **10**, pp. 193-2164-10-193.

TIROSH, I. and BARKAI, N., 2007. Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome biology,* **8**(4), pp. R50.

TOUFIGHI, K., BRADY, S.M., AUSTIN, R., LY, E. and PROVART, N.J., 2005. The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses. *The Plant Journal : for cell and molecular biology,* **43**(1), pp. 153-163.

TRAPNELL, C., WILLIAMS, B.A., PERTEA, G., MORTAZAVI, A., KWAN, G., VAN BAREN, M.J., SALZBERG, S.L., WOLD, B.J. and PACHTER, L., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology,* **28**(5), pp. 511-515.

TSANKOV, A.M., THOMPSON, D.A., SOCHA, A., REGEV, A. and RANDO, O.J., 2010. The role of nucleosome positioning in the evolution of gene regulation. *PLoS biology,* **8**(7), pp. e1000414.

TURUNEN, O., SEELKE, R. and MACOSKO, J., 2009. In silico evidence for functional specialization after genome duplication in yeast. *FEMS yeast research,* **9**(1), pp. 16-31.

VAN DE PEER, Y., MAERE, S. and MEYER, A., 2009. The evolutionary significance of ancient genome duplications. *Nature reviews.Genetics,* **10**(10), pp. 725-732.

VIAENE, T., VEKEMANS, D., BECKER, A., MELZER, S. and GEUTEN, K., 2010. Expression divergence of the AGL6 MADS domain transcription factor lineage after a core eudicot duplication suggests functional diversification. *BMC plant biology,* **10**, pp. 148.

VISION, T.J., BROWN, D.G. and TANKSLEY, S.D., 2000. The origins of genomic duplications in Arabidopsis. *Science (New York, N.Y.),* **290**(5499), pp. 2114-2117.

VOGEL, C. and MARCOTTE, E.M., 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews.Genetics,* **13**(4), pp. 227-232.

WANG, D., SUNG, H.M., WANG, T.Y., HUANG, C.J., YANG, P., CHANG, T., WANG, Y.C., TSENG, D.L., WU, J.P., LEE, T.C., SHIH, M.C. and LI, W.H., 2007. Expression evolution in yeast genes of single-input modules is mainly due to changes in trans-acting factors. *Genome research,* .

WANG, M., ZHANG, X., ZHAO, H., WANG, Q. and PAN, Y., 2009. FoxO gene family evolution in vertebrates. *BMC evolutionary biology,* **9**, pp. 222-2148-9-222.

WANG, R., CHONG, K. and WANG, T., 2006. Divergence in spatial expression patterns and in response to stimuli of tandem-repeat paralogues encoding a novel class of proline-rich proteins in Oryza sativa. *Journal of experimental botany,* **57**(11), pp. 2887-2897.

WANG, Z., DONG, X., DING, G. and LI, Y., 2010. Comparing the retention mechanisms of tandem duplicates and retrogenes in human and mouse genomes. *Genetics, selection, evolution : GSE,* **42**, pp. 24.

WHITEHEAD, A. and CRAWFORD, D.L., 2006a. Neutral and adaptive variation in gene expression. *Proceedings of the National Academy of Sciences of the United States of America,* **103**(14), pp. 5425-5430.

WHITEHEAD, A. and CRAWFORD, D.L., 2006b. Variation within and among species in gene expression: raw material for evolution. *Molecular ecology,* **15**(5), pp. 1197-1211.

WOODY, O.Z., DOXEY, A.C. and MCCONKEY, B.J., 2008. Assessing the evolution of gene expression using microarray data. *Evolutionary bioinformatics online,* **4**, pp. 139-152.

XING, Y., OUYANG, Z., KAPUR, K., SCOTT, M.P. and WONG, W.H., 2007. Assessing the conservation of Mammalian gene expression using high-density exon arrays. *Molecular biology and evolution,* **24**(6), pp. 1283-1285.

YANG, J., SU, A.I. and LI, W.H., 2005. Gene expression evolves faster in narrowly than in broadly expressed mammalian genes. *Molecular biology and evolution,* **22**(10), pp. 2113-2118.

ZHAO, Q., FANG, F., SHAN, Y., SUI, Z., ZHAO, B., LIANG, Z., ZHANG, L. and ZHANG, Y., 2017. In-Depth Proteome Coverage by Improving Efficiency for Membrane Proteome Analysis. *Analytical Chemistry,* **89**(10), pp. 5179-5185.

Appendix A: Alternative identifiers for Apis mellifera

Table A.1: Alternative identifiers for yellow genes (Chapter 4)

Gene Identifier	Database Identifier	Alternative mRNA accession
mrjp-1	GB14888	NM_001011579.1
mrjp-2	GB16246	NM_001011580.1
mrjp-3	GB16459	NM_001011601.1
mrjp-4	GB11768	NM_001011610.1
mrjp-5	GB10622	NM_001011599.1
mrjp-6	GB13789	NM_001011622.1
mrjp-7	GB11022	NM_001014429.1
mrjp-8	GB14639	ACD84799.1
mrjp-9	GB16324	NM_001024697.1
yellow-y	GB19464	
yellow-e3	GB18089	ADW82101.1
yellow-h	GB18654	XM_006558929.2
yellow-e	GB17225	XM_003249378.3
yellow-x1	GB18300	XM_006564945.2
yellow-f	GB17489	
yellow-b	GB16705	
yellow-g	GB10842	XM_006558944.2
yellow-g2	GB18218	XM_006558943.2
yellow-x2	GB19132	

Appendix B: Summaries of *Sus scrofa* Analysis through BranchOut

Table B.1: Complete listing of tissue repesentation amongst most strongly scoring BranchOut reconstructions.

Tissue Sample	Representation in High-Scoring Reconstructions
Thymus.P3.F	29
Jejeunum.P3.F	26
Cortex_.prefrontal..P3.F	25
Jejeunum.P4.M	25
Blood_1	23
Hind_Brain_.medulla..P4.M	23
Liver.P3.F	23
Cerebellum.P4.M	22
Cortex_.prefrontal..P4.M	22
Placenta.F	22
Tongue_.dermal_layer..P4.M	22
Caecum..apical..P4.M	21
Colon_.distal..P3.F	21
Spinal_cord_.lower..P2.F	21
Testis_.adult..M	21
Colon_.distal..P4.M	20
Colon_.proximal..P3.F	20
Fallopian_tube.P3.F	19
Gall_bladder.P3.F	19
Kidney_.medulla..P4.M	19
Lung_Parenchyma.P3.F	19
Rectum.P4.M	19
Skeletal_muscle_.leg..P3.F	19
Tongue_.dermal_layer..P3.F	19
Blood_2	18
Duodenum.P4.M	18
Hind_Brain_.medulla..P3.F	18
Pancreas.P4.M	18
Rectum.P3.F	18
Spinal_cord_.lower..P3.F	18
Cerebellum.P3.F	17
Heart.P4.M	17
Oesophagus_.lower_third..P3.F	17
Oesophagus_.lower_third..P4.M	17
Penis.P4.M	17

Stomach_.antrum..P3.F	17
Bladder.P2.F	16
Gall_bladder.P4.M	16
Liver.P4.M	16
Stomach_.fundus..P3.F	16
Uterus.P3.F	16
Vas_deferens.P4.M.1	16
Bone_marrow.P4.M	15
Duodenum.P3.F	15
Mesenteric_lymph_node.P4.M	15
Alveolar_macrophage_2	14
Kidney_.cortex..P3.F	14
Optic_nerve.P3.F	14
Ovary.P3.F	14
Pancreas.P3.F	14
Retina.sclera.P3.F	14
Retina.sclera.P4.M	14
Spinal_cord_.upper..P3.F	14
Stomach_.fundus..P4.M	14
Trachea.P3.F	14
Trachea.P4.M	14
Abdominal_aorta.P4.M	13
Caecum..mid..P3.F	13
Kidney_.cortex..P4.M	13
Kidney_.medulla..P3.F	13
Lung_Parenchyma.P4.M	13
MD_macrophage_7h_LPS	13
Mesenteric_lymph_node.P3.F	13
Pylorus_.smooth_muscle..P3.F	13
Pylorus_.smooth_muscle..P4.M	13
Spleen.P4.M	13
Testis_.juvenile..P4.M	13
Thymus.P4.M	13
Thyroid.P4.M	13
Ureter.P4.M	13
Abdominal_aorta.P3.F	12
Adrenal_gland_.cortex..P4.M	12
Alveolar_macrophage_1	12
Caecum..mid..P4.M	12
Cornea.iris.P3.F	12
Ileum.P4.M	12

Optic_nerve.P4.M	12
Skin_.head..P4.M	12
Spleen.P3.F	12
Thyroid.P3.F	12
Vagina.P3.F	12
Adrenal_gland_.cortex..P3.F	11
BMD_macrophage_unstimulated	11
Bone_marrow.P3.F	11
Caecum..apical..P3.F	11
Cervix.P3.F	11
Ileum.P3.F	11
Inferior_vena_cava.P4.M	11
MD_macrophage_unstimulated	11
Spinal_cord_.upper..P2.F	11
BMD_macrophage_7h_LPS	10
Pituitary.P1.F	10
Salivary_glands_.submandibular..P3.F	10
Salivary_glands_.submandibular..P4.M	10
Bladder.P4.M	9
Oesophagus_.upper_third..P3.F	9
Oesophagus_.upper_third..P4.M	9
Epididymis.P4.M	8
Ureter.P3.F	8
Snout_tendon.P4.M	7
Colon_.proximal..P4.M	6
Stomach_.antrum..P4.M	6
Skeletal_muscle_.leg..P4.M	4
Vas_deferens.P4.M	4

Note: A reconstruction was considered “high-scoring” if the state-change ratio (comparing the randomized assignments to observed label assignments) was 1.5 or over.

Table B.2: Complete listing of gene family representation amongst most strongly scoring BranchOut reconstructions.

Protein Family Identifier and Description	Representation in High-Scoring Reconstructions
PF02736: Myosin.N.terminal.SH3.like.domain	34
PF00386: C1q.domain	32
PF00026: Eukaryotic.aspartyl.protease	26
PF00244: X14.3.3.protein	25
PF06512: Sodium.ion.transport.associated	21
PF00175: Oxidoreductase.NAD.binding.domain	19
PF00503: G.protein.alpha.subunit	18
PF00030: Beta.Gamma.crystallin	17
PF00160: Cyclophilin.type.peptidyl.prolyl.cis.trans.isomerase.CLD	17
PF00188: Cysteine.rich.secretory.protein.family	17
PF00010: Helix.loop.helix.DNA.binding.domain	15
PF01582: TIR.domain	15
PF02463: RecF.RecN.SMC.N.terminal.domain	15
PF00040: Fibronectin.type.II.domain	14
PF00191: Annexin	14
PF01569: PAP2.superfamily	14
PF01534: Frizzled.Smoothened.family.membrane.region	13
PF00022: Actin	12
PF00055: Laminin.N.terminal..Domain.VI.	12
PF00086: Thyroglobulin.type.1.repeat	12
PF00629: MAM.domain	12
PF00735: Septin	12
PF01266: FAD.dependent.oxidoreductase	12
PF01436: NHL.repeat	12
PF01462: Leucine.rich.repeat.N.terminal.domain	12
PF01576: Myosin.tail	12
PF03062: MBOAT.family	12
PF03114: BAR.domain	12
PF00474: Sodium.solute.symporter.family	11
PF00533: BRCA1.C.Terminus..BRCT..domain	11
PF01421: Reprolysin..M12B..family.zinc.metalloprotease	11
PF02815: MIR.domain	11
PF06747: CHCH.domain	11
PF00038: Intermediate.filament.protein	10

PF00249: Myb.like.DNA.binding.domain	10
PF00270: DEAD.DEAH.box.helicase	10
PF00324: Amino.acid.permease	10
PF00335: Tetraspanin.family	10
PF00888: Cullin.family	10
PF01823: MAC.Perforin.domain	10
PF00003: X7.transmembrane.sweet.taste.receptor.of.3.GCPR	9
PF00079: Serpin..serine.protease.inhibitor.	9
PF00149: Calcineurin.like.phosphoesterase	9
PF00225: Kinesin.motor.domain	9
PF00357: Integrin.alpha.cytoplasmic.region	9
PF00481: Protein.phosphatase.2C	9
PF00688: TGF.beta.propeptide	9
PF01369: Sec7.domain	9
PF01694: Rhomboid.family	9
PF02798: Glutathione.S.transferase..N.terminal.domain	9
PF03133: Tubulin.tyrosine.ligase.family	9
PF03143: Elongation.factor.Tu.C.terminal.domain	9
PF04851: Type.III.restriction.enzyme..res.subunit	9
PF07546: EMI.domain	9
PF00300: Phosphoglycerate.mutase.family	8
PF00626: Gelsolin.repeat	8
PF00690: Cation.transporter.ATPase..N.terminus	8
PF01979: Amidohydrolase.family	8
PF08242: Methyltransferase.domain	8
PF00246: Zinc.carboxypeptidase	7
PF00313: X.Cold.shock..DNA.binding.domain	7
PF00612: IQ.calmodulin.binding.motif	7
PF00644: Poly.ADP.ribose..polymerase.catalytic.domain	7
PF00702: haloacid.dehalogenase.like.hydrolase	7
PF01094: Receptor.family.ligand.binding.region	7
PF01388: ARID.BRIGHT.DNA.binding.domain	7
PF02932: Neurotransmitter.gated.ion.channel.transmembrane.region	7
PF07562: Nine.Cysteines.Domain.of.family.3.GPCR	7
PF07992: Pyridine.nucleotide.disulphide.oxidoreductase	7
PF09279: Phosphoinositide.specific.phospholipase.C..efhand.like	7
PF00029: Connexin	6
PF00092: von.Willebrand.factor.type.A.domain	6
PF00104: Ligand.binding.domain.of.nuclear.hormone.receptor	6
PF00105: Zinc.finger..C4.type..two.domains.	6

PF00147: Fibrinogen.beta.and.gamma.chains..C.terminal.globular.domain	6
PF00151: Lipase	6
PF00155: Aminotransferase.class.I.and.II	6
PF00167: Fibroblast.growth.factor	6
PF00200: Disintegrin	6
PF00271: Helicase.conserved.C.terminal.domain	6
PF00530: Scavenger.receptor.cysteine.rich.domain	6
PF00569: Zinc.finger..ZZ.type	6
PF00648: Calpain.family.cysteine.protease	6
PF00689: Cation.transporting.ATPase..C.terminus	6
PF00899: ThiF.family	6
PF01023: S.100.ICaBP.type.calcium.binding.domain	6
PF01399: PCI.domain	6
PF01535: PPR.repeat	6
PF01562: Reprolysin.family.propeptide	6
PF01759: UNC.6.NTR.C345C.module	6
PF02493: MORN.repeat	6
PF03826: OAR.domain	6
PF07525: SOCS.box	6
PF07717: Domain.of.unknown.function..DUF1605.	6
PF00005: ABC.transporter	5
PF00031: Cystatin.domain	5
PF00035: Double.stranded.RNA.binding.motif	5
PF00043: Glutathione.S.transferase..C.terminal.domain	5
PF00083: Sugar..and.other..transporter	5
PF00129: Class.I.Histocompatibility.antigen..domains.alpha.1.and.2	5
PF00156: Phosphoribosyl.transferase.domain	5
PF00171: Aldehyde.dehydrogenase.family	5
PF00209: Sodium.neurotransmitter.symporter.family	5
PF00211: Adenylate.and.Guanylate.cyclase.catalytic.domain	5
PF00230: Major.intrinsic.protein	5
PF00250: Fork.head.domain	5
PF00388: Phosphatidylinositol.specific.phospholipase.C..X.domain	5
PF00454: Phosphatidylinositol.3..and.4.kinase	5
PF00777: Glycosyltransferase.family.29..sialyltransferase.	5
PF01390: SEA.domain	5
PF01433: Peptidase.family.M1	5
PF01490: Transmembrane.amino.acid.transporter.protein	5

PF01663: Type.I.phosphodiesterase...nucleotide.pyrophosphatase	5
PF02191: Olfactomedin.like.domain	5
PF02214: K..channel.tetramerisation.domain	5
PF02338: OTU.like.cysteine.protease	5
PF02883: Adaptin.C.terminal.domain	5
PF02984: Cyclin..C.terminal.domain	5
PF03151: Triose.phosphate.Transporter.family	5
PF03372: Endonuclease.Exonuclease.phosphatase.family	5
PF03765: CRAL.TRIO..N.terminus	5
PF03810: Importin.beta.N.terminal.domain	5
PF03953: Tubulin.C.terminal.domain	5
PF04408: Helicase.associated.domain..HA2.	5
PF05739: SNARE.domain	5
PF07690: Major.Facilitator.Superfamily	5
PF08516: ADAM.cysteine.rich	5
PF00017: SH2.domain	4
PF00021: u.PAR.Ly.6.domain	4
PF00050: Kazal.type.serine.protease.inhibitor.domain	4
PF00060: Ligand.gated.ion.channel	4
PF00070: Pyridine.nucleotide.disulphide.oxidoreductase	4
PF00091: Tubulin.FtsZ.family..GTPase.domain	4
PF00100: Zona.pellucida.like.domain	4
PF00106: short.chain.dehydrogenase	4
PF00107: Zinc.binding.dehydrogenase	4
PF00110: wnt.family	4
PF00118: TCP.1.cpn60.chaperonin.family	4
PF00254: FKBP.type.peptidyl.prolyl.cis.trans.isomerase	4
PF00293: NUDIX.domain	4
PF00413: Matrixin	4
PF00536: SAM.domain..Sterile.alpha.motif.	4
PF00581: Rhodanese.like.domain	4
PF00616: GTPase.activator.protein.for.Ras.like.GTPase	4
PF00619: Caspase.recruitment.domain	4
PF00625: Guanylate.kinase	4
PF00643: B.box.zinc.finger	4
PF00650: CRAL.TRIO.domain	4
PF00656: Caspase.domain	4
PF00754: F5.8.type.C.domain	4
PF00788: Ras.association..RalGDS.AF.6..domain	4
PF00856: SET.domain	4

PF00884: Sulfatase	4
PF01105: emp24.gp25L.p24.family.GOLD	4
PF01217: Clathrin.adaptor.complex.small.chain	4
PF01302: CAP.Gly.domain	4
PF01336: OB.fold.nucleic.acid.binding.domain	4
PF01370: NAD.dependent.epimerase.dehydratase.family	4
PF01404: Ephrin.receptor.ligand.binding.domain	4
PF01839: FG.GAP.repeat	4
PF02037: SAP.domain	4
PF02178: AT.hook.motif	4
PF04969: CS.domain	4
PF08205: CD80.like.C2.set.immunoglobulin.domain	4
PF08686: PLAC..protease.and.lacunin..domain	4
PF00012: Hsp70.protein	3
PF00020: TNFR.NGFR.cysteine.rich.region	3
PF00025: ADP.ribosylation.factor.family	3
PF00028: Cadherin.domain	3
PF00059: Lectin.C.type.domain	3
PF00076: RNA.recognition.motif...a.k.a..RRM..RBD..or.RNP.domain.	3
PF00085: Thioredoxin	3
PF00112: Papain.family.cysteine.protease	3
PF00178: Ets.domain	3
PF00194: Eukaryotic.type.carbonic.anhydrase	3
PF00219: Insulin.like.growth.factor.binding.protein	3
PF00373: FERM.central.domain	3
PF00431: CUB.domain	3
PF00498: FHA.domain	3
PF00505: HMG..high.mobility.group..box	3
PF00515: Tetratricopeptide.repeat	3
PF00531: Death.domain	3
PF00566: TBC.domain	3
PF00617: RasGEF.domain	3
PF00664: ABC.transporter.transmembrane.region	3
PF00795: Carbon.nitrogen.hydrolase	3
PF00999: Sodium.hydrogen.exchanger.family	3
PF01007: Inward.rectifier.potassium.channel	3
PF01145: SPFH.domain...Band.7.family	3
PF01392: Fz.domain	3
PF01412: Putative.GTPase.activating.protein.for.Arf	3
PF01437: Plexin.repeat	3

PF01463: Leucine.rich.repeat.C.terminal.domain	3
PF01602: Adaptin.N.terminal.region	3
PF01753: MYND.finger	3
PF01794: Ferric.reductase.like.transmembrane.component	3
PF01825: Latrophilin.CL.1.like.GPS.domain	3
PF02205: WH2.motif	3
PF02210: Laminin.G.domain	3
PF02225: PA.domain	3
PF02820: mbt.repeat	3
PF03171: X2OG.Fe.II..oxygenase.superfamily	3
PF03456: uDENN.domain	3
PF05729: NACHT.domain	3
PF07647: SAM.domain..Sterile.alpha.motif.	3
PF08240: Alcohol.dehydrogenase.GroES.like.domain	3
PF08736: FERM.adjacent..FA.	3
PF00004: ATPase.family.associated.with.various.cellular.activities..AAA.	2
PF00009: Elongation.factor.Tu.GTP.binding.domain	2
PF00019: Transforming.growth.factor.beta.like.domain	2
PF00027: Cyclic.nucleotide.binding.domain	2
PF00045: Hemopexin	2
PF00051: Kringle.domain	2
PF00057: Low.density.lipoprotein.receptor.domain.class.A	2
PF00058: Low.density.lipoprotein.receptor.repeat.class.B	2
PF00067: Cytochrome.P450	2
PF00090: Thrombospondin.type.1.domain	2
PF00093: von.Willebrand.factor.type.C.domain	2
PF00095: WAP.type..Whey.Acidic.Protein...four.disulfide.core.	2
PF00125: Core.histone.H2A.H2B.H3.H4	2
PF00153: Mitochondrial.carrier.protein	2
PF00173: Cytochrome.b5.like.Heme.Steroid.binding.domain	2
PF00233: X3.5..cyclic.nucleotide.phosphodiesterase	2
PF00397: WW.domain	2
PF00433: Protein.kinase.C.terminal.domain	2
PF00439: Bromodomain	2
PF00452: Apoptosis.regulator.proteins..Bcl.2.family	2
PF00501: AMP.binding.enzyme	2
PF00567: Tudor.domain	2
PF00610: Domain.found.in.Dishevelled..Egl.10..and.Pleckstrin..DEP.	2
PF00611: Fes.CIP4.homology.domain	2

PF00618: motif	2
PF00632: HECT.domain..ubiquitin.transferase.	2
PF00641: Zn.finger.in.Ran.binding.protein.and.others	2
PF00652: Ricin.type.beta.trefoil.lectin.domain	2
PF00989: PAS.fold	2
PF01049: Cadherin.cytoplasmic.region	2
PF01344: Kelch.motif	2
PF01477: PLAT.LH2.domain	2
PF01485: IBR.domain	2
PF01585: G.patch.domain	2
PF01833: IPT.TIG.domain	2
PF02759: RUN.domain	2
PF02770: Acyl.CoA.dehydrogenase..middle.domain	2
PF02809: Ubiquitin.interaction.motif	2
PF02828: L27.domain	2
PF02931: Neurotransmitter.gated.ion.channel.ligand.binding.domain	2
PF04089: BRICHOS.domain	2
PF05986: ADAM.TS.Spacer.1	2
PF07648: Kazal.type.serine.protease.inhibitor.domain	2
PF07885: Ion.channel	2
PF07974: EGF.like.domain	2
PF08441: Integrin.alpha	2
PF08447: PAS.fold	2
PF00002: X7.transmembrane.receptor..Secretin.family.	1
PF00014: Kunitz.Bovine.pancreatic.trypsin.inhibitor.domain	1
PF00048: Small.cytokines..intecrine.chemokine...interleukin.8.like	1
PF00053: Laminin.EGF.like..Domains.III.and.V.	1
PF00098: Zinc.knuckle	1
PF00102: Protein.tyrosine.phosphatase	1
PF00122: E1.E2.ATPase	1
PF00134: Cyclin..N.terminal.domain	1
PF00168: C2.domain	1
PF00170: bZIP.transcription.factor	1
PF00179: Ubiquitin.conjugating.enzyme	1
PF00226: DnaJ.domain	1
PF00387: Phosphatidylinositol.specific.phospholipase.C..Y.domain	1
PF00415: Regulator.of.chromosome.condensation..RCC1..repeat	1
PF00443: Ubiquitin.carboxyl.terminal.hydrolase	1
PF00514: Armadillo.beta.catenin.like.repeat	1

PF00520: Ion.transport.protein	1
PF00615: Regulator.of.G.protein.signaling.domain	1
PF00621: RhoGEF.domain	1
PF00627: UBA.TS.N.domain	1
PF00640: Phosphotyrosine.interaction.domain..PTB.PID.	1
PF00646: F.box.domain	1
PF00651: BTB.POZ.domain	1
PF00685: Sulfotransferase.domain	1
PF00787: PX.domain	1
PF00822: PMP.22.EMP.MP20.Claudin.family	1
PF01064: Activin.types.I.and.II.receptor.domain	1
PF01403: Sema.domain	1
PF01471: Putative.peptidoglycan.binding.domain	1
PF01553: Acyltransferase	1
PF01762: Galactosyltransferase	1
PF01926: GTPase.of.unknown.function	1
PF02141: DENN..AEX.3..domain	1
PF02373: JmjC.domain	1
PF03144: Elongation.factor.Tu.domain.2	1
PF03455: dDENN.domain	1
PF06602: Myotubularin.related	1
PF07719: Tetratricopeptide.repeat	1
PF08028: Acyl.CoA.dehydrogenase..C.terminal.domain	1
PF08659: KR.domain	1

Note: A reconstruction was considered “high-scoring” if the state-change ratio (comparing the randomized assignments to observed label assignments) was 1.5 or over.

Appendix C: Summaries of *Bos taurus* Analysis through BranchOut

Table C.1: Complete listing of tissue representation amongst most strongly scoring BranchOut reconstructions.

Tissue Sample	Representation in High-Scoring Reconstructions
ampula (contralateral to CL)	15
Atrium	14
Kidney Medulla	14
uterine endometrium - caruncular (contralateral to CL)	14
Ascending colon	13
uterine endometrium - intercaruncular (contralateral to CL)	13
ampula (ipsilateral to CL)	12
Super bull Testis	12
Anterior Pituitary	11
Biceps femoris (bottom/outside round)	11
Cerebellum	11
Cerebral cortex	11
Descending Colon	11
Duodenum	11
follicle 4	11
isthmus (ipsilateral to CL)	11
Midbrain	11
Pineal Gland	11
Thalamus	11
Trachea	11
vas deferens	11
Bone Marrow	10
caput epididymis	10
Internal Tongue Muscle	10
SME	10
Tongue Superficial	10
Ventricle	10
bladder	9
Caecum	9
cervical lining	9
Fornix vagina	9
Gall Bladder	9
Ileum	9

infundibulum (ipsilateral to CL)	9
isthmus (contralateral to CL)	9
Longissimus dorsi (ribeye/loin)	9
Lymph Nodes	9
mesenteric lymph node	9
prostrate	9
Urethra	9
uterine endometrium - intercaruncular (ipsilateral to CL)	9
Adrenal	8
Aorta	8
Diencephalon	8
Frontal Cortex	8
Jejunum	8
KPH fat	8
Liver	8
Pigment Epithelium eye	8
Pons	8
Salivary Gland	8
Temporal Cortex	8
Ant. Eye	7
Hippocampus	7
Infraspinatus (top blade or flat iron from shoulder)	7
Infundibulum (contralateral to CL)	7
Kidney Cortex	7
Posterior Pituitary	7
Rectus femoris (center of the knuckle/sirloin tip)	7
Spleen	7
Sub-cutaneous Fat	7
Thyroid	7
Choroid plexus	6
Diaphragm	6
follicle 2	6
mammary gland fat	6
Nasal Mucosa	6
Omasum	6
Reticulum	6
Rumen	6
Rumen Papillae	6
Triceps brachii (shoulder clod)	6
uterine myometrium	6

white blood cells (wBC)	6
Abomasum	5
corpus epididymis	5
Esophagus	5
follicle 1	5
left Lung	5
Mammary gland	5
normal outer eye layer 1 cm from cancer eye	5
Supraspinatus (mock tender from shoulder)	5
255d lactating mammary gland	4
Corpus Luteum (if present, estimate d of cycle)	4
Gluteus medius (top sirloin)	4
Larynx Cartilage	4
Ureter	4
Cancer Eye	3
Semimembranosus (top/inside round)	3
Semitendinosus (eye of round)	3
lower tongue	2

Note: A reconstruction was considered “high-scoring” if the state-change ratio (comparing the randomized assignments to observed label assignments) was 1.5 or over.

Table C.2: Complete listing of gene family representation amongst most strongly scoring BranchOut reconstructions.

Protein Family Identifier and Description	Representation in High-Scoring Reconstructions
PF00134: Cyclin..N.terminal.domain	21
PF00102: Protein.tyrosine.phosphatase	20
PF00063: Myosin.head..motor.domain.	17
PF07654: Immunoglobulin.C1.set.domain	16
PF00091: Tubulin.FtsZ.family..GTPase.domain	14
PF00104: Ligand.binding.domain.of.nuclear.hormone.receptor	14
PF00481: Protein.phosphatase.2C	14
PF00149: Calcineurin.like.phosphoesterase	13
PF00153: Mitochondrial.carrier.protein	13
PF00789: UBX.domain	13
PF01553: Acyltransferase	13
PF12937: F.box.like	13
PF00013: KH.domain	12
PF00098: Zinc.knuckle	12
PF00105: Zinc.finger..C4.type..two.domains.	12
PF00373: FERM.central.domain	12
PF00615: Regulator.of.G.protein.signaling.domain	12
PF01363: FYVE.zinc.finger	12
PF13516: Leucine.Rich.repeat	11
PF13923: Zinc.finger..C3HC4.type..RING.finger.	11
PF00097: Zinc.finger..C3HC4.type..RING.finger.	10
PF00226: DnaJ.domain	10
PF05773: RWD.domain	10
PF07645: Calcium.binding.EGF.domain	10
PF00025: ADP.ribosylation.factor.family	9
PF00625: Guanylate.kinase	9
PF00685: Sulfotransferase.domain	9
PF00808: Histone.like.transcription.factor..CBF.NF.Y..and.archaeal.histone	9
PF00928: Adaptor.complexes.medium.subunit.family	9
PF03953: Tubulin.C.terminal.domain	9
PF13637: Ankyrin.repeats..many.copies.	9
PF00632: HECT.domain..ubiquitin.transferase.	8
PF07525: SOCS.box	8
PF00092: von.Willebrand.factor.type.A.domain	7
PF00125: Core.histone.H2A.H2B.H3.H4	7
PF00178: Ets.domain	7

PF00617: RasGEF.domain	7
PF00651: BTB.POZ.domain	7
PF01436: NHL.repeat	7
PF07719: Tetratricopeptide.repeat	7
PF13414: TPR.repeat	7
PF13833: EF.hand.domain.pair	7
PF00001: X7.transmembrane.receptor..rhodopsin.family.	6
PF00004: ATPase.family.associated.with.various.cellular.activities..AAA.	6
PF00046: Homeobox.domain	6
PF00397: WW.domain	6
PF00782: Dual.specificity.phosphatase..catalytic.domain	6
PF00787: PX.domain	6
PF04212: MIT..microtubule.interacting.and.transport..domain	6
PF05347: Complex.1.protein..LYR.family.	6
PF05739: SNARE.domain	6
PF07717: Oligonucleotide.oligosaccharide.binding..OB..fold	6
PF13857: Ankyrin.repeats..many.copies.	6
PF00009: Elongation.factor.Tu.GTP.binding.domain	5
PF00085: Thioredoxin	5
PF00169: PH.domain	5
PF00777: Glycosyltransferase.family.29..sialyltransferase.	5
PF01284: Membrane.associating.domain	5
PF02214: BTB.POZ.domain	5
PF04408: Helicase.associated.domain..HA2.	5
PF07686: Immunoglobulin.V.set.domain	5
PF12799: Leucine.Rich.repeats..2.copies.	5
PF13405: EF.hand.domain	5
PF13424: Tetratricopeptide.repeat	5
PF00017: SH2.domain	4
PF00076: RNA.recognition.motif...a.k.a..RRM..RBD..or.RNP.domain.	4
PF00155: Aminotransferase.class.I.and.II	4
PF00179: Ubiquitin.conjugating.enzyme	4
PF00241: Cofilin.tropomyosin.type.actin.binding.protein	4
PF00433: Protein.kinase.C.terminal.domain	4
PF00501: AMP.binding.enzyme	4
PF01529: DHHC.palmitoyltransferase	4
PF01585: G.patch.domain	4
PF03372: Endonuclease.Exonuclease.phosphatase.family	4
PF07707: BTB.And.C.terminal.Kelch	4

PF13202: EF.hand	4
PF13499: EF.hand.domain.pair	4
PF13893: RNA.recognition.motif...a.k.a..RRM..RBD..or.RNP.domain.	4
PF13894: C2H2.type.zinc.finger	4
PF00010: Helix.loop.helix.DNA.binding.domain	3
PF00160: Cyclophilin.type.peptidyl.prolyl.cis.trans.isomerase.CLD	3
PF00168: C2.domain	3
PF00170: bZIP.transcription.factor	3
PF00240: Ubiquitin.family	3
PF00270: DEAD.DEAH.box.helicase	3
PF00612: IQ.calmodulin.binding.motif	3
PF00899: ThiF.family	3
PF01485: IBR.domain	3
PF03144: Elongation.factor.Tu.domain.2	3
PF04969: CS.domain	3
PF07653: Variant.SH3.domain	3
PF08205: CD80.like.C2.set.immunoglobulin.domain	3
PF13895: Immunoglobulin.domain	3
PF13920: Zinc.finger..C3HC4.type..RING.finger.	3
PF00005: ABC.transporter	2
PF00018: SH3.domain	2
PF00293: NUDIX.domain	2
PF00335: Tetraspanin.family	2
PF00595: PDZ.domain..Also.known.as.DHR.or.GLGF.	2
PF00622: SPRY.domain	2
PF00643: B.box.zinc.finger	2
PF00753: Metallo.beta.lactamase.superfamily	2
PF01217: Clathrin.adaptor.complex.small.chain	2
PF01344: Kelch.motif	2
PF01423: LSM.domain	2
PF01926: X50S.ribosome.binding.GTPase	2
PF03357: Snf7	2
PF12697: Alpha.beta.hydrolase.family	2
PF13639: Ring.finger.domain	2
PF13849:	2
PF13855: Leucine.rich.repeat	2
PF00008: EGF.like.domain	1
PF00023: Ankyrin.repeat	1
PF00071: Ras.family	1
PF00106: short.chain.dehydrogenase	1

PF00130: Phorbol.esters.diacylglycerol.binding.domain..C1.domain.	1
PF00307: Calponin.homology..CH..domain	1
PF00443: Ubiquitin.carboxyl.terminal.hydrolase	1
PF00581: Rhodanese.like.domain	1
PF00621: RhoGEF.domain	1
PF00627: UBA.TS.N.domain	1
PF00641: Zn.finger.in.Ran.binding.protein.and.others	1
PF00856: SET.domain	1
PF01391: Collagen.triple.helix.repeat..20.copies.	1
PF01399: PCI.domain	1
PF01437: Plexin.repeat	1
PF02023: SCAN.domain	1
PF02535: ZIP.Zinc.transporter	1
PF03810: Importin.beta.N.terminal.domain	1
PF07690: Major.Facilitator.Superfamily	1
PF14259: RNA.recognition.motif..a.k.a..RRM..RBD..or.RNP.domain.	1

Note: A reconstruction was considered "high-scoring" if the state-change ratio (comparing the randomized assignments to observed label assignments) was 1.5 or over.

Appendix D: Sus scrofa Families Included in Study of Robustness

Table D.1: Families included in comparison of preprocessing algorithms

Protein Family Identifier	Protein Family Description	Number of Genes
PF00003	X7.transmembrane.sweet.taste.receptor.of.3.GCPR	[12]
PF00009	Elongation.factor.Tu.GTP.binding.domain	[15]
PF00012	Hsp70.protein	[10]
PF00013	KH.domain	[16]
PF00019	Transforming.growth.factor.beta.like.domain	[14]
PF00020	TNFR.NGFR.cysteine.rich.region	[14]
PF00021	u.PAR.Ly.6.domain	[7]
PF00022	Actin	[18]
PF00025	ADP.ribosylation.factor.family	[16]
PF00026	Eukaryotic.aspartyl.protease	[7]
PF00027	Cyclic.nucleotide.binding.domain	[18]
PF00029	Connexin	[10]
PF00030	Beta.Gamma.crystallin	[7]
PF00031	Cystatin.domain	[10]
PF00035	Double.stranded.RNA.binding.motif	[13]

PF00043	Glutathione.S.transferase..C.terminal.domain	[10]
PF00045	Hemopexin	[13]
PF00048	Small.cytokines..intecrine.chemokine...interleukin.8.like	[11]
PF00050	Kazal.type.serine.protease.inhibitor.domain	[19]
PF00051	Kringle.domain	[12]
PF00055	Laminin.N.terminal..Domain.VI.	[9]
PF00058	Low.density.lipoprotein.receptor.repeat.class.B	[9]
PF00060	Ligand.gated.ion.channel	[11]
PF00061	Lipocalin...cytosolic.fatty.acid.binding.protein.family	[23]
PF00067	Cytochrome.P450	[37]
PF00070	Pyridine.nucleotide.disulphide.oxidoreductase	[8]
PF00079	Serpin..serine.protease.inhibitor.	[25]
PF00083	Sugar..and.other..transporter	[19]
PF00085	Thioredoxin	[16]
PF00086	Thyroglobulin.type.1.repeat	[12]
PF00091	Tubulin.FtsZ.family..GTPase.domain	[16]
PF00093	von.Willebrand.factor.type.C.domain	[13]
PF00095	WAP.type..Whey.Acidic.Protein...four.disulfide.core.	[9]
PF00098	Zinc.knuckle	[11]
PF00100	Zona.pellucida.like.domain	[9]
PF00105	Zinc.finger..C4.type..two.domains.	[23]
PF00106	short.chain.dehydrogenase	[36]
PF00107	Zinc.binding.dehydrogenase	[9]
PF00110	wnt.family	[10]
PF00112	Papain.family.cysteine.protease	[8]
PF00118	TCP.1.cpn60.chaperonin.family	[12]
PF00122	E1.E2.ATPase	[13]
PF00125	Core.histone.H2A.H2B.H3.H4	[15]
PF00129	Class.I.Histocompatibility.antigen..domains.alpha.1.and.2	[10]
PF00134	Cyclin..N.terminal.domain	[16]
PF00147	Fibrinogen.beta.and.gamma.chains..C.terminal.globular.domain	[17]
PF00149	Calcineurin.like.phosphoesterase	[14]
PF00151	Lipase	[9]
PF00153	Mitochondrial.carrier.protein	[32]
PF00155	Aminotransferase.class.I.and.II	[13]
PF00156	Phosphoribosyl.transferase.domain	[8]
PF00167	Fibroblast.growth.factor	[12]
PF00170	bZIP.transcription.factor	[15]
PF00171	Aldehyde.dehydrogenase.family	[10]
PF00173	Cytochrome.b5.like.Heme.Steroid.binding.domain	[10]

PF00175	Oxidoreductase.NAD.binding.domain	[9]
PF00178	Ets.domain	[12]
PF00188	Cysteine.rich.secretory.protein.family	[10]
PF00191	Annexin	[8]
PF00194	Eukaryotic.type.carbonic.anhydrase	[11]
PF00200	Disintegrin	[24]
PF00209	Sodium.neurotransmitter.symporter.family	[11]
PF00211	Adenylate.and.Guanylate.cyclase.catalytic.domain	[14]
PF00219	Insulin.like.growth.factor.binding.protein	[8]
PF00227	Proteasome.subunit	[13]
PF00230	Major.intrinsic.protein	[7]
PF00233	X3.5..cyclic.nucleotide.phosphodiesterase	[11]
PF00240	Ubiquitin.family	[19]
PF00244	X14.3.3.protein	[7]
PF00246	Zinc.carboxypeptidase	[16]
PF00250	Fork.head.domain	[24]
PF00254	FKBP.type.peptidyl.prolyl.cis.trans.isomerase	[9]
PF00293	NUDIX.domain	[10]
PF00300	Histidine.phosphatase.superfamily..branch.1.	[7]
PF00313	X.Cold.shock..DNA.binding.domain	[9]
PF00324	Amino.acid.permease	[17]
PF00335	Tetraspanin.family	[15]
PF00357	Integrin.alpha.cytoplasmic.region	[9]
PF00378	Enoyl.CoA.hydratase.isomerase.family	[8]
PF00386	C1q.domain	[18]
PF00387	Phosphatidylinositol.specific.phospholipase.C..Y.domain	[10]
PF00388	Phosphatidylinositol.specific.phospholipase.C..X.domain	[10]
PF00413	Matrixin	[15]
PF00441	Acyl.CoA.dehydrogenase..C.terminal.domain	[10]
PF00452	Apoptosis.regulator.proteins..Bcl.2.family	[10]
PF00474	Sodium.solute.symporter.family	[8]
PF00481	Protein.phosphatase.2C	[10]
PF00498	FHA.domain	[15]
PF00501	AMP.binding.enzyme	[15]
PF00503	G.protein.alpha.subunit	[10]
PF00514	Armadillo.beta.catenin.like.repeat	[16]
PF00530	Scavenger.receptor.cysteine.rich.domain	[13]
PF00531	Death.domain	[14]
PF00533	BRCA1.C.Terminus..BRCT..domain	[11]
PF00535	Glycosyl.transferase.family.2	[15]
PF00561	alpha.beta.hydrolase.fold	[18]

PF00566	Rab.GTPase.TBC.domain	[21]
PF00567	Tudor.domain	[8]
PF00569	Zinc.finger..ZZ.type	[9]
PF00571	CBS.domain	[12]
PF00581	Rhodanese.like.domain	[9]
PF00583	Acetyltransferase..GNAT..family	[13]
PF00610	Domain.found.in.Dishevelled..Egl.10..and.Pleckstrin..DEP.	[10]
PF00611	Fes.CIP4..and.EFC.F.BAR.homology.domain	[8]
PF00615	Regulator.of.G.protein.signaling.domain	[19]
PF00616	GTPase.activator.protein.for.Ras.like.GTPase	[9]
PF00618	RasGEF.N.terminal.motif	[15]
PF00619	Caspase.recruitment.domain	[7]
PF00625	Guanylate.kinase	[13]
PF00626	Gelsolin.repeat	[9]
PF00629	MAM.domain	[9]
PF00640	Phosphotyrosine.interaction.domain..PTB.PID.	[13]
PF00644	Poly.ADP.ribose..polymerase.catalytic.domain	[10]
PF00648	Calpain.family.cysteine.protease	[7]
PF00650	CRAL.TRIO.domain	[10]
PF00652	Ricin.type.beta.trefoil.lectin.domain	[15]
PF00656	Caspase.domain	[9]
PF00685	Sulfotransferase.domain	[9]
PF00688	TGF.beta.propeptide	[9]
PF00689	Cation.transporting.ATPase..C.terminus	[11]
PF00690	Cation.transporter.ATPase..N.terminus	[9]
PF00735	Septin	[8]
PF00754	F5.8.type.C.domain	[9]
PF00777	Glycosyltransferase.family.29..sialyltransferase.	[10]
PF00780	CNH.domain	[11]
PF00782	Dual.specificity.phosphatase..catalytic.domain	[23]
PF00795	Carbon.nitrogen.hydrolase	[7]
PF00822	PMP.22.EMP.MP20.Claudin.family	[20]
PF00855	PWWP.domain	[12]
PF00884	Sulfatase	[8]
PF00888	Cullin.family	[7]
PF00899	ThiF.family	[8]
PF00907	T.box	[11]
PF00989	PAS.fold	[12]
PF00999	Sodium.hydrogen.exchanger.family	[11]
PF01007	Inward.rectifier.potassium.channel	[10]
PF01023	S.100.ICaBP.type.calcium.binding.domain	[8]

PF01049	Cadherin.cytoplasmic.region	[10]
PF01064	Activin.types.I.and.II.receptor.domain	[9]
PF01094	Receptor.family.ligand.binding.region	[19]
PF01105	emp24.gp25L.p24.family.GOLD	[9]
PF01145	SPFH.domain...Band.7.family	[10]
PF01217	Clathrin.adaptor.complex.small.chain	[8]
PF01284	Membrane.associating.domain	[20]
PF01302	CAP.Gly.domain	[9]
PF01336	OB.fold.nucleic.acid.binding.domain	[10]
PF01369	Sec7.domain	[10]
PF01370	NAD.dependent.epimerase.dehydratase.family	[8]
PF01388	ARID.BRIGHT.DNA.binding.domain	[11]
PF01390	SEA.domain	[12]
PF01392	Fz.domain	[14]
PF01399	PCI.domain	[10]
PF01404	Ephrin.receptor.ligand.binding.domain	[11]
PF01412	Putative.GTPase.activating.protein.for.Arf	[14]
PF01423	LSM.domain	[7]
PF01429	Methyl.CpG.binding.domain	[7]
PF01433	Peptidase.family.M1	[7]
PF01436	NHL.repeat	[7]
PF01454	MAGE.family	[23]
PF01471	Putative.peptidoglycan.binding.domain	[13]
PF01485	IBR.domain	[10]
PF01490	Transmembrane.amino.acid.transporter.protein	[9]
PF01529	DHHC.palmitoyltransferase	[16]
PF01534	Frizzled.Smoothened.family.membrane.region	[7]
PF01535	PPR.repeat	[7]
PF01553	Acyltransferase	[9]
PF01569	PAP2.superfamily	[11]
PF01576	Myosin.tail	[10]
PF01582	TIR.domain	[8]
PF01585	G.patch.domain	[13]
PF01602	Adaptin.N.terminal.region	[10]
PF01663	Type.I.phosphodiesterase...nucleotide.pyrophosphatase	[7]
PF01694	Rhomboid.family	[7]
PF01753	MYND.finger	[11]
PF01759	UNC.6.NTR.C345C.module	[8]
PF01762	Galactosyltransferase	[7]
PF01794	Ferric.reductase.like.transmembrane.component	[8]
PF01823	MAC.Perforin.domain	[8]

PF01839	FG.GAP.repeat	[15]
PF01852	START.domain	[9]
PF01926	X50S.ribosome.binding.GTPase	[10]
PF01979	Amidohydrolase.family	[7]
PF02037	SAP.domain	[11]
PF02141	DENN..AEX.3..domain	[11]
PF02191	Olfactomedin.like.domain	[8]
PF02205	WH2.motif	[10]
PF02225	PA.domain	[9]
PF02338	OTU.like.cysteine.protease	[9]
PF02463	RecF.RecN.SMC.N.terminal.domain	[8]
PF02493	MORN.repeat	[9]
PF02518	Histidine.kinase...DNA.gyrase.B...and.HSP90.like.ATPase	[12]
PF02736	Myosin.N.terminal.SH3.like.domain	[7]
PF02759	RUN.domain	[11]
PF02770	Acyl.CoA.dehydrogenase..middle.domain	[12]
PF02793	Hormone.receptor.domain	[16]
PF02798	Glutathione.S.transferase..N.terminal.domain	[11]
PF02820	mbt.repeat	[7]
PF02828	L27.domain	[8]
PF02864	STAT.protein..DNA.binding.domain	[7]
PF02883	Adaptin.C.terminal.domain	[7]
PF02893	GRAM.domain	[13]
PF02931	Neurotransmitter.gated.ion.channel.ligand.binding.domain	[25]
PF02932	Neurotransmitter.gated.ion.channel.transmembrane.region	[23]
PF02984	Cyclin..C.terminal.domain	[8]
PF02991	Autophagy.protein.Atg8.ubiquitin.like	[7]
PF03006	Haemolysin.III.related	[11]
PF03062	MBOAT..membrane.bound.O.acyltransferase.family	[7]
PF03114	BAR.domain	[8]
PF03133	Tubulin.tyrosine.ligase.family	[9]
PF03143	Elongation.factor.Tu.C.terminal.domain	[9]
PF03144	Elongation.factor.Tu.domain.2	[14]
PF03151	Triose.phosphate.Transporter.family	[10]
PF03171	X2OG.Fe.II..oxygenase.superfamily	[10]
PF03372	Endonuclease.Exonuclease.phosphatase.family	[18]
PF03455	dDENN.domain	[11]
PF03456	uDENN.domain	[12]
PF03765	CRAL.TRIO..N.terminal.domain	[9]
PF03810	Importin.beta.N.terminal.domain	[8]
PF03826	OAR.domain	[8]

PF03953	Tubulin.C.terminal.domain	[15]
PF04089	BRICHOS.domain	[7]
PF04408	Helicase.associated.domain..HA2.	[11]
PF04851	Type.III.restriction.enzyme..res.subunit	[7]
PF04969	CS.domain	[9]
PF05729	NACHT.domain	[7]
PF05739	SNARE.domain	[10]
PF05986	ADAM.TS.Spacer.1	[13]
PF06512	Sodium.ion.transport.associated	[7]
PF06602	Myotubularin.like.phosphatase.domain	[8]
PF06747	CHCH.domain	[7]
PF07525	SOCS.box	[22]
PF07546	EMI.domain	[8]
PF07562	Nine.Cysteines.Domain.of.family.3.GPCR	[8]
PF07648	Kazal.type.serine.protease.inhibitor.domain	[22]
PF07654	Immunoglobulin.C1.set.domain	[22]
PF07716	Basic.region.leucine.zipper	[17]
PF07717	Oligonucleotide.oligosaccharide.binding..OB..fold	[10]
PF07992	Pyridine.nucleotide.disulphide.oxidoreductase	[11]
PF08028	Acyl.CoA.dehydrogenase..C.terminal.domain	[9]
PF08240	Alcohol.dehydrogenase.GroES.like.domain	[9]
PF08241	Methyltransferase.domain	[13]
PF08242	Methyltransferase.domain	[10]
PF08441	Integrin.alpha	[14]
PF08447	PAS.fold	[15]
PF08516	ADAM.cysteine.rich	[22]
PF08659	KR.domain	[9]
PF08686	PLAC..protease.and.lacunin..domain	[7]
PF08736	FERM.adjacent..FA.	[8]
PF09279	Phosphoinositide.specific.phospholipase.C..efhand.like	[8]