Novelty or Nuisance? Where Lineage-Specific Genes Come From and Why It Matters

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Committee on Higher Degrees in Biophysics

have examined a dissertation entitled

Novelty or Nuisance?
Where Lineage-Specific Genes Come From
and Why It Matters

presented by **Caroline M. Weisman**

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature ___ _____

Typed name: Prof. Michael M. Desai

Signature _____ _____

Typed name: Prof. Cassandra G. Extavour

Signature _____ _____

Typed name: Prof. Michael T. Laub

Signature _____ _____

Typed name: Prof. John R. Wakeley

Date: August 13, 2021

Novelty or Nuisance?

Where Lineage-Specific Genes Come From and Why It Matters

A dissertation presented by

Caroline M. Weisman

to

The Committee on Higher Degrees in Biophysics

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in the subject of Biophysics

Harvard University

Cambridge, Massachusetts

August 2021

Dissertation Advisors: Sean Eddy, Andrew Murray Caroline M. Weisman

Novelty or Nuisance? Where Lineage-Specific Genes Come From and Why It Matters

Abstract

"Lineage-specific" genes, defined operationally as those lacking detected homologs outside of a narrow group of related species, are widely interpreted as novel genes. In this capacity, they have become the subject of significant evolutionary interest seeking to understand the nature and consequences of genetic novelty. Are all or most lineage-specific genes truly evolutionarily novel? Or might there be other, more mundane, reasons that these genes lack homologs beyond their lineage? Here, I explore two sources of lineage-specific genes that do not involve novelty. The first is variation in the genome annotation methods that are used for different species in a comparative analysis; the second is the failure of homology detection algorithms to identify homologs that do actually exist outside of a given lineage. I find evidence that these contribute large numbers of the total lineage-specific genes, suggesting that interpreting lineage-specific genes as novel is not generally reliable. I explore a particular example in which such an interpretation may have confused attempts to understand the evolution of a novel animal trait, and show how an analysis method developed here informs a different interpretation, altering conclusions about evolutionary history. Finally, I review what is known about de novo genes, a particularly poorly-understood type of novel gene.

Table of Contents

Acknowledgements

More thanks are due than I can list; what follows is surely an incomplete accounting.

I am grateful to my advisors, Andrew Murray and Sean Eddy, for their advice, support, and permissiveness; to my committee members, Michael Desai, Cassandra Extavour, Michael Laub, and John Wakeley; to members of the Murray and Eddy labs, especially Beverly Neugeboren, Nichole Wespe, Tom Jones, and Nick Carter for their technical support; to faculty and staff in the Center for Systems Biology, the Center for Quantitative Biology, the Department of Molecular and Cellular Biology, and the Program in Biophysics, especially Michele Jakoulov, Jim Hogle, Bridget Queenan, Linda Kefalas, and Peter Arvidson, who provided enormous support well beyond the logistical; to my cohort in the Biophysics program; to all members of the Department, who contributed to an unparalleled intellectual environment and to unparalleled Friday happy hours; to the students and staff of Dunster House, my home for the final five years of graduate school; and to my partner, Brian Claus, and my parents, Bruce Weisman and Kate Beckingham, for their patience, loyalty, love, and unflagging support.

Dedication

To my parents, who led me to science: not deliberately, by pressure or expectation, but quietly, by their shining example.

Introduction: Homology, novelty, and lineage-specific genes

Homology is a central notion in evolutionary biology. Despite vast diversity in organismal form and function, all life on Earth shares a common origin. Many genes present in this common ancestor have been carried forward into its modern descendants: genomes spanning even the farthest reaches of the tree of life are rife with components recognizable as extremely similar to one another, a phenomenon that we interpret as evidence of this common descent. Even more striking than how many genes disparate organisms share is how many *functions* they have in common. It is not just the parts list that evolution has retained: it is how they run the machine. All but necessary in any discussion of the pervasiveness of homology is Jacques Monod's famous adage: "What is true for *E. coli* is true for the elephant." This quip now represents a sentiment taken so much for granted that it has acquired the feel of a quaint platitude. *Everyone knows* that most things in biology are homologous. It is our default expectation, so omnipresent that it fades from conscious awareness, the nose on our face as we peer into nature.

Beyond repeated empirical demonstration of the broad conservation of genes and the functions that they encode, compelling a priori intuitions bolster the conviction that most biological processes have homologs, and thereby evolutionary roots, elsewhere. Even structures and functions that *appear* new are much likelier to result from modifications of components of existing ones than to have been created from entirely new components. The prospect of getting useful biology from nothing simply seems an astronomically tall order compared to getting it from *something*. It must, of course, have happened once or a few times, at the very origin of life, in a process that has been likened to a molecular "big bang" [1]; but, with this initial hurdle mounted and some cellular machinery in place, the argument goes, it becomes vastly easier, and

so vastly more likely, for evolution to modify and re-use existing components than to invent processes or structures totally anew [1]. Consistent with this line of thinking, early work estimated that the number of evolutionarily distinct protein families was much smaller than the total number of proteins: perhaps on the order of a thousand [2]. Again, a compulsory adage, this time from Monod's partner Francois Jacob: "Evolution does not produce novelties from scratch" [3]. In the next section, on "Molecular Tinkering," the prospect of novel proteins in particular prompts a pointed follow-up: "the probability that a functional protein would appear de novo by random association of amino acids is practically zero" [3]. Jacob belabored the supremacy of homology over novelty in the case of proteins, the cell's basic constituents, in particular. Here, many of us agree, it is especially hard to imagine how novelty could arise, how function might emerge from nothing.

That all genes ought to have homologs was put systematically to the test with the advent of genome sequencing projects. In the completed *S. cerevisiae* genome sequence, 30% of inferred genes "had no clear-cut sequence homologs in any organism" [4]. Other genomes from across the tree of life yielded similar results [5, 6]. It was hoped that this was an artifact of too-sparse sampling, and that increasing database size would squash the problem of these so-called "orphan" proteins [7]. While broader sequencing has somewhat decreased their abundance, a persistent minority of proteins in any species still remains orphaned, lacking detectable homologs anywhere else in the tree of life [8, 9]. The problem deepens when so-called "lineage-specific" or "taxonomically-restricted" proteins are considered: every taxon, young or old, seems to have its own suite of unique proteins, not found in any species beyond its borders [8, 9]. I will refer to both of these classes as "lineage-specific genes," with orphans being the base case in which the lineage is a single species. I will define them throughout this thesis as is standard in

the field: purely operationally, as those that lack detected homologs beyond the specified taxonomic boundary. I note that strict adherence to this operational definition can at times produce unintuitive statements: for example, a gene with no *detected* homologs beyond a lineage, but for which there is circumstantial evidence pointing toward such homologs nonetheless existing, will still be referred to as lineage-specific. Though at times unwieldy, this terminology is consistent with use in the field, and so I adhere to it here.

Lineage-specific genes have remained quite abundant, despite the rapid recent increase in sequencing density across the tree of life. A few quantitative examples: 23% of *Caenorhabditis elegans* genes are specific to the genus [10]; 6% of honey bee (*Apis mellifera*) genes specific to insects [11]; 25% of ash tree (*Fraxinus excelsior*) genes specific to the species [12]; and 1% of human genes specific to primates [13].

What are these genes? Lacking homologs, they appear on face to be evolutionarily novel. This interpretation rapidly dominated: in many circles, lineage-specificity became, and remains, essentially synonymous with novelty [9, 14-28]. The prospect of such abundant genetic novelty, so starkly in contrast to the unflashy truism of pervasive homology, captured imaginations and sparked a rush to study these novel genes. With available genome sequences from increasingly diverse organisms mounting, an obvious and fairly simple approach to learning about these genes was to computationally, systematically and at large scale, identify them in different taxa and then analyze them in some way. A cottage industry taking this approach sprung up: a lineage was selected; lineage-specific genes, lacking outgroup homologs, were identified, usually with a BLASTP search; and their properties were analyzed, again usually computationally and at large scale [10, 13, 18, 29-45]. (The term “phylostratigraphy” was introduced to refer to this approach when used to identify genes specific to lineages of varying ages, inferred to have been born at

corresponding “strata” in evolution [32].) The results of these analyses *of lineage-specific genes* – the numbers of genes in different taxa, the properties of their sequences, any hints of apparent functions – are often explicitly interpreted as results *about novel genes* [13, 18, 29-36].

In precisely what way are these genes thought to be novel? Different proposals have been made. An early and influential hypothesis held that these genes are duplicates that have gained an entirely new function: although they have homologs in the formal sense, these bear no meaningful resemblance, entirely different in function and unrecognizable by homology search algorithms in sequence. In the proposed model for how this occurs, in a period of relaxed selective constraint following the duplication event, these genes undergo rapid evolution in which they rapidly sample mutational space and thereby acquire a novel function. Selective constraint then reemerges to conserve this function, slowing gene evolution and allowing detection of homologs that emerge after this point and so share the new function. The result of this process is that these genes are lineage-specific, with the extent of the lineage corresponding to the evolutionary moment at which their new function as gained [21]. Another version of novelty attributed to lineage-specific genes, simpler but arguably more exotic, is that they are “de novo” genes, emerged from previously noncoding DNA [21, 22]. Individual studies sometimes clearly adopt one of these interpretations, usually without justification, but sometimes do not [9, 14-28]. The field therefore strikes me as somewhat confused on this point, a lack of clarity that I interpret as reflecting that the focal claim – what drives interest in these genes -- seems not to be anything about the molecular nature of the origin of lineage-specific genes per se, but just the overall notion they are in *some way* evolutionary novel. I will use the term “novel genes” to indicate this kind of meaningful evolutionary novelty implied by the novelty hypothesis.

I will refer to these and similar interpretations collectively as the "novelty hypothesis": again, the view that lineage-specific genes are all, or overwhelmingly, meaningfully novel. *A lack of detected homologs alone* is taken as a reliable indicator of novelty. Terminologically, the novelty hypothesis holds that all lineage-specific genes are novel genes.

Conclusions about novel genes drawn in studies that adopt the novelty hypothesis are tantalizing. While results from studies that take this approach are reviewed extensively elsewhere [22, 46, 47], a brief sampling includes: that the properties of de novo genes are distinct from those of more conserved genes, in that they are shorter [23, 48-52], faster-evolving [23, 48-52], lower-expressed [18, 23], more tissue-specific [18, 23], and more disordered [53]; that in animals, de novo genes are enriched in particular tissues like brain and testis; that the emergence rate of novel genes has varied across evolutionary time, and has peaked in paleontologically significant eras like the Cambrian Explosion [34]; and, of course, that de novo genes are shockingly abundant, repeatedly reported to number hundreds to thousands in any individual species, and to emerge at a rate comparable to or higher than gene duplicates [21, 22].

But these conclusions, and all attempts to study novel genes by way of studying lineage-specific genes, are only as good as the assumption on which they are based: the novelty hypothesis. *Are all, or even most, lineage-specific genes actually novel?* If the answer is no, these results do not reflect the true biology of novel genes. Any signal of real novelty will be polluted by the noise of whatever other genes have wound up in the mix.

To understand whether studies of lineage-specific genes can be interpreted as bearing on the properties of novel genes, one needs to identify other sources of lineage-specific genes and know what proportion of total lineage-specific genes they contribute. Two outcomes are possible. Lineage-specific genes may be comprised exclusively or largely of truly novel genes.

Alternatively, they could be comprised exclusively or largely of genes that appear lineage-specific for reasons *that have nothing to do with novelty*. In this case, we cannot study novel genes by studying lineage-specific genes, and cannot trust existing conclusions about novel genes arrived at by doing so.

I take up this question here. In Chapters One and Two, I describe two possible sources of lineage-specific genes that have nothing to do with evolutionary novelty, and determine approximately what proportion of lineage-specific genes they seem to comprise. To briefly foreshadow, the first source is the use of different genome annotation methods for different species within a single comparative analysis, which produces lineage-specific genes merely due to some of those different methods mistakenly including or excluding the same sequence in the different annotations, making it appear as though it is present only in a subset of species. The second source is the failure of homology detection algorithms like BLAST to identify homologs outside of the given lineage even when they are present. I find evidence that both sources contribute substantial fractions of the total lineage-specific genes. From this, I conclude that the novelty hypothesis is not generally reliable: most lineage-specific genes are not novel, and so I cannot study novel genes by studying lineage-specific genes.

In Chapter Three, I identify a particular instance in which adherence to the novelty hypothesis has confused attempts to address an outstanding question in animal evolution, and demonstrate how a revised understanding of the relevant lineage-specific genes, provided by a method developed here, can correct this confusion.

The first three chapters are disappointing to those, including myself, who seek an understanding of genetic novelty. This was the initial aim of the work described here. Its primary conclusion is a negative one: that most studies of lineage-specific genes likely do not

meaningfully bear on novel genes. As a small step toward some positive insight into novel genes, Chapter Four presents a review aiming to comprehensively summarize more reliable data regarding what is known about the origins and functions of the most radical molecular novelties: de novo genes. A brief and fair summary of this chapter, and indeed of this thesis as a whole, is: I do not know very much. If this work amounts to anything, I hope it is to encourage more direct and reliable study of this exciting class of genes.

Chapter 1: The contribution of genome annotation errors to analyses of lineage-specific genes

Introduction

The rapid decline in cost of genome sequencing has provided fuel for sequence-based attempts to identify lineage-specific genes [10, 13, 18, 29-45]. Conceptually, the methodology underlying these attempts is straightforward: one identifies lineage-specific genes by searching for homologs of the genes from species in the focal lineage in all outgroup species . Genes for which homologs cannot be found are considered lineage-specific [8, 47].

Such an analysis requires one to define the set of genes within the focal lineage. Generally, this is done by relying on a particular annotation of the focal lineage genomes: a hypothesis for what genes they encode. Only genes included in this annotation are used in the homology search [32, 34, 46], and so results depend crucially on the annotation.

The process of producing an accurate genome annotation – one that includes all genes in the genome, and omits all non-genic sequences -- is by no means a solved problem [54]. A large number of genome annotation methods are in widespread use, with no clear consensus about which is the most accurate in any given situation or what the false positive, false negative, and trade-off between the two looks like [55]. Common methods have broad similarities, but differ in important details, as do particular implementations of them [54]. Among annotation methods in widespread use are custom pipelines at large bioinformatic databases (NCBI [56], Ensembl [57]), hand-curated model organism annotation (Flybase [58], Wormbase [59]), crowdsourced

annotation (VectorBase [60]), and idiosyncratic combinations of a subset of the available software packages that perform some or all of the steps inherent in annotation (Maker [61], PASA [62]), run with custom parameters chosen by individual researchers. A few major differences in these methods are worth noting. Some of these methods are purely sequence-based, while others bring to bear expression data like RNA-seq. When used, such expression data are produced at varying sequencing depths and from varying numbers of tissues or conditions. Some methods incorporate purely de novo predictions, while others require additional support for genes from either homology or expression.

Insofar as different annotation methods exist and disagree in their predictions, they cannot all be right. It may also be the case that none of them are right. There is therefore undeniably some rate of error in the genome annotations that underlie comparative genome analyses. Previous work has suggested two ways that errors in genome annotations may produce spurious lineage-specific genes. In one, a gene is annotated in the lineage, but its homologs are incorrectly omitted in outgroups [63-65]. In another, a *non-genic* sequence is incorrectly annotated in the lineage, but is correctly omitted in outgroups [53].

Both of these cases are caused by erroneous differential annotation: homologous sequences are inconsistently annotated in the ingroup and outgroup, when they should actually either both be included or both omitted. Because different annotation methods can disagree about what sequences are genes, I wondered whether these errors could be particularly abundant when different annotation methods are used on different species within a study. I term this "annotation heterogeneity"; indeed, it is extremely common in comparative genome analyses, including those seeking to identify lineage-specific genes [14, 15, 23, 33, 51]. Although previous work has highlighted how genome annotation errors may cause spurious lineage-specific genes [63-65]

[53], the concept of annotation heterogeneity, and its potential to cause or enhance the rate of spurious lineage-specific genes within these error modes, has not been explored.

Here, I explore the effects of annotation heterogeneity on the inferred number of lineage-specific genes, hypothesizing that increased heterogeneity should falsely inflate the apparent number of such genes. In six case studies, I compare the number of lineage-specific genes when all species are annotated with the same method ("uniform annotations") to when they are annotated with different methods ("heterogeneous annotation"). I also investigate the possibility that annotation heterogeneity is responsible for or contributes to reported correlations between the inferred age of lineage-specific genes (with the age of the lineage corresponding to the age of the gene) and characteristics like length, evolutionary rate, and GC content, which have been used to propose models of novel gene evolution [18, 23, 51-53].

Identifying an appropriate dataset: taxa in which each species has a genome assembly with genome annotations from two methods

To isolate and assess the effect of annotation heterogeneity on the apparent number of lineage-specific genes, I sought groups of species for which a) all species had been annotated using the same single method, and b) the same genome assembly of each species had been independently annotated using some different second method. This allowed us to compare the apparent number of genes specific to a lineage within the group when all species annotations were generated by the same method ("uniform annotation") to the number when annotations were generated by different methods ("heterogeneous annotation"). I identified four groups of five species, each less than approximately 60 My old: cichlids, primates, bats, and rodents.

Assembly accessions, links to both annotations, annotation sources, and a brief description of annotation methods are listed in Supplemental Table 1.

We used existing annotations from common methods instead of generating my own to make results representative of real studies. Accordingly, annotations were taken from a variety of standard sources: bioinformatic databases (NCBI [56], UCSC [66], Ensembl [57]), sequencing consortia (Bat1K [67]), and individual research groups [68]. These and similar sources are generally treated as functionally equivalent: explanation is not generally offered for which is used, and the effect of sources on results is not generally tested or discussed. my initial expectation was therefore that these annotations would be highly similar.

Different annotations of the same genome assembly result in hundreds to thousands of proteins unique to one method

For annotation heterogeneity to affect inferences of lineage-specific genes, different annotation methods applied to the same genome sequence must produce different sets of proteins. Specifically, they must differ in the proteins to which they show significant similarity in a homology search.

To get a sense of the magnitude of these differences, for each species, I performed a reciprocal homology search between the two annotations from different methods. For each species' two annotations, I used a BLASTP search [69] (version 2.6.0) of all proteins in each annotation to ask if a significantly similar (E=0.001) protein was present in the other annotation. I used BLASTP with a relatively permissive E-value threshold rather than requiring identical proteins to permit common minor annotation differences, like start/stop site or splice structure, that do not affect homology search results.

In the pairs of annotations considered here, the percentage of proteins in one annotation missing a similar sequence in the other varies between 0.6% and 9.7%, and the absolute number

between 197 and 4143. Most annotations include many hundreds to thousands of proteins that have no similar sequence in the other annotation (Table 1).

Table 1.1: Percentages and numbers of genes in one annotation of a genome assembly lacking a significantly similar gene (by BLASTP, E=0.001) in another annotation of the same assembly.

Species	**Annotation 1 source**	**Annotation 2 source**	**Number/percent of genes in annotation 1 not found in annotation 2**	**Number/percent of genes in annotation 2 not found in annotation 1**
Cichlid fish				
Metraclimia zebra	Broad Institute	NCBI Eukaryotic annotation pipeline	3592/6.9%	706/1.8%
Pundamilia nyererei	Broad Institute	NCBI Eukaryotic annotation pipeline	3276/7.7%	668/1.7%
Astatotilapia burtoni	Broad Institute	NCBI Eukaryotic annotation pipeline	4110/7.8%	799/1.8%

Table 1.1 (Continued)

Neolamprologus brichardi	Broad Institute	NCBI Eukaryotic annotation pipeline	3568/9.7%	755/2.4%
Oreochromis niloticus	Broad Institute	NCBI Eukaryotic annotation pipeline	4143/6.2%	765/1.6%
Primates				
Macaca fascicularis	Ensembl "full genebuild"	NCBI Eukaryotic annotation pipeline	1510/3.3%	1044/1.7%
Macaca nemestrina	Ensembl "full genebuild"	NCBI Eukaryotic annotation pipeline	1295/2.8%	1623/2.4%
Mandrillus leucophaeus	Ensembl "full genebuild"	NCBI Eukaryotic annotation pipeline	1406/3.4%	693/1.8%

Table 1.1 (Continued)

Rhinopithecus bieti	Ensembl "full genebuild"	NCBI Eukaryotic annotation pipeline	1233/2.8%	1476/3.0%
Cebus imitator	Ensembl "full genebuild"	NCBI Eukaryotic annotation pipeline	926/1.7%	602/1.5%
Rodents				
Mus musculus	Ensembl "full genebuild"	NCBI Eukaryotic annotation pipeline	1523/2.2%	471/0.6%
Mus caroli	UCSC	NCBI Eukaryotic annotation pipeline	1496/3.0%	590/1.2%
Mus pahari	UCSC	NCBI Eukaryotic annotation pipeline	1596/3.2%	413/1.0%

Table 1.1 (Continued)

Rattus norwegicus	Ensembl "mixed genebuild"	NCBI Eukaryotic annotation pipeline	458/1.6%	716/1.3%
Peromyscus maniculatus	Ensembl "full genebuild"	NCBI Eukaryotic annotation pipeline	267/0.9%	517/1.1%
Bats				
Myotis lucifugus	Ensembl "full genebuild"	NCBI Eukaryotic annotation pipeline	197/1.0%	622/1.4%
Myotis brandtii	Beijing Genomics Institute	NCBI Eukaryotic annotation pipeline	867/4.5%	1370/3.4%
Myotis myotis	Bat1K	NCBI Eukaryotic annotation pipeline	3448/7.5%	704/1.2%

Table 1.1 (Continued)

Molossus molossus	Bat1K	NCBI Eukaryotic annotation pipeline	3107/5.8%	486/1.1%
Pteropus alecto	Beijing Genomics Institute	NCBI Eukaryotic annotation pipeline	1338/6.8%	955/2.4%

While these values do not translate directly into changes in apparent numbers of lineage-specific genes, they demonstrate the extent to which using different annotation methods affects the resulting set of proteins from the perspective of homology detection.

Different patterns of annotation heterogeneity may differently affect the inferred number of lineage-specific genes

When different methods have been used to annotate different species in a comparative analysis, there are different possible patterns in which the different methods may be arranged on the topology of the species tree. I hypothesized that different such patterns may differently affect how many lineage-specific genes are inferred. In particular, because any sequence annotated in the focal lineage and not annotated in the outgroups will appear to be a lineage-specific gene, I hypothesized that the differences in annotation methods between the lineage and outgroups would be of particular impact. Three such classes of annotation heterogeneity patterns are fairly common in existing studies, and so I focused on these, described below.

In the first pattern, the entire focal lineage is annotated using a single method, and all outgroups are annotated with a different single method. I refer to this as "phyletic annotation," and it occurs in at least two scenarios. Studies that newly sequence a lineage often use a custom method to annotate ingroups, and may then compare them to outgroup annotations from a single other source (e.g. Ensembl). Additionally, studies using existing annotations may encounter a correlation between taxon and annotation method because groups that produce annotations often select species taxonomically (e.g. studies of particular taxa, sequencing consortia/database initiatives for particular taxa) [18, 70]. If use of the same annotation method produces more

similar sets of proteins, phyletic annotation should maximize differences between the proteins in the lineage and outgroups; I therefore hypothesized that this pattern would produce the highest number of spurious lineage-specific genes.

In the second pattern, the ingroup is annotated with a single method, and the outgroups with a mixture of methods. I refer to this as "semi-phyletic" annotation. This occurs in scenarios similar to phyletic annotation, but where outgroup annotations are available from multiple sources (e.g. a mix of Ensembl and NCBI) [23, 27, 50-52, 70-72]. While any difference in annotation method between ingroups and outgroups risks some spurious lineage-specific genes, compared to phyletic annotation, semi-phyletic annotation should increase the chance that a gene from the lineage is annotated in at least one outgroup. I therefore hypothesized that this pattern would produce spurious lineage-specific genes, but fewer than would phyletic annotation.

In the third pattern, different annotation methods are used for different species in the analysis, but there is no systematic variation between those used for the ingroup and outgroup. I refer to this as "unpatterned" annotation heterogeneity. This commonly arises when existing annotations for species are available from a variety of sources, but without taxonomic bias [11, 14, 33, 50, 73-75]. Of the three patterns described here, this one minimizes annotation differences between ingroups and outgroups. While such differences should produce some spurious lineage-specific genes, I hypothesized that, among the three considered here, this method harbors the lowest potential for spurious lineage-specific genes.

Annotating a lineage with one method and outgroups with a different method ("phyletic annotation") greatly increases the apparent number of lineage-specific genes

We tested the impact of phyletic annotation on the apparent number of lineage-specific genes. I used two young (approximately 50 My old [68, 76]) taxa in which each species had been annotated by the same two methods: five cichlids, annotated both by a group at the Broad and NCBI; and five primates, annotated both by Ensembl and NCBI (Table 1). These pairs of sources exemplify the two scenarios above: a research group's lineage annotations compared to outgroup annotations from an established database, and lineage and outgroup annotations taken from two different established databases.

In each group, I selected a focal species based on the species topology: the cichlid *Metraclimia zebra*, and the primate *Macaca fascicularis* (Figure 1.1, Figure 1.2). I then asked how many of this species' proteins appear to be specific to lineages of different ages within the group, based on having no significant similarity to any proteins from outgroups to that lineage in a BLASTP search (Methods). For example, if a protein had no significant similarity to proteins from any other species, it was considered an 'orphan'; if it had no significant similarity to proteins from any species other than its two nearest neighbors, it was considered specific to that three-species lineage; and so on (Figure 1.1, Figure 1.2).

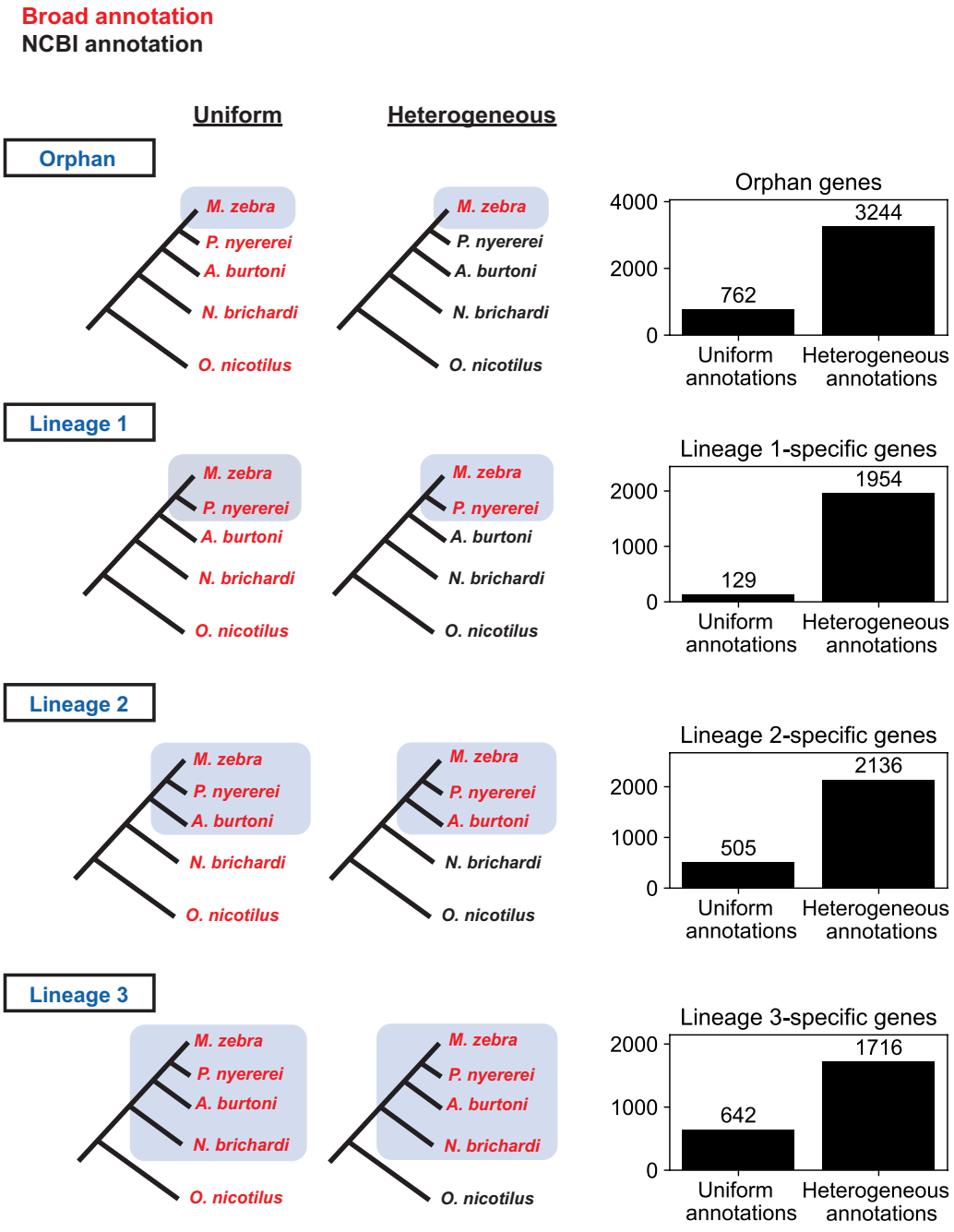

Figure 1.1: Comparison of the number of lineage-specific genes in cichlid fish when uniform annotations are used to when heterogeneous (phyletic) annotations are used. Each row shows genes specific to the lineage indicated by the blue box on the species trees. Species outside of the

(Continued) blue box are the outgroups in each analysis. The colors of species names depict which annotation source was used in the uniform (left tree) and heterogeneous (right) annotation schemes. The bar graphs depict the numbers of genes that appear specific to the given lineage in the uniform (left bar) and heterogeneous (right bar) annotation schemes. When phyletic heterogeneous annotations are used, there is a large increase in the number of apparent lineage-specific genes.

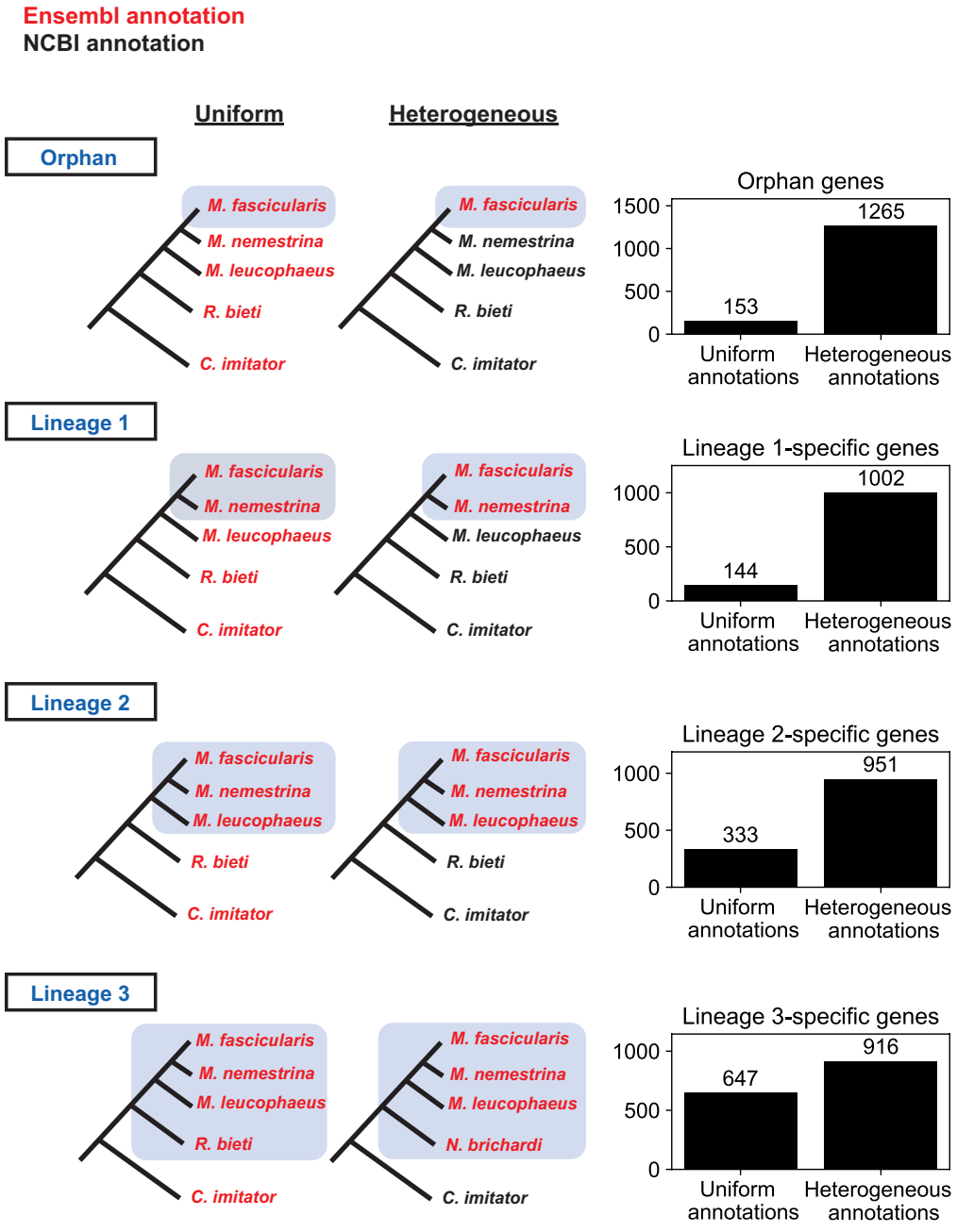

Figure 1.2: Comparison of the number of lineage-specific genes in primates when uniform annotations are used to when heterogeneous (phyletic) annotations are used. Each row shows genes specific to the lineage indicated by the blue box on the species trees. Species outside of the blue box are the outgroups in each analysis. The colors of species names depict which annotation source was used in the uniform (left tree) and heterogeneous (right) annotation

(Continued) schemes. The bar graphs depict the numbers of genes that appear specific to the given lineage in the uniform (left bar) and heterogeneous (right bar) annotation schemes. When phyletic heterogeneous annotations are used, there is a large increase in the number of apparent lineage-specific genes.

We first determined the number of genes specific to each lineage when the same annotation method was used for all species. For each lineage, I then replaced the annotations of all outgroups with the alternative annotation method, such that annotations were phyletic: all species in the lineage annotated by one method and all outgroup species by a second, different method. I then again determined the number of genes specific to each lineage. The difference between these two cases is the number of genes that appear lineage-specific due only to annotation heterogeneity.

Figure 1.1 shows the results of this analysis for cichlids, comparing uniform Broad annotations to a phyletic mixture of Broad and NCBI annotations. Figure 1.2 shows results for primates, comparing uniform Ensembl annotations to a phyletic mixture of Ensembl and NCBI annotations.

Compared to when all species were annotated by the same method, phyletic annotation heterogeneity consistently caused a large increase in the number of genes that appear to be specific to the lineage in question. In all lineages in both groups, annotation heterogeneity produced hundreds to thousands of lineage-specific genes. This corresponds to a percent increase ranging between 50% and 6000%, with typical values around 400%.

We also performed the same analysis in both lineages after reversing the pattern of the two annotation sources on the tree: beginning with the uniform set of NCBI annotations, and then considering heterogeneous annotations by mixing in annotations from the Broad or Ensembl in the cichlid and primate lineages, respectively. Results from these analyses were qualitatively similar, and are shown in Figures 1.3 and 1.4.

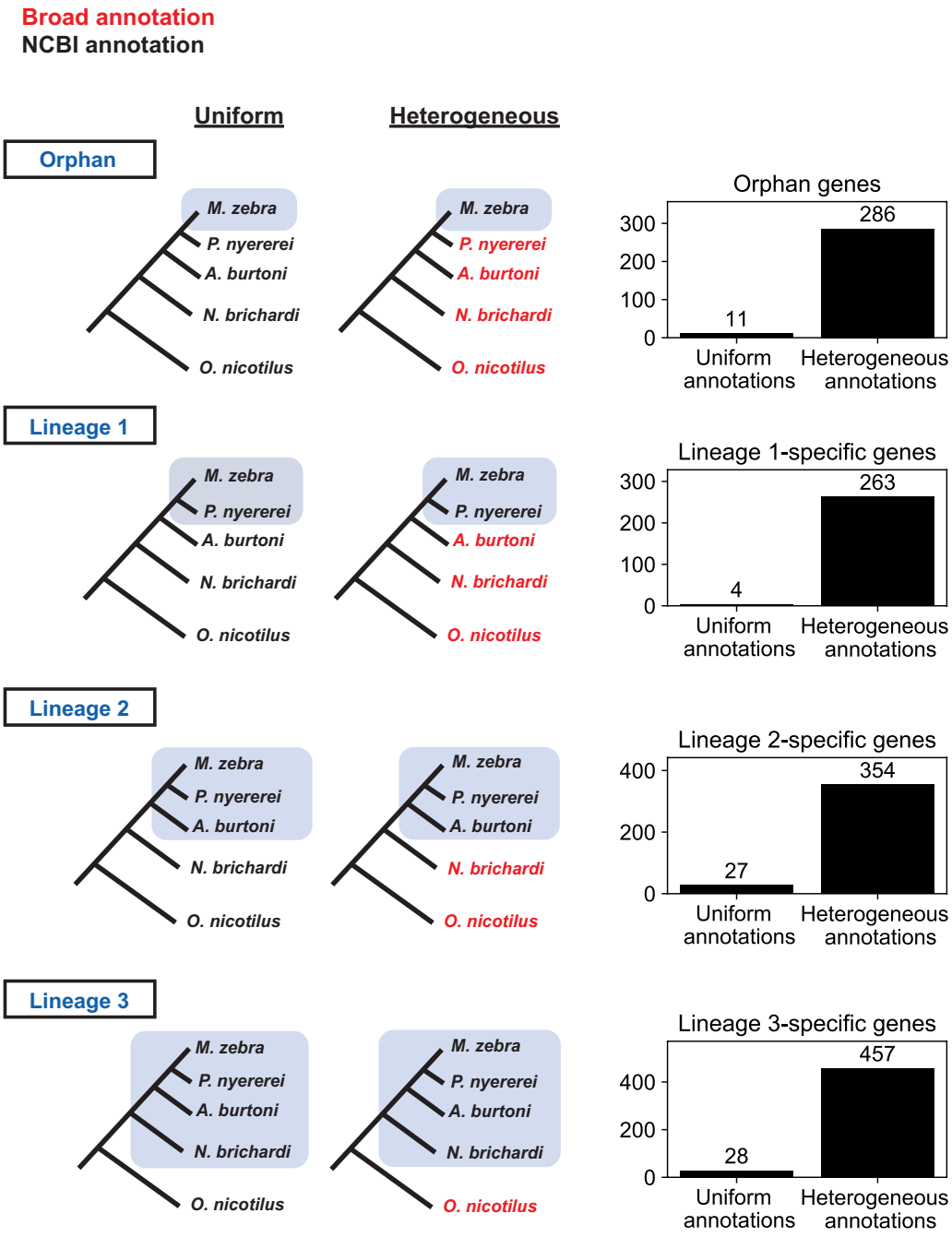

Figure 1.3: Comparison of the number of lineage-specific genes in cichlid fish when uniform annotations are used to when heterogeneous (phyletic) annotations are used. Each row shows genes specific to the lineage indicated by the blue box on the species trees. Species outside of the blue box are the outgroups in each analysis. The colors of species names depict which annotation

(Continued) source was used in the uniform (left tree) and heterogeneous (right) annotation schemes. These are switched relative to Figure 1.1 The bar graphs depict the numbers of genes that appear specific to the given lineage in the uniform (left bar) and heterogeneous (right bar) annotation schemes. When phyletic heterogeneous annotations are used, there is a large increase in the number of apparent lineage-specific genes.

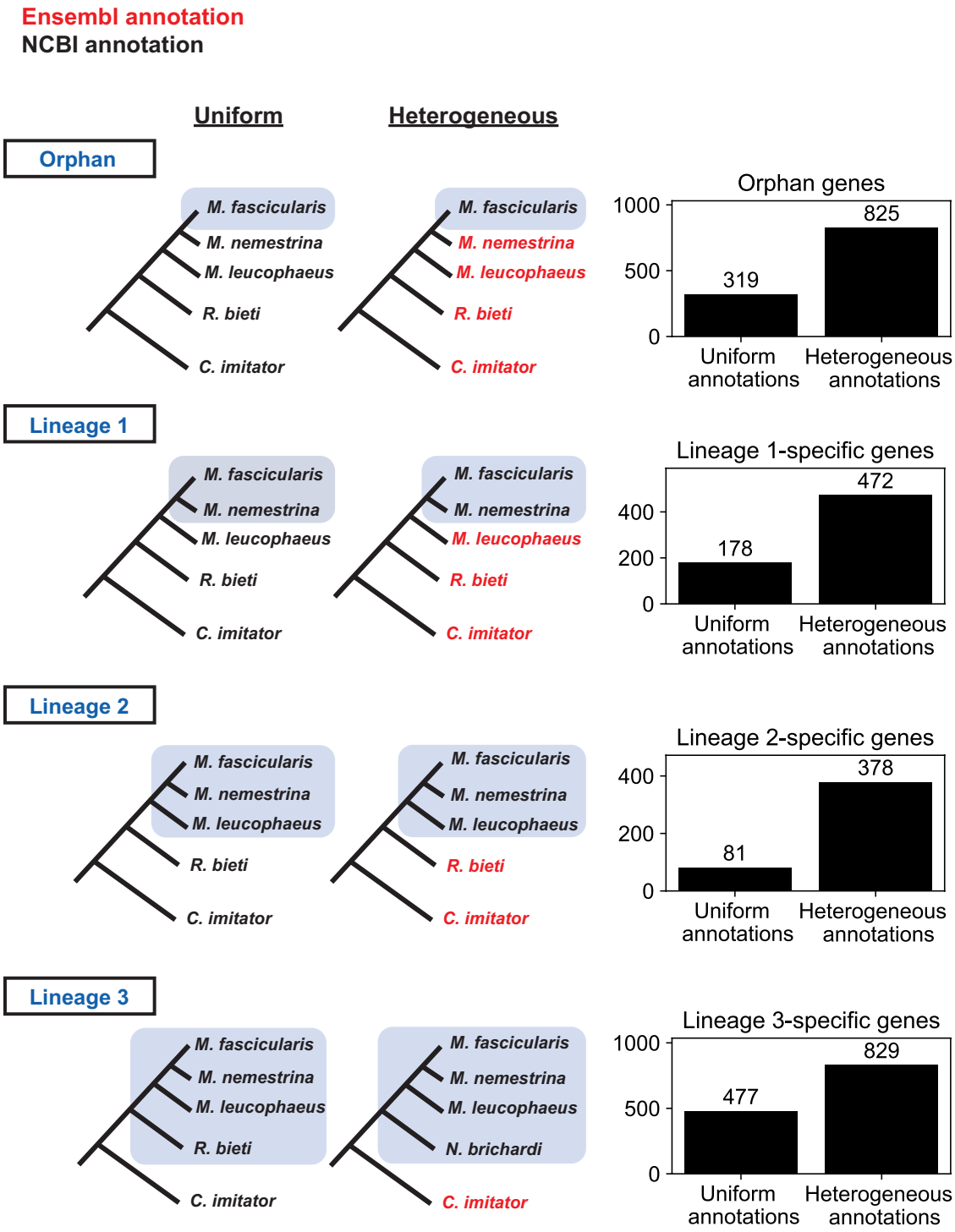

Figure 1.4: Comparison of the number of lineage-specific genes in primates when uniform annotations are used to when heterogeneous (phyletic) annotations are used. Each row shows genes specific to the lineage indicated by the blue box on the species trees. Species

(Continued) outside of the blue box are the outgroups in each analysis. The colors of species names depict which annotation source was used in the uniform (left tree) and heterogeneous (right) annotation schemes. These are switched relative to Figure 1.2. The bar graphs depict the numbers of genes that (Continued) appear specific to the given lineage in the uniform (left bar) and heterogeneous (right bar) annotation schemes. When phyletic heterogeneous annotations are used, there is a large increase in the number of apparent lineage-specific genes.

Considered in reverse, in almost all cases, a majority of genes that appear to be lineage-specific when heterogeneous annotations are used disappear if annotation methods are standardized. Very often, this majority is overwhelming.

Annotating a lineage with one method and outgroups with a mixture of other methods ("semi-phyletic annotation") increases the apparent number of lineage-specific genes

We tested the impact of semi-phyletic annotation on the apparent number of lineage-specific genes. I used two young (approximately 60 My old [77, 78]) taxa in which every species had been annotated by the same single method, and in addition had been annotated by some second, different method. The first is five rodents, annotated both by a) NCBI and b) a mixture of UCSC and two methods from Ensembl. The second is five bats, annotated by both a) NCBI and a mixture of Beijing Genomics Institute and the Bat1K consortium (Table 1). I chose these species and annotations by first determining the availability of one set of uniform annotations and then pulling a second set of heterogeneous annotations from wherever I found them to be available. This design was intended to emulate as closely as possible the approach taken in real studies, which often draw on available annotations regardless of source.

We repeated the same procedure as above to compare the number of genes that appear to be specific to each lineage within the group when uniform annotations and semi-phyletic annotations were used. Results are shown in Figures 1.5 (rodents) and 1.6 (bats).

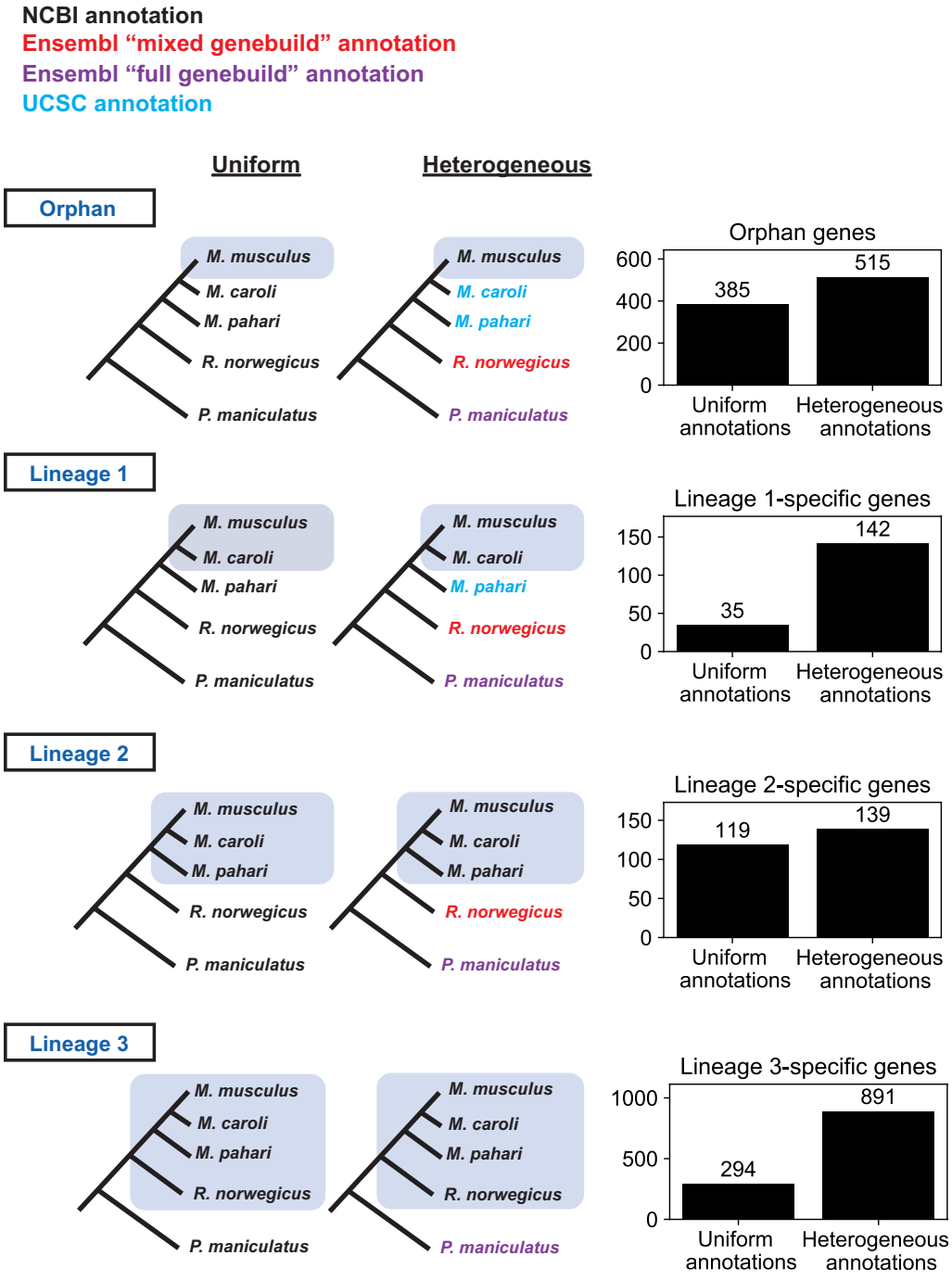

Figure 1.5: Comparison of the number of lineage-specific genes in rodents when uniform annotations are used to when heterogeneous (semi-phyletic) annotations are used. Each row shows genes specific to the lineage indicated by the blue box on the species trees. Species outside of the blue box are the outgroups in each analysis. The colors of species names depict which

(Continued) annotation source was used in the uniform (left tree) and heterogeneous (right) annotation schemes. The bar graphs depict the numbers of genes that appear specific to the given lineage in the uniform (left bar) and heterogeneous (right bar) annotation schemes. When semi-phyletic heterogeneous annotations are used, there is an increase in the number of apparent lineage-specific genes.

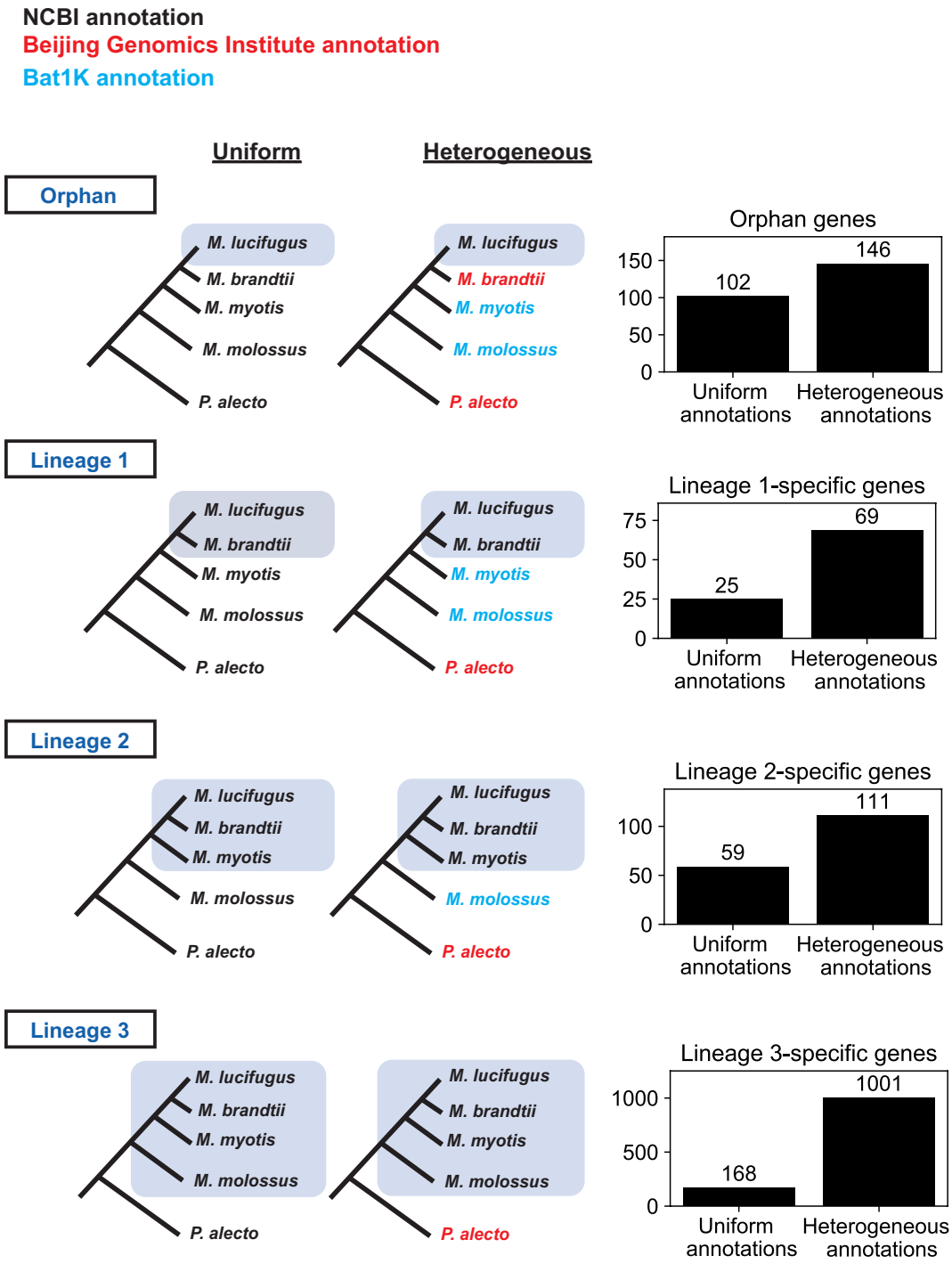

Figure 1.6: Comparison of the number of lineage-specific genes in bats when uniform annotations are used to when heterogeneous (semi-phyletic) annotations are used. Each row shows genes specific to the lineage indicated by the blue box on the species trees. Species outside

(Continued) of the blue box are the outgroups in each analysis. The colors of species names depict which annotation source was used in the uniform (left tree) and heterogeneous (right) annotation schemes. The bar graphs depict the numbers of genes that appear specific to the given lineage in the uniform (left bar) and heterogeneous (right bar) annotation schemes. When semi-phyletic heterogeneous annotations are used, there is an increase in the number of apparent lineage-specific genes.

Semi-phyletic annotation heterogeneity caused an increase in the number of apparent lineage-specific genes in all lineages in both groups. The magnitude of this effect varied between lineages, ranging from tens to hundreds of additional lineage-specific genes. This corresponds to a percent increase of between 20% and 600%. Though this effect size is smaller than for phyletic annotation, semi-phyletic annotation still causes an increase in the apparent number of lineage-specific genes.

Annotating species with a mixture of methods lacking taxonomic bias ("unpatterned annotation") increases the apparent number of lineage-specific genes

We sought to test the impact of unpatterned annotation heterogeneity on the apparent number of lineage-specific genes. I used the same two species groups from the section above, for which I had already identified a) one uniform set of annotations and b) one unpatterned heterogeneous set of annotations. I repeated the same procedure to compare the number of genes that appear to be specific to each lineage within the group when uniform annotations were used to when heterogeneous annotations were used. However, instead of systematically varying the annotation used to produce a semi-phyletic annotation pattern for each lineage (Figures 1.5, 1.6), I used the full set of heterogeneous annotations for all lineages (Figures 1.7, 1.8), producing unpatterned annotation heterogeneity.

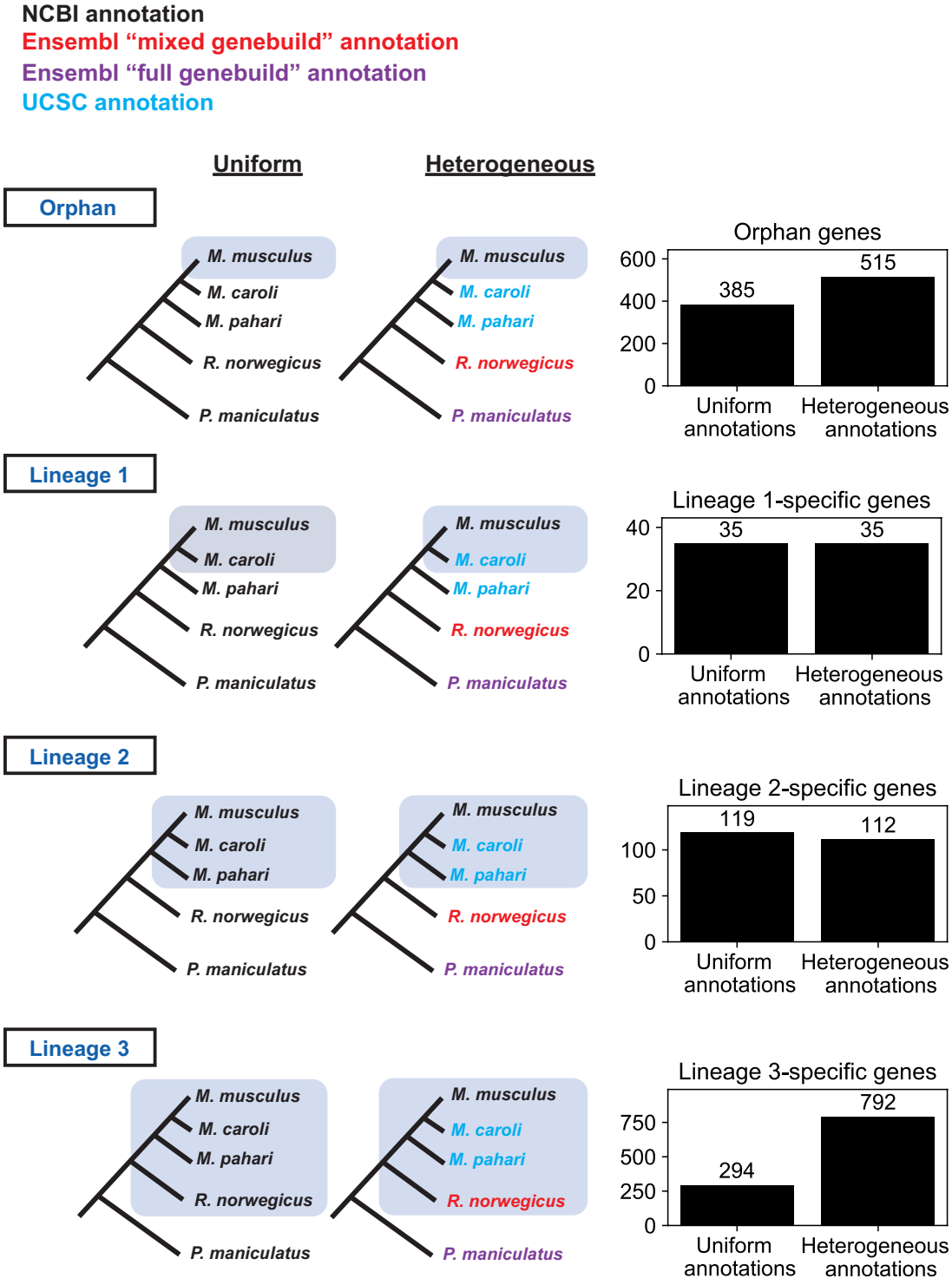

Figure 1.7: Comparison of the number of lineage-specific genes in rodents when uniform annotations are used to when heterogeneous (unpatterned) annotations are used. Each row shows genes specific to the lineage indicated by the blue box on the species trees. Species outside

(Continued) of the blue box are the outgroups in each analysis. The colors of species names depict which annotation source was used in the uniform (left tree) and heterogeneous (right) annotation schemes. The bar graphs depict the numbers of genes that appear specific to the given lineage in the uniform (left bar) and heterogeneous (right bar) annotation schemes. When unpatterned heterogeneous annotations are used, the number of apparent lineage-specific genes tends to increase.

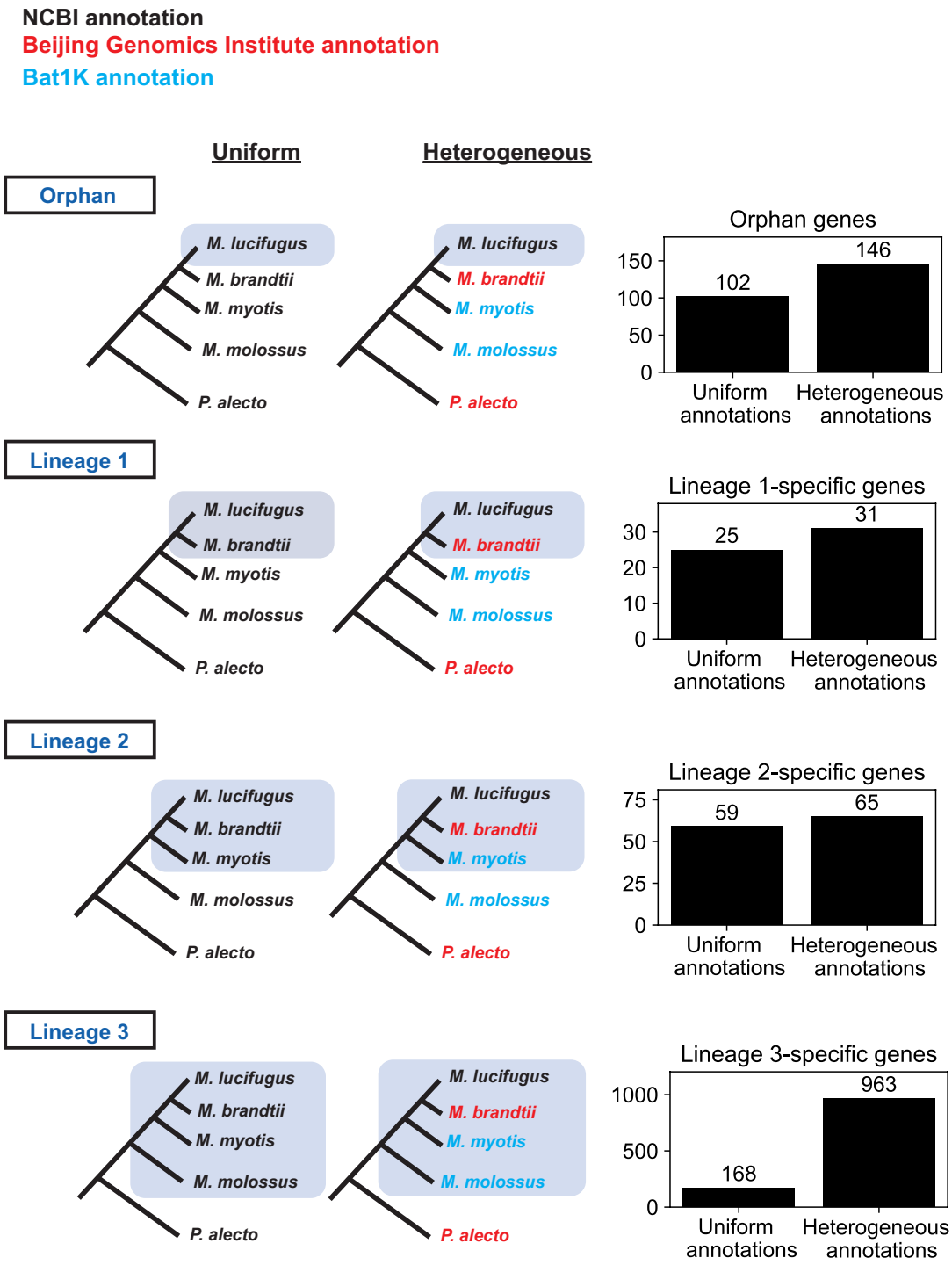

Figure 1.8: Comparison of the number of lineage-specific genes in bats when uniform annotations are used to when heterogeneous (unpatterned) annotations are used. Each row shows genes specific to the lineage indicated by the blue box on the species trees. Species outside

(Continued) of the blue box are the outgroups in each analysis. The colors of species names depict which annotation source was used in the uniform (left tree) and heterogeneous (right) annotation schemes. The bar graphs depict the numbers of genes that appear specific to the given lineage in the uniform (left bar) and heterogeneous (right bar) annotation schemes. When unpatterned heterogeneous annotations are used, the number of apparent lineage-specific genes increases.

Results for rodent and bats are shown in Figures 1.7 and 1.8. Unpatterned annotation heterogeneity tended to produce an increase in the apparent number of lineage-specific genes. The magnitude of this effect was smaller and more inconsistent than for the two kinds of annotation heterogeneity considered above. One lineage in mice showed a slight decrease, and typical increases were in the 10s of genes, though one lineage in each group showed a large increase of between 500 and 800 genes. These increases correspond to percent increases ranging between 10% and 500%, with typical values around 30%. Unpatterned annotation heterogeneity therefore also causes increases in the apparent number of lineage-specific genes, which are sometimes, though not as frequently as in other patterns of annotation heterogeneity, substantial.

Could annotation heterogeneity produce observed correlations between apparent gene age and sequence characteristics?

Sequences that appear to be genes specific to lineages of varying evolutionary ages within a phylogeny are commonly inferred to have been born at the base of that lineage. Genes specific to older lineages are therefore considered older in age, and those specific to younger lineages considered younger. It has become common in comparative analyses of lineage-specific genes to classify apparent lineage-specific genes by age in this manner, and to then characterize how genes of different ages differ in various characteristics, like gene length, evolutionary rate, expression level, GC content, or structural content. For example, these analyses have consistently found that the youngest genes are shorter, faster-evolving, more lowly-expressed and more tissue-specific than the oldest genes, with a fairly smooth continuum displayed by age classes in between [22, 48, 52]. Other characteristics, most notably GC content and structural disorder,

have been the subject of more controversy, with different results reported in different taxonomic groups and from different methods [23, 53, 64]. Such analyses have been used to propose, in a Hertzsprung-Russell diagram-like manner, a narrative of how genes change as they age, with accompanying suggestions of the forces dictating their origin and evolution: new genes emerge from the open reading frames present by chance, and therefore likely to be short, lowly-expressed, and poorly-conserved, in intergenic DNA, and gradually come to acquire more "gene-like" properties as they gain biological functions.

Different annotation methods differ in detail, but are similar in their shared reliance on a suite of sequence characteristics to distinguish genic from non-genic sequences. These characteristics are just those that are displayed by sequences for which there is little doubt of their coding status: that is, highly conserved, or "old," genes. Sequences with properties that match those found in highly conserved genes – fairly long, highly and broadly expressed, fairly slow-evolving, with high codon adaptation indices – are therefore more likely to be annotated by any given method. I hypothesized that homologs of sequences higher in these properties are therefore more likely than sequences lower in these properties to be annotated in multiple species, causing sequences with such properties to appear older than those without, even if they are similarly conserved throughout the taxonomy. I further hypothesized that, while this effect should affect analyses in which all species have been annotated with the same method, it may be exacerbated in the case of annotation heterogeneity. This would be true if the probability of different methods annotating homologous sequences is more dependent on these sequence features than is the probability of the same method annotating homologous sequences, which I considered plausible.

In one analysis, annotation heterogeneity changes the relationship between gene age and gene length, exon number, expression level, and GC content

We aimed to test whether annotation heterogeneity does indeed produce or increase the magnitude of the above-described correlations between apparent gene age and gene characteristics compared to the case of uniform annotations.

We first considered the four rodent species described above, using the same datasets and methods to partition genes from the focal species *Mus musculus* into three "gene age" categories: orphan genes, which had no detectable homologs outside of *M. musculus*; lineage-specific genes, which had detectable homologs only in *M. caroli, M. pahari,* and/or *R. norvegicus;* and conserved genes, which had detectable homologs in the earliest-diverging *P. maniculatus*. I then found the lengths, exon numbers, GC contents, expression level, codon adaptation indices [79] and/or hexamer scores [80], and degree of structural disorder [81] for all genes (Methods). I performed this procedure using both the uniform annotation set and the heterogeneous annotation set described above, and asked whether any correlations of these properties with gene age classification were introduced, changed, or exacerbated in the heterogeneous annotations compared to the uniform annotations.

We found that the relationship of three of these quantities with gene age was qualitatively different when the uniform versus heterogeneous annotations were used. First, the widely-observed positive relationships between apparent gene age and both gene length and exon number (themselves clearly correlated quantities) was present but fairly mild with the uniform annotations, and was substantially exacerbated when considering heterogeneous annotations. Second, a negative relationship between age and GC content was present when using uniform annotations, but disappeared when heterogeneous annotations were used. Finally, the widely-

observed positive relationship between age and expression level was absent when uniform annotations were used, appearing only under heterogeneous annotations. The magnitude of these differences between age classes are roughly comparable to those observed in published analyses of lineage-specific genes [18, 23, 51-53]. For other quantities, only mild quantitative differences were present (Table 2).

Changing the annotations used in any comparative genomic analysis should change the its results, and so some degree of quantitative difference between these pairs of annotations is expected. Whether the qualitative differences in trends that I observe here are statistically significant within this analysis, and, more importantly, whether they belie a general effect operating in other studies, is not yet clear. I conclude that some of the reported trends between gene age and various gene characteristics may be due to annotation heterogeneity.

Table 1.2: Mean and median values of several characteristics for genes of different apparent ages using uniform and heterogeneous annotations.

Uniform annotations (mean/median)

	Orphan	Lineage-specific	Conserved
Length (aa)	219/204	269/155	687/493
Exon number	2/52	5.2/3	12.4/9
GC%	59/57	53/53	52/52
% Disordered	44/45	37/37	33/31
Average expression level across tissues (FPKM)	69/23	208/24	137/19
Hexamer score	0.26/0.26	0.27/0.28	0.26/0.27

Heterogeneous (unpatterned) annotations (mean/median)

	Orphan	Lineage-specific	Conserved
Length (aa)	241/214	367/240	689/494
Exon number	2.9/2	6.6/4	12.5/9
GC%	58/58	50/49	52/52
% Disordered	44/49	39/39	33/31
Average expression level across tissues (FPKM)	55/29	97/19	138/18
Hexamer score	0.25/0.25	0.28/0.28	0.26/0.27

Conclusions

Here, I investigated the effects of using taxa in which different species' genomes have been annotated using a mixture of different methods ("annotation heterogeneity") on analyses of genes specific to restricted lineages within those taxa. I considered six case studies of four lineages each, taken from four different species groups. I found that, compared to when all species were annotated with the same method, annotation heterogeneity consistently increased the number of apparent lineage-specific genes.

The number of apparent lineage-specific genes produced by annotation heterogeneity depended to some extent on the particular pattern in which the different annotation methods were arranged on the species topology, increasing as the annotation methods used within the lineage became more distinct overall from those used in outgroups. Effect sizes ranged from increases of tens to several thousands of genes, corresponding to percent increases of between 10% and 6000%. In many cases, a large majority of the total genes that appeared lineage-specific in the heterogeneous annotation analysis were due *only* to annotation heterogeneity, disappearing when uniform annotations were used.

Annotation heterogeneity is common in studies of lineage-specific genes. My results suggest that the numbers of lineage-specific genes found in such studies are likely inflated, potentially significantly, by the effects I find here.

Recent work, including that described in Chapter 2, has shown that homology detection failure, in which homology searches fail to detect homologs that are actually present in outgroups, can also produce spurious lineage-specific genes [82, 83]. Annotation heterogeneity joins homology detection failure as a second such source. Unlike homology detection failure,

whose incidence increases with evolutionary time, I show here that annotation heterogeneity can cause large numbers of spurious genes appearing specific to very young lineages (<60 million years old). This may explain the surprisingly large number of young lineage-specific genes found in many studies; for example, one analysis found, in two taxa, a peak in lineage-specific genes that emerged within the last ~50 million years that is larger than at any other point in the taxa's evolutionary history [34].

Annotation heterogeneity is common not just in studies of lineage-specific genes, but in comparative genomic studies at large [6, 84-88]. It may have effects on other types of analyses; for example, it could cause the false appearance of lineage-specific gene loss.

We find some evidence that annotation heterogeneity can alter the qualitative trends between the apparent age of lineage-specific genes and features like their length, expression level, and GC content. Such trends have been used to draw inferences how and from what new genes are born, and how they change over time [23, 53][1]. If such trends were shown to be entirely or in part a technical artifact of annotation heterogeneity, such evolutionary inferences would be called into questions. The data presented here is not at all conclusive on this point. I demonstrate only, through one example, that such an effect is possible, and leave open whether it is a general or sizable enough phenomenon to have such consequences. Though I do not address this explicitly here, I also propose the related concept of differences in annotations *between* studies may account for the conflicting reports of correlation of particular gene properties like GC content and disorder with gene age [23, 53, 89].

What is the true number of lineage-specific genes? My results do not answer this question: I find only that annotation heterogeneity inflates results compared to uniform

[1]

annotations. These annotations may themselves still overestimate the number of lineage-specific genes. Conversely, they may underestimate the number of lineage-specific genes. That most annotation methods omit some real lineage-specific genes seems likely, as they generally rely on features, like homology to known genes, and sequence properties like length, expression level, and codon optimization, that are likely weaker in novel genes. The relative magnitude of these effects, and so the true total number of lineage-specific genes, remains unknown.

Importantly, studies affected more or less by these different types of errors will disagree about not just *how many* lineage-specific genes there are, but also *which* they are. When annotation methods disagree, which is the most accurate? I do not attempt to address this question: my results only illustrate a consequence of these disagreements. I consider annotation accuracy a question primarily accountable to experimental data. Whether a sequence is transcribed, translated, and effects a biologically important function in a given species is of ultimate importance in determining whether it is a gene in that species. In light of these results, I suggest an increased emphasis on these metrics in studies of lineage-specific genes. Some recent papers have opted to avoid reliance on existing annotations entirely, instead using ribosome profiling occupancy to define in the set of analyzed sequences uniformly in all organisms under consideration [90-92]. While ribosome occupancy is not a not perfect proxy for function, it is closer than purely computational annotations or annotations guided only by transcriptomic data, and using a single method to annotate all genomes under consideration avoids the risk of artifacts described here. This type of approach seems to us like a promising way forward.

Methods

Identifying lineage-specific genes

For each species group, I defined a gene as specific to a particular lineage if a search using BLASTP [69] version 6.2.0 had no similar gene at a significance threshold of E=0.001 in any species in each group that was an outgroup to that lineage. I did not require that a gene be present in all members of the lineage to be specific to that lineage: a gene was defined as specific to a lineage based on the most distant species in which it was detected. For example, if a gene in *M. musculus* was detected only in *R. norvegicus*, it was defined as specific to that lineage; if a gene in *M. musculus* was detected in *M. caroli, M. pahari, and R. norvegicus,* it was also defined as specific to that same lineage. If a gene was found in the oldest member of the species group, it was considered "conserved" and so not counted as any kind of lineage-specific gene. This way of classifying lineage-specificity coheres with standard practice [32].

Computation of percent disorder, expression level, and hexamer score for *Mus musculus* genes

For all sequences in the NCBI annotation of *M. musculus* used here, I computed the hexamer score using version 1.2.2. of CPAT [80], run with default settings. I used model files prebuilt for *M. musculus* (mus_hexamer and mus_logit version 1.2.2, downloaded from https://sourceforge.net/projects/rna-cpat/files/v1.2.2/prebuilt_model/).

We computed the percent disorder of these sequences using IUPred [81] version 2a, downloaded from https://iupred2a.elte.hu/, using "long" mode.

We computed the expression level of these sequences by calculating the average FPKM in one replicate of several tissues available from a published *M. musculus* RNA-seq tissue atlas [93]. The datasets used here were downloaded from the NCBI SRA as ERR2130614, ERR2130626, ERR2130638, ERR2130640, ERR2130648, and ERR21306. I aligned reads from

these datasets to the M. musculus genome assembly underlying the annotation used here (GCF_000001635.26_GRCm38.p6_genomic.fna) using Bowtie version 1.1.1 [94] and calculated FPKM for each sequence using FeatureCounts [95] as available in SubRead version 1.5.1 [96].

Chapter 2: The contribution of homology detection errors to analyses of lineage-specific genes

This chapter is based heavily on a published manuscript [82], but is here somewhat elaborated with additional explanation, discussion, and analyses.

Introduction

As discussed in the introductory chapter, the prevailing common interpretation of lineage-specific genes is that they are in some way meaningfully biologically "novel." In the previous chapter, I explored one possible alternative explanation, having nothing to do with novelty, for what these genes are: that they are merely technical artifacts resulting from different annotation methods being used for different species in the analysis.

Another explanation for a lineage-specific gene having nothing to do with novelty is that homologs of the gene *do* exist outside of the lineage, but that computational similarity searches (e.g. BLAST) have merely failed to detect those homologs. I refer to this scenario as homology detection failure. Why would this occur? As homologs diverge in sequence from one another, the statistical significance of their similarity declines. Over evolutionary time, with a constant rate of sequence evolution, the degree of similarity is expected to eventually fall below the chosen significance threshold, resulting in a failure to detect the homolog. Some lineage-specific genes may just be those for which this happens to have occurred relatively quickly, even in the absence of any novelty-generating evolutionary mechanisms.

The possibility of homology detection failure has long been recognized [4, 97-102], but the key questions of how many and which lineage-specific genes are best explained by it remains unclear. Previous work has aimed to explicitly simulate the evolution of each gene to predict

whether or not its homologs would be detectable at a given evolutionary distance [48, 98, 99, 103-106]. These approaches depend on the choice of an evolutionary model and a method for fitting the large number of parameters within that model; results have proven sensitive to these choices, varying widely within the same taxon [48, 98, 99, 103-105, 107]. In the absence of clear data benchmarking the accuracy of different methods that take this approach, significant controversy remains about the motivating question of how many, and which, lineage-specific genes could be due to homology detection failure [48, 98, 99, 103-105, 107]. In addition, the parameter richness and computational complexity of these approaches makes them practically difficult to apply to new genes in new taxa. A tool to easily determine whether any given lineage-specific gene in any given taxon could be due to homology detection failure would be useful, allowing researchers to focus further characterization – for example, functional experiments – on the genes likeliest to be novel, minimizing the possibility of results based on technical artifact.

Here, I further explore this question: for how many, and which, lineage-specific genes is homology detection failure a viable explanation? I present a method for evaluating whether homology detection failure is sufficient to account for a given lineage-specific gene. This method is based on a simple two-parameter mathematical model that estimates the probability that a homolog would be detected at a specified evolutionary distance if it were evolving at a constant rate under standard, novelty-free evolutionary processes. I apply this method to lineage-specific genes in insects and yeasts, assessing for how many such genes in these taxa homology detection failure is a sufficient explanation. I make a preliminary assessment of features of the genes that reject the explanation of detection failure by this test, making them stronger candidates for

potential novelty. Finally, I make this method, which I call abSENSE, available for studies of lineage-specific genes in arbitrary taxa, and describe ongoing efforts to increase its ease of use.

Mathematical formulation of a null model of homolog detectability decline as a function of evolutionary distance

We developed a formal test of the null hypothesis that homology detection failure is sufficient to explain the lineage-specificity of a gene. Specifically, I model the scenario in which the gene actually existed in the common ancestor of all species, evolved at a constant rate, and has homologs outside the clade in which its homologs are detected that appear to be absent solely due to homology detection failure. This is an evolutionary null model: it invokes no processes beyond the simple and novelty-free scenario of orthologs diverging from a common ancestor and evolving at a constant rate.

Because of its use in previous work on lineage-specific genes and in sequence analysis more broadly, I use BLASTP as the search program used to detect homologs here. In search programs like BLAST, sequence similarity is used to infer homology between two genes. Such programs report a similarity score (referred to as “bitscore” by BLAST) between a pair of sequences, as well as the number of sequences that would be expected to achieve that similarity score by chance (an E value). When this number falls below a significance threshold (e.g. $E < 0.001$), statistically significant similarity is interpreted as evidence that the two genes are homologous. The similarity score therefore directly determines whether a homolog is successfully detected in a search.

The key idea in my method is to predict how the similarity score between two homologs evolving according to my null model is expected to decline as a function of the evolutionary

distance between them. I can then ask whether a given gene's lack of detectable homologs outside of the lineage is expected under this null evolutionary model. Previous work on this problem has simulated the evolution of each gene [48, 98, 99, 105]. In preliminary work, I explored similar ideas. Such an approach requires selection of many evolutionary parameters, which has led to questions about the sensitivity of results to these details [48, 99, 103, 104, 107]. my preliminary work recapitulated this sensitivity. In addition, I found the parameter inference and simulation procedures involved in these approaches to be unwieldy to perform. I wondered if a different approach might be both more robust to parameter uncertainty and simpler in practice.

We chose instead to analytically model how the similarity score between two homologs decays with the evolutionary distance between them. my model assumes that a similarity score S between two homologs [69] is proportional to the number of amino acid sites that are identical in these two homologs. After the homologs diverge from their common ancestor, I assume that they undergo a substitution-only mutation process, in which each site in the proteins mutates into a non-identical site at the same protein-specific rate R per unit of evolutionary time. Neglecting the possibility of reversion, the probability that a site will *not* undergo a mutation within a time t following homolog divergence is e^{-Rt}, as given by the Poisson distribution. Given a constant number N_0 of total sites in the protein, the number of sites that remain identical at that time t is binomially distributed with mean $N_0 e^{-Rt}$. If each identical site contributes the same amount c to the total similarity score, then the mean similarity score at time t is $S(t) = cN_0e^{-Rt}$. The variance of the similarity score at time t is $\sigma^2 = cN_0(1-e^{-Rt})(e^{-Rt})$, from the variance of a binomial distribution. I refer to cN_0 as L, as these two parameters only appear as a product. This model makes many simplifying approximations. It abstracts away effects of the substitution score matrix, insertion and deletion scores, and local versus global sequence alignment. It also

approximates the evolution of all sites in a protein as evolving identically, at the same protein-specific rate *R.*

With this model, I can predict similarity scores for a given gene if I have three inputs: a gene from a chosen focal species, the similarity scores of successfully identified homologs of the gene in a few other species *(S),* and the evolutionary distances between the focal species and these other species *(t).* As described in the following section and methods, I precalculate these evolutionary distances *t* from an aggregate of many genes from the set of species under consideration, and therefore they do not depend on the particular gene under consideration. I use these inputs to find the gene-specific values of the parameters *L* and *R* that produce the best fit to my equation describing how similarity scores decline within the species where homologs were detected. I then use these parameters to extrapolate and predict the expected similarity score of hypothetical homologs of the gene at evolutionary distances beyond those of the species whose homologs were used in the parameter fitting. Given an E-value threshold, this predicted similarity score, and the expected variance of the similarity score, I can estimate the probability that a homolog will be undetected at these longer evolutionary distances. In the analyses that follow, I use a relatively permissive E-value threshold of 0.001.

Null model adequately describes the decay of ortholog detectability with evolutionary distance

We applied my model to genes of the yeast *Saccharomyces cerevisiae* and the fly *Drosophila melanogaster* and their orthologs in several fungal and insect outgroups respectively. I focus on the fungi and insects because their genomes are well-annotated, they have closely related and well-annotated sister species, and they have been the focus of previous work on

lineage-specific genes [21, 23, 36, 38, 51, 108]. For *S. cerevisiae, I* included 11 fungal species spanning a divergence time of ~600 million years [109]; for *D. melanogaster*, I included 21 insect species spanning a divergence time of ~400 million years [110]. These species are listed in Figure 2.1.

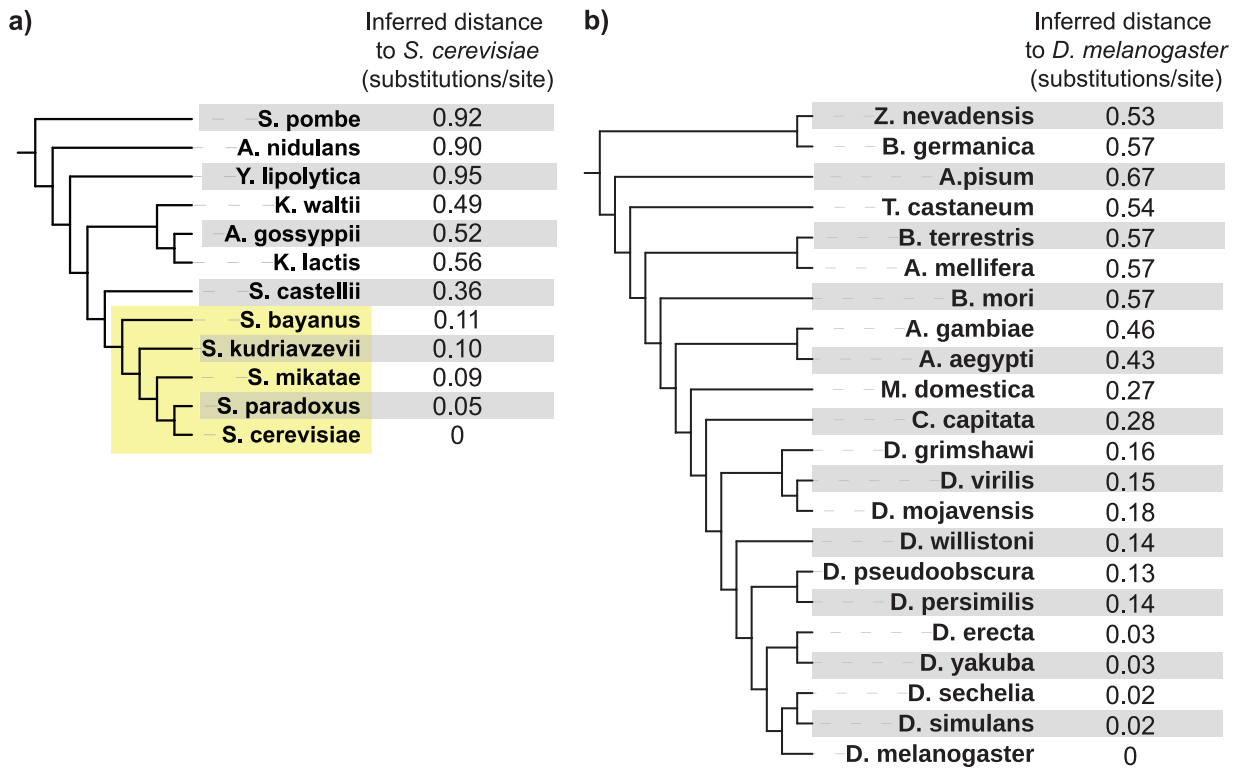

Figure 2.1: Inferred evolutionary distances between each fungal species and *S. cerevisiae* (a) and each insect species and *D. melanogaster* (b). The tree topologies for these taxa are based on previously published studies [110, 111] and were not calculated here; branch lengths are not to scale. The fungal *sensu stricto* lineage, referenced frequently in the text, is shaded in yellow.

Before using my null model to ask whether it explains the lack of detected homologs of lineage-specific genes, I confirmed that it is a good approximation of how similarity scores decay with evolutionary distance. To do this, I tested how well the model represents the decay of similarity scores of general *S. cerevisiae* and *D. melanogaster* genes in increasingly distant species. If the model fits this decay well for most genes, it is likely a good representation of the minimal evolutionary process in the null hypothesis, and can therefore detect deviations from that process.

To obtain evolutionary distances from the focal species (*t* values), I used 102 genes from the Benchmarking Universal Single Copy Ortholog ("BUSCO") [112] database to calculate evolutionary distances in substitutions/site between *S. cerevisiae* and each of the 11 other fungi, and 125 BUSCO genes to calculate evolutionary distances between *D.* melanogaster and each of the 21 other insects (Methods). In both taxa, to show that distances can be reliably computed using a small number of genes, I also re-calculated these distances using two random subsets of 15 BUSCO genes. Distances computed from these different gene sets were similar (Table 1).

Table 2.1: Computed evolutionary distances for the two lineages considered here.

Species	Inferred distance to S. cerevisiae (substitutions/site), 102 BUSCOs	Inferred distance to S. cerevisiae (substitutions/site), 15 BUSCOs subset 1	Inferred distance to S. cerevisiae (substitutions/site), 15 BUSCO subset 2
S. cerevisiae	0	0	0
S. paradoxus	0.05	0.05	0.05
S. mikatae	0.09	0.09	0.09
S. kudriavzevii	0.1	0.11	0.1
S. bayanus	0.11	0.12	0.11
S. castellii	0.36	0.39	0.44
K. waltii	0.49	0.49	0.51
A. gossypii	0.52	0.52	0.54
K. lactis	0.52	0.56	0.58
A. nidulans	1.02	0.9	0.99
S. pombe	1	0.92	0.95
Y. lipolytica	0.9	0.95	0.89

Table 2.1

(Continued)

Species	Inferred distance to D. melanogaster (substitutions/site), 125 BUSCOs	Inferred distance to D. melanogaster (substitutions/site), 15 BUSCO subset 1	Inferred distance to D. melanogaster (substitutions/site), 15 BUSCO subset 2
D. melanogaster	0	0	0
D. simulans	0.02	0.02	0.02
D. sechelia	0.02	0.02	0.02
D. erecta	0.04	0.03	0.04
D. yakuba	0.04	0.03	0.04
D. pseudoobscura	0.14	0.13	0.11
D. persimilis	0.14	0.14	0.12
D. willistoni	0.16	0.14	0.16
D. virilis	0.17	0.15	0.15
D. grimshawi	0.18	0.16	0.16
D. mojavensis	0.18	0.18	0.17
M. domestica	0.3	0.27	0.29
C. capitata	0.31	0.28	0.33

Table 2.1 (Continued)

A. aegypti	0.48	0.43	0.48
A. gambiae	0.49	0.46	0.5
Z. nevadensis	0.58	0.53	0.6
T. castaneum	0.58	0.54	0.58
A. mellifera	0.62	0.57	0.62
B. germanica	0.62	0.57	0.65
B. terrestris	0.62	0.57	0.63
B. mori	0.62	0.57	0.6
A. pisum	0.69	0.67	0.67

Figure 2.1 shows evolutionary distances inferred from one of the 15 gene sets between the focal organism *S. cerevisiae* and the 11 other fungi, and between the focal organism *D. melanogaster* and the 21 other insects. For reference, Figure 2.1 depicts these distances along with a topology taken from previous phylogenetic studies of these taxa [110, 111]; branch lengths are not to scale. I use these distances, computed from one of the 15 gene subsets, in all results presented in the main text below. I opted to use distances from 15-gene subsets instead of from the full set of 100+ BUSCO genes to demonstrate that the method I describe here produces reliable results using only a small number of genes, which lowers the barrier to use for other researchers seeking to apply the method to different taxa.

We next took all annotated *S. cerevisiae* and *D. melanogaster* proteins (Supplemental Table 2) and calculated the similarity scores of their detectable orthologs in each of the 11 other fungal and 21 other insect outgroup species respectively. (For *S. cerevisiae* and *D. melanogaster,* the score is the comparison of the protein with itself.) I identified orthologs using reciprocal best BLASTP search with a threshold of $E < 0.001$ (Methods). Reciprocal best BLASTP is not a perfect means of distinguishing orthologs from paralogs, and results in some genes failing to be assigned to orthologs in some species, but, based on my results below, suffices for the purpose, and is easy to do at scale.

With these similarity scores (S) and evolutionary distances (t) in hand, I tested how well my model explains the observed decline in similarity scores with increasing evolutionary distance in fungal and insect genes. Fits to the model for three example genes from each taxon are illustrated in Figure 2.2.

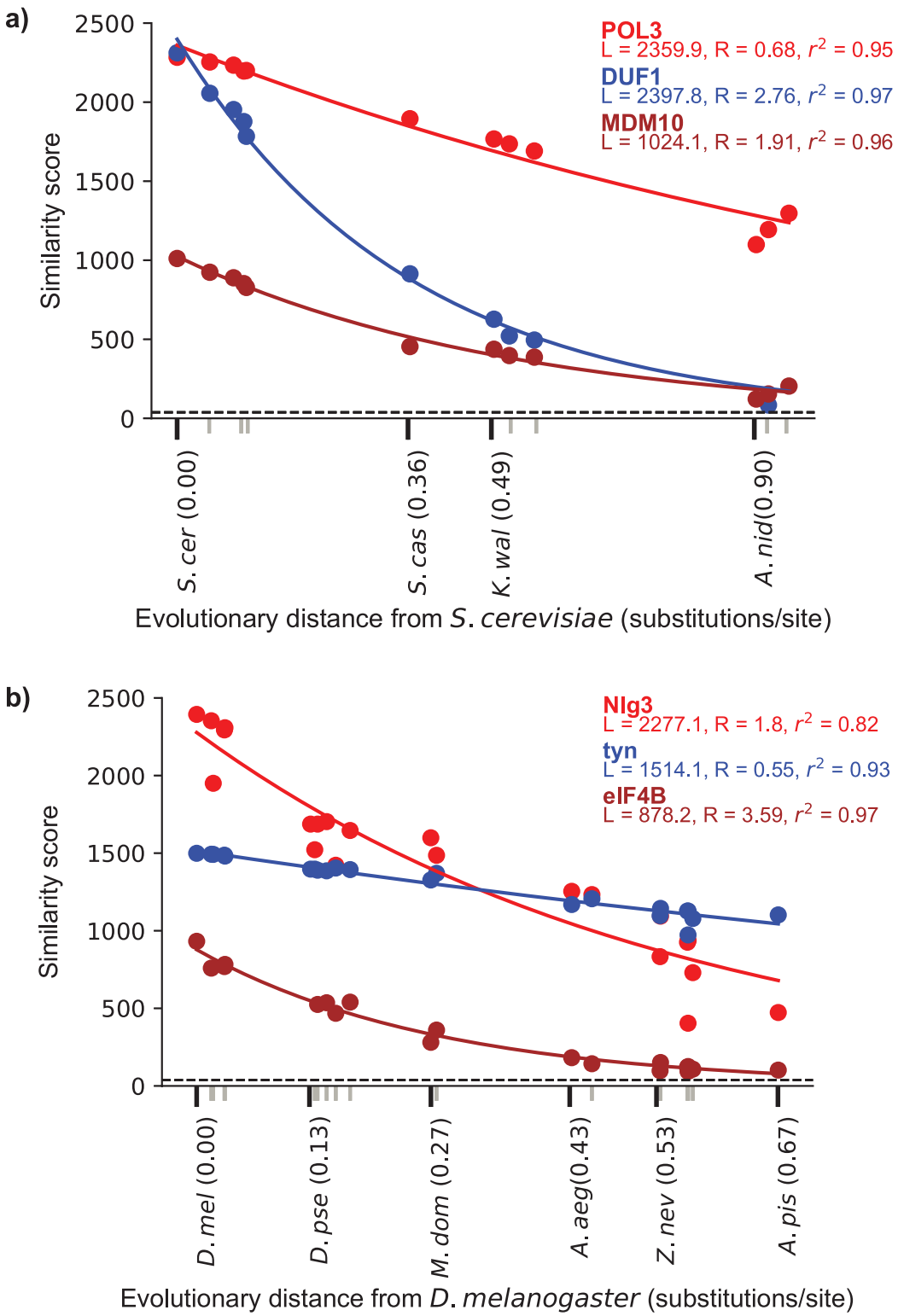

Figure 2.2: Depictions of the fit of the null model of similarity score, as defined in the text, decline with evolutionary distance for three representative proteins from *S. cerevisiae* (a) and *D.*

(Continued) *melanogaster* (b). Colored points represent the BLASTP score between the protein and its ortholog in the species that is at the evolutionary distance indicated on the x axis. Tick marks on the x axis represent each of the species used here. For visual clarity, only some species names and evolutionary distances are included, indicated with black tick marks; gray tick marks represent the other unlabeled species. The dashed line represents the detectability threshold, the score below which an ortholog would be undetected at my chosen E-value of 0.001. The best fit values of L and R are shown for each protein. The r^2 value is also shown and was calculated from a linear regression of the log of the similarity score versus evolutionary distance.

We aimed to assess the fit of my model systematically. my model predicts a linear relationship between the log of ortholog similarity scores and evolutionary distance. For all proteins in both clades, I therefore performed a linear regression of the log of each protein's similarity score, ln $S(t)$, against the inferred evolutionary distance to the focal species, t, and computing the square of the Pearson correlation coefficient (r^2), which measures how much of the variance in ln $S(t)$ is explained by t.

The model predicts similarity scores reasonably well. The mean and median r^2 were 0.92 and 0.95 for similarity scores of *S. cerevisiae* genes. I repeated this with *D. melanogaster* proteins and their orthologs in the other insects, where the mean and median r^2 were 0.84 and 0.91 for similarity scores of *D. melanogaster* genes. Results were similar using the two other sets of estimated distances (Figure 2.3), indicating their insensitivity to minor differences in the choice of gene sets used.

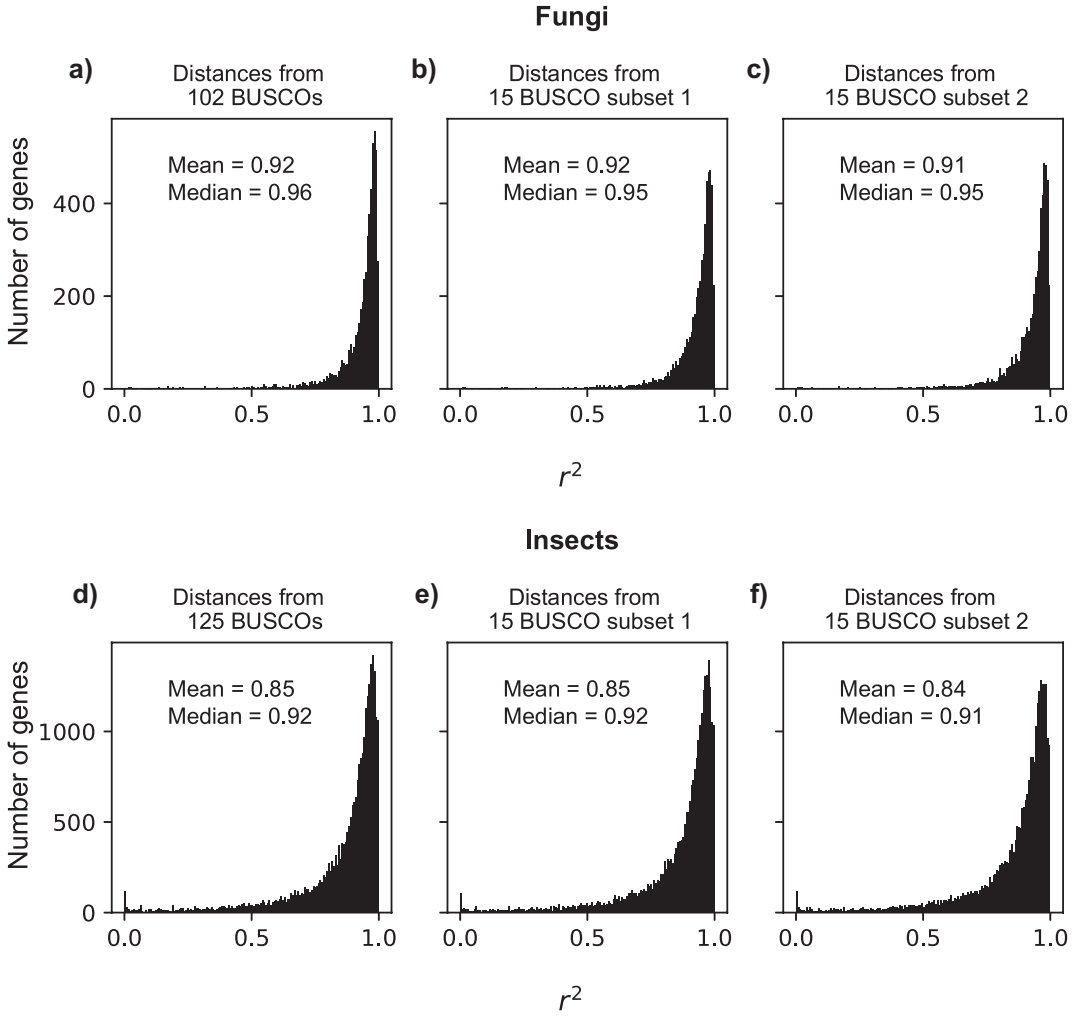

Figure 2.3: r^2 distributions for the fit to the model of *S. cerevisiae* and *D. melanogaster* genes using evolutionary distances derived from 3 sets of genes**.** **a**: *S. cerevisiae* genes with distances derived from 102 BUSCOs. **b**: *S. cerevisiae* genes with distances derived from a randomly selected subset of 15 of the BUSCOs used in a. **c**: *S. cerevisiae* genes with distances derived from a second randomly selected subset of 15 of the BUSCOs used in a. **d**: *D. melanogaster* genes with distances derived from 125 BUSCOs. **e**: *D. melanogaster* genes with distances derived from a

(Continued) randomly selected subset of 15 of the BUSCOs used in d. **f**: *D. melanogaster* genes with distances derived from a second randomly selected subset of 15 of the BUSCOs used in d. In d-f, the peak near $r^2 = 0$ is comprised of genes with orthologs identifiable only in a subset of the closely related *Drosophilid* flies, such that their sequences are identical or nearly identical in all species, except 1 or 2 in which a large chunk of the *melanogaster* protein is absent from the annotation, resulting in almost none of the variance in score (of which there is none, save this large event) being explained by divergence time. I consider this an artifact of the method, as it only appears in the limited cases where the sequences in question are almost totally identical.

As well as considering the fit of each gene to the expected value of the model, I tested how well my estimate for the variance of the similarity score captured the observed scatter around this expected value. To do this, for the ortholog of each *S. cerevisiae* gene in each species, I calculated the difference between the actual and expected similarity score and expressed it as a multiple of the predicted standard deviation $\sigma = \sqrt{L}(1-e^{-Rt})(e^{-Rt})$ of the similarity score (a Z score). I expect these Z-scores to follow a normal distribution if my model's estimated variance is correct, which is roughly what I observe. Approximately 92% of *S. cerevisiae* orthologs have observed scores within 3 s.d. of the prediction; for a standard normal distribution, 99% are expected. 7% of scores are below three s.d., and 1% are above three s.d. Results in *D. melanogaster* are similar: 88% have observed scores within 3 s.d., 8% below, and 4% above. I attribute the skew toward predicted scores that are higher than observed scores to the fact that my model neglects how insertions and deletions may disrupt the length of a local alignment. Results were similar when using the two other sets of estimated distances (Figure 2.4).

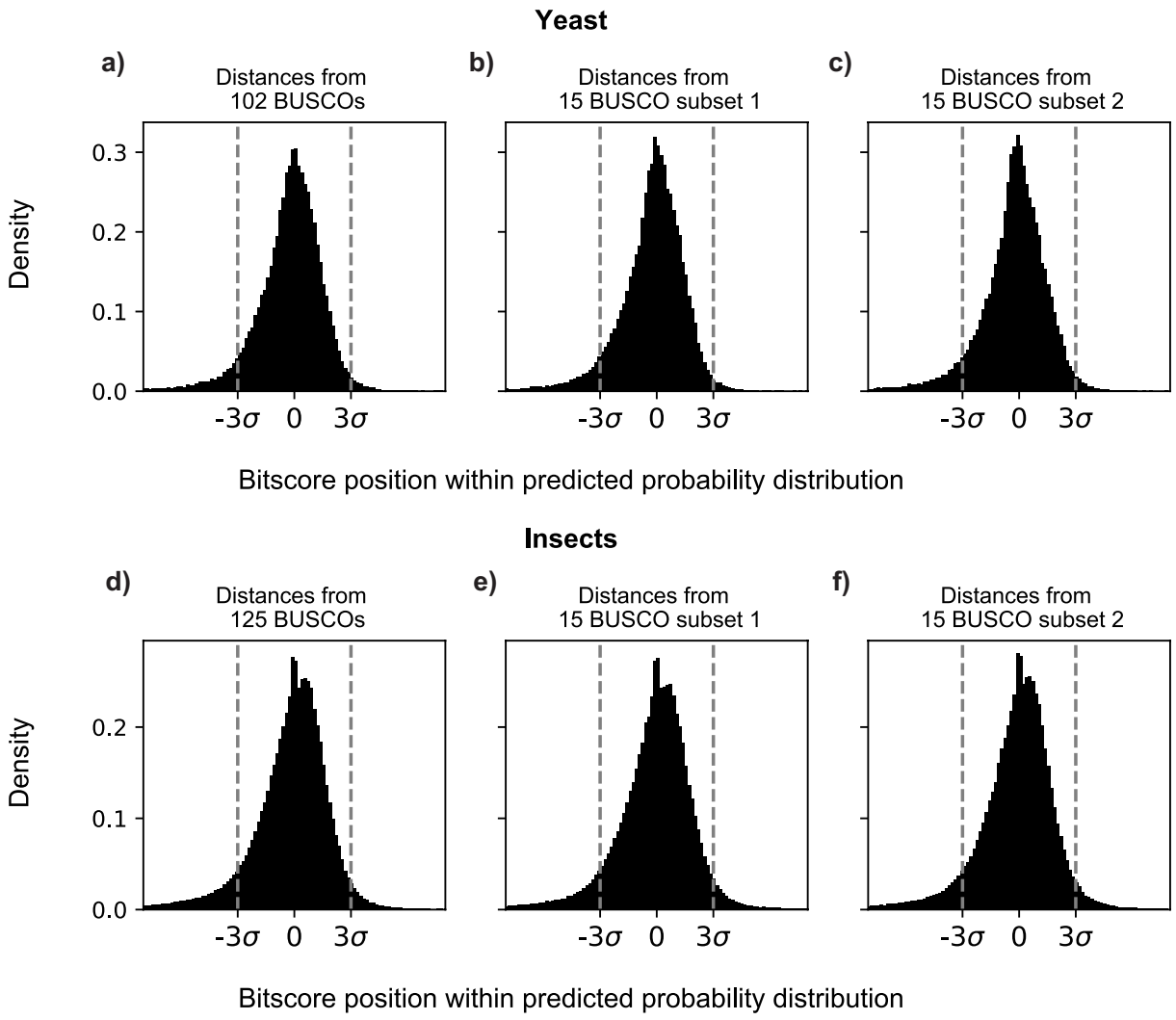

Figure 2.4: Distribution of position of BLASTP scores between *S. cerevisiae* and outgroup yeast (top) and *D. melanogaster* and outgroup insects (bottom) relative to the predicted confidence interval. 0 indicates that the score has the same value as the best fit to the model; multiples of sigma indicate that the score is that many standard deviations above or below the best-fit value. **a**: *S. cerevisiae* genes with distances derived from 102 BUSCOs. **b**: *S. cerevisiae* genes with distances derived from a randomly selected subset of 15 of the BUSCOs used in a. **c**: *S. cerevisiae* genes with distances derived from a second randomly selected subset of 15 of

(Continued) the BUSCOs used in a. **d**: *D. melanogaster* genes with distances derived from 125 BUSCOs. **e**: *D. melanogaster* genes with distances derived from a randomly selected subset of 15 of the BUSCOs used in d. **f**: *D. melanogaster* genes with distances derived from a second randomly selected subset of 15 of the BUSCOs used in d.

We asked whether the best-fit values of the parameters *L* and *R* found for the fungal proteins are correlated with the interpretation of these parameters in my model. I expect values of *L* to be related to gene length, and values of *R* to be related to evolutionary rate. Using comparisons to *S. cerevisiae* genes, I plotted *L* versus gene length and *R* vs. maximum likelihood estimates of evolutionary distance in substitutions/site in multiple alignments of proteins from *S. cerevisiae* and the four most closely related species. The *L* parameter is indeed highly correlated with gene length ($r^2 = 0.99$), and *R* is more weakly correlated with gene-specific evolutionary rate ($r^2 = 0.47$) (Figure 2.5).

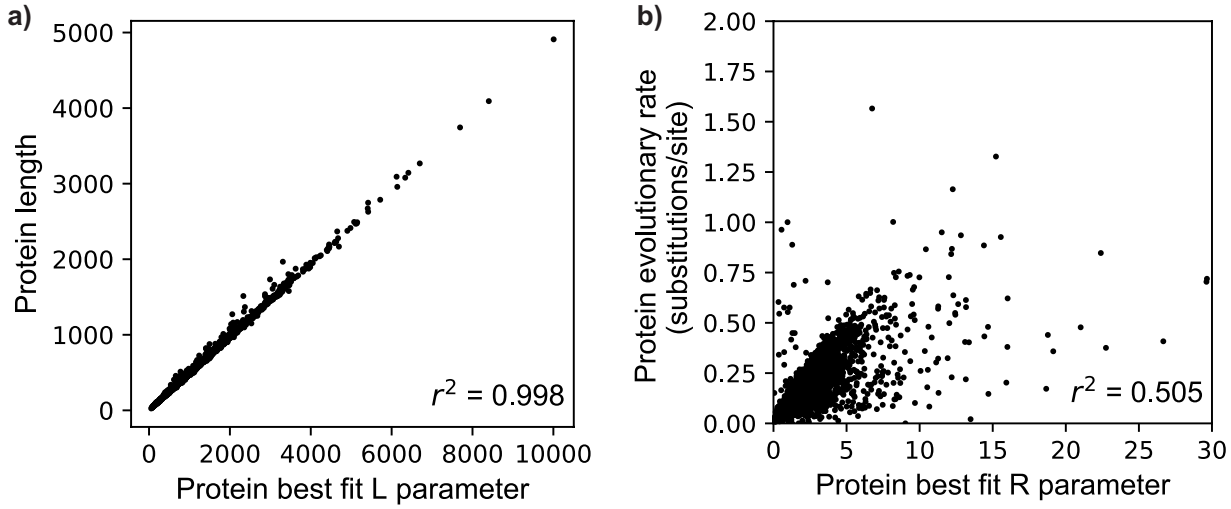

Figure 2.5: Correlation between best-fit parameters and gene properties in yeast. **a**: Correlation between each *S. cerevisiae* protein's best-fit value of *L* and its length in amino acids. The *L* parameter is consistently larger than the length due to most identical alignment positions contributing a score larger than 1 according to the scoring scheme used here (BLOSUM62). **b**: Correlation between each *S. cerevisiae* protein's best-fit value of *R* and its relative evolutionary

(Continued) rate in substitutions per site from alignments made from it and its orthologs in the other four *sensu stricto* yeasts.

We attribute some of this lower correlation to the fact that *R,* which describes how quickly score declines, includes the effects of insertions and deletions as well as substitutions, while standard measures of evolutionary rate derived from alignments (like the gene evolutionary rates I calculated to compare to *R*) only consider substitutions. The distributions of the estimated *L* and *R* parameters across all genes are long-tailed and approximately log-normal (Figure 2.6), consistent with other analyses of distributions of gene length [113] and evolutionary rate [114].

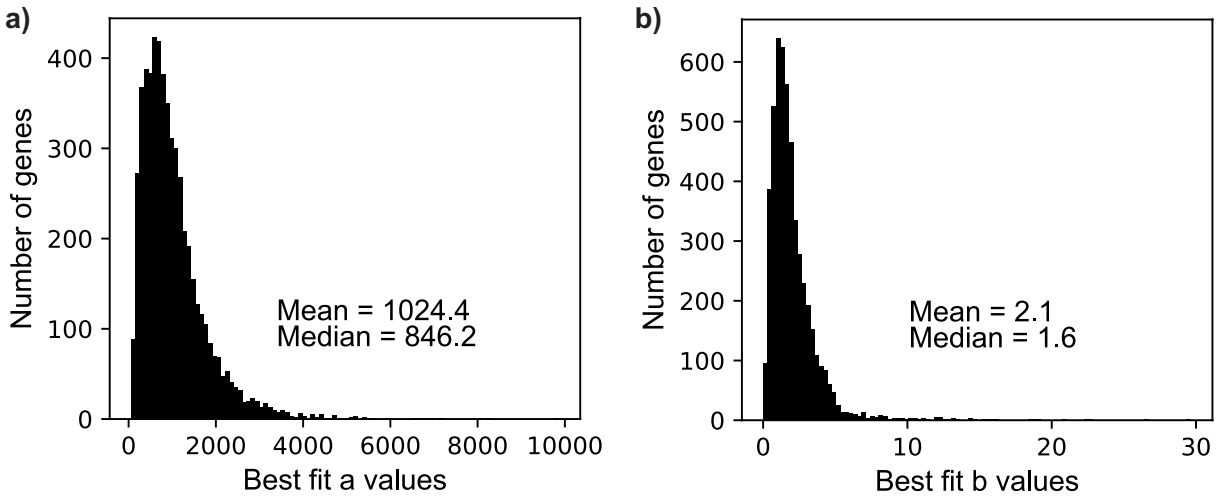

Figure 2.6: Distribution of best-fit parameter values for all *S. cerevisiae* proteins. a: Distribution of the best-fit *a* values for all *S. cerevisiae* proteins. b: Distribution of the best-fit *b* values for all *S. cerevisiae* proteins.

Many lineage-specific genes can be explained by homology detection failure

Having validated my null model for similarity score decline, I then used the model to ask my central question: what proportion of lineage-specific genes is homology detection failure alone sufficient to explain?

We first considered annotated *S. cerevisiae* proteins that are lineage-specific to the *sensu stricto* yeasts, a young lineage sharing a common ancestor ~20 Mya containing the five species *S. cerevisiae, S. paradoxus, S. mikatae, S. bayanus,* and *S. uvarum* (Figure 2.1), which has been the focus of previous work on lineage-specific genes [23, 51]. I identified 375 such *sensu stricto*-specific genes, defined as having homologs detectable by BLASTP in at least one of these species but lacking detectable homologs in the nearest outgroup *S. castellii* or in any other outgroups according to a permissive E-value threshold of 0.001 (Methods). Between 40 and 70%

of *sensu stricto* specific genes identified in two previous studies are included in this set [23, 51]. The remainder are either ORFs not used in my initial search because they are marked as dubious in both the Saccharomyces Genome Database and Refseq and so have been removed from the *S. cerevisiae* Refseq annotation, or because I detected homologs outside of the *sensu strictos*, likely due to my permissive E-value threshold. Since my detectability model is regression-based, a minimum of 3 observed homologs (including the gene in the focal species) are required; for example, I could not perform this computation on the *S. cerevisiae* gene *BSC4* [115], proposed to have a very recent de novo origin and thus only found in *S. cerevisiae*. I applied my model to the 155 such *sensu stricto*-specific proteins.

For each of these 155 lineage-specific genes, *I* used the best-fit values of the *L* and *R* parameters found above to extrapolate and predict the score of an ortholog at the evolutionary distance of *S. castellii* under the null model. Using parameters from the *sensu stricto* lineage to extrapolate to more distant species corresponds to assuming that these two groups of orthologs have evolved in the same manner since their divergence from their common ancestor. Finally, I calculated the probability that a homolog at the evolutionary distance of *S. castellii* would be detected, P(detected | null model, $t_{castellii}$), by using my model for similarity score variance to generate a probability distribution for the score and computing the percentage of the probability mass in this distribution below my chosen detectability threshold (corresponding to an E-value of 0.001).

This analysis is illustrated for one example of a *sensu stricto*-restricted *S. cerevisiae* protein, Uli1, in Figure 2.7.

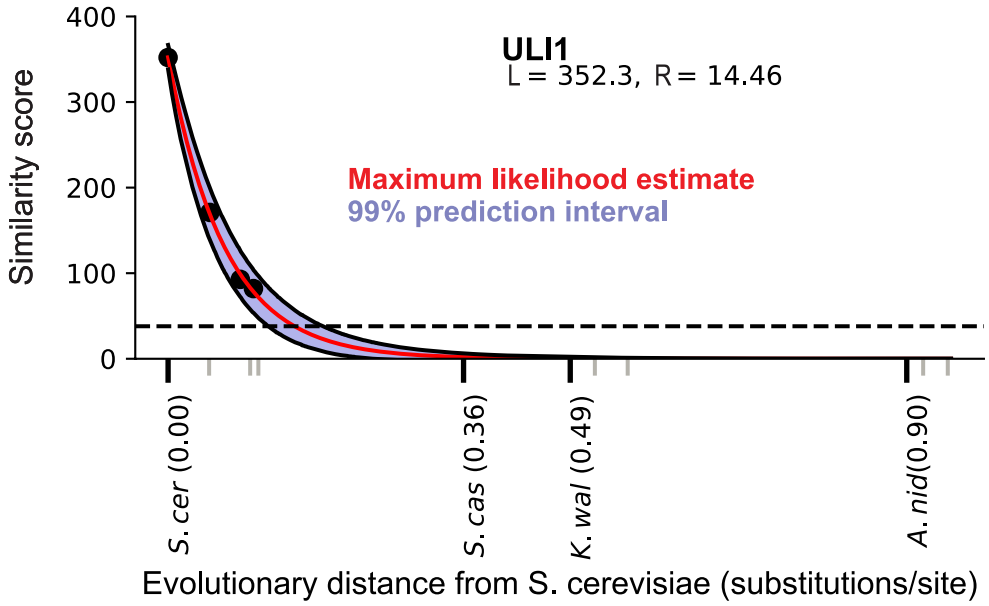

Figure 2.7: Illustration of the prediction of detectability decline for the *S. cerevisiae* protein Uli1, displayed as in Figure 2.2. At the evolutionary distance of the nearest outgroup *S. castellii*, the entire prediction interval lies below the detectability threshold, indicating a ~0% probability that an ortholog would be detected under the null model even if an *S. castellii* ortholog were present.

Uli1 has been implicated in the unfolded protein response [116], making it one of only a few *sensu stricto* specific genes with experimental evidence of function, and its lineage-specificity has prompted previous studies to propose that it originated de novo [23, 51]. However, I find that the probability that an ortholog of this gene would be detectable in *S. castellii*, P(detected | null model, $t_{castellii}$), is approximately 0, indicating that a null evolutionary model is sufficient to explain the lineage specificity of this short and rapidly-evolving gene.

The result of performing this test on all of the 155 *sensu stricto*-specific genes amenable to my analysis is shown in Figure 2.8a, which depicts the distribution of probabilities of detecting a homolog in the outgroup *S. castellii* given the null model and the evolutionary distance between *S. cerevisiae* and *S. castelli*, P(detected | null model, $t_{castellii}$).

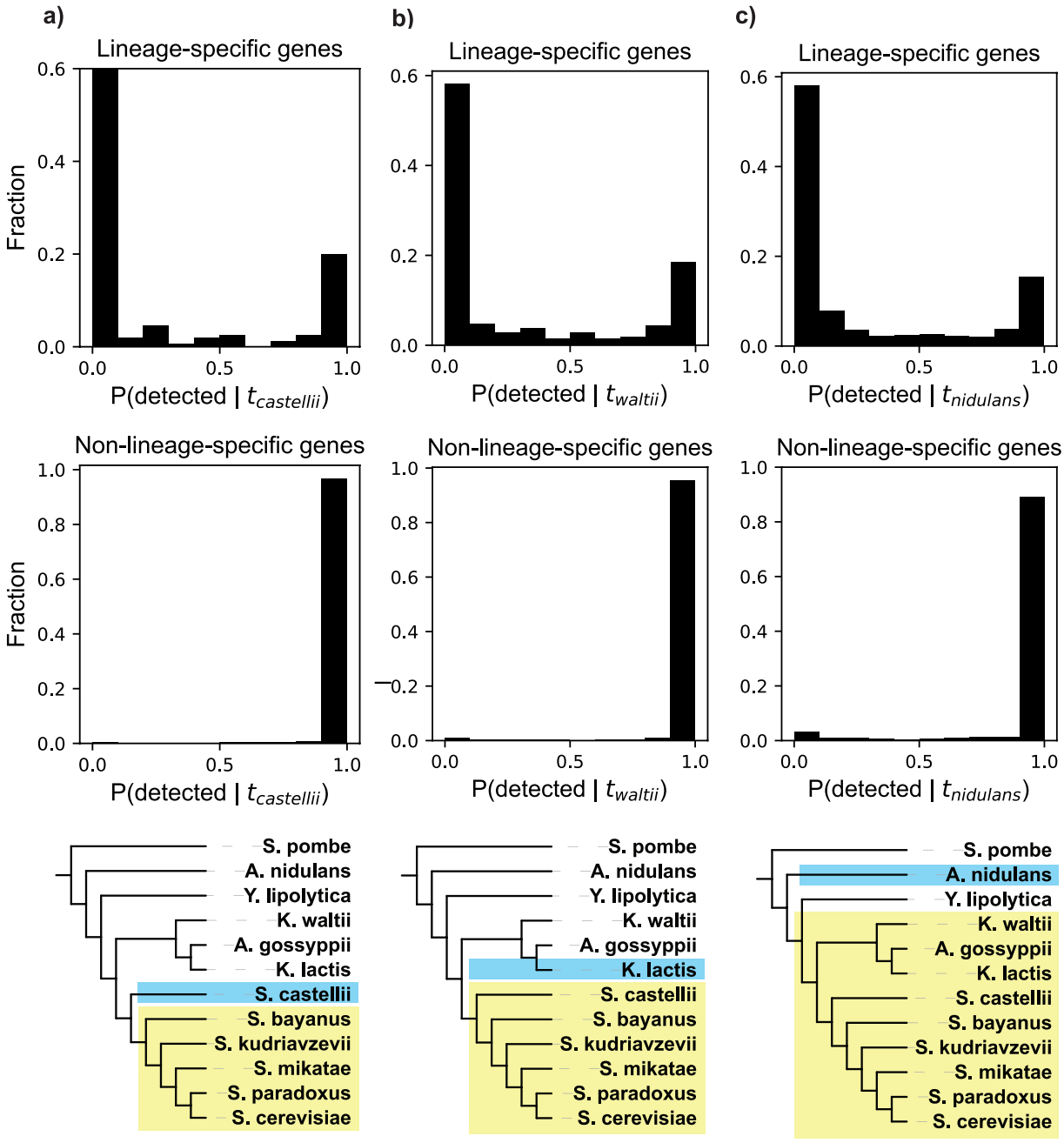

Figure 2.8: Distributions of detectability prediction results for three yeast lineages (a, b, c). Top: results for all lineage specific genes. Middle: results of the same analysis for all non-lineage specific genes, which serve as a positive control. These genes, which have detectable orthologs outside of the lineage, should be predicted to be detected, which they largely are. Bottom:

(Continued) Depiction of the lineage (yellow) and closest outgroup (blue) considered in the analyses in the corresponding column. In c), note that *Y. lipolytica* is the topological outgroup to the shaded lineage, but is not the closest species by evolutionary distance (branch lengths are not to scale).

Many genes have a very high probability of being undetected, and a majority are more likely to be undetected than detected: 55% have P(detected | null model, $t_{castellii}$) below 0.05, and 73% percent below 0.5. This implies that homology detection failure is sufficient to explain a large number, potentially a majority, of these lineage-specific genes. Homologs of these genes only being detected in *sensu stricto* species does not require invoking evolutionary novelty.

We repeated this procedure for *D. melanogaster* genes restricted to the *Drosophila* genus. This young lineage shared a common ancestor ~70 Mya, with the housefly *M. domestica* as the nearest outgroup in my analyses (Fig 9a). I identified 1611 *Drosophila*-restricted genes (Methods), of which 1278 had the two identified orthologs in the *Drosophila* lineage required for my analysis. Again, many of these *Drosophila*-restricted genes are very likely to be undetected: 46% percent have values of P(detected | null model, $t_{domestica}$) below 0.05, and 76% percent are below 0.5. Homology detection failure is therefore also sufficient to explain many lineage-specific genes in this group.

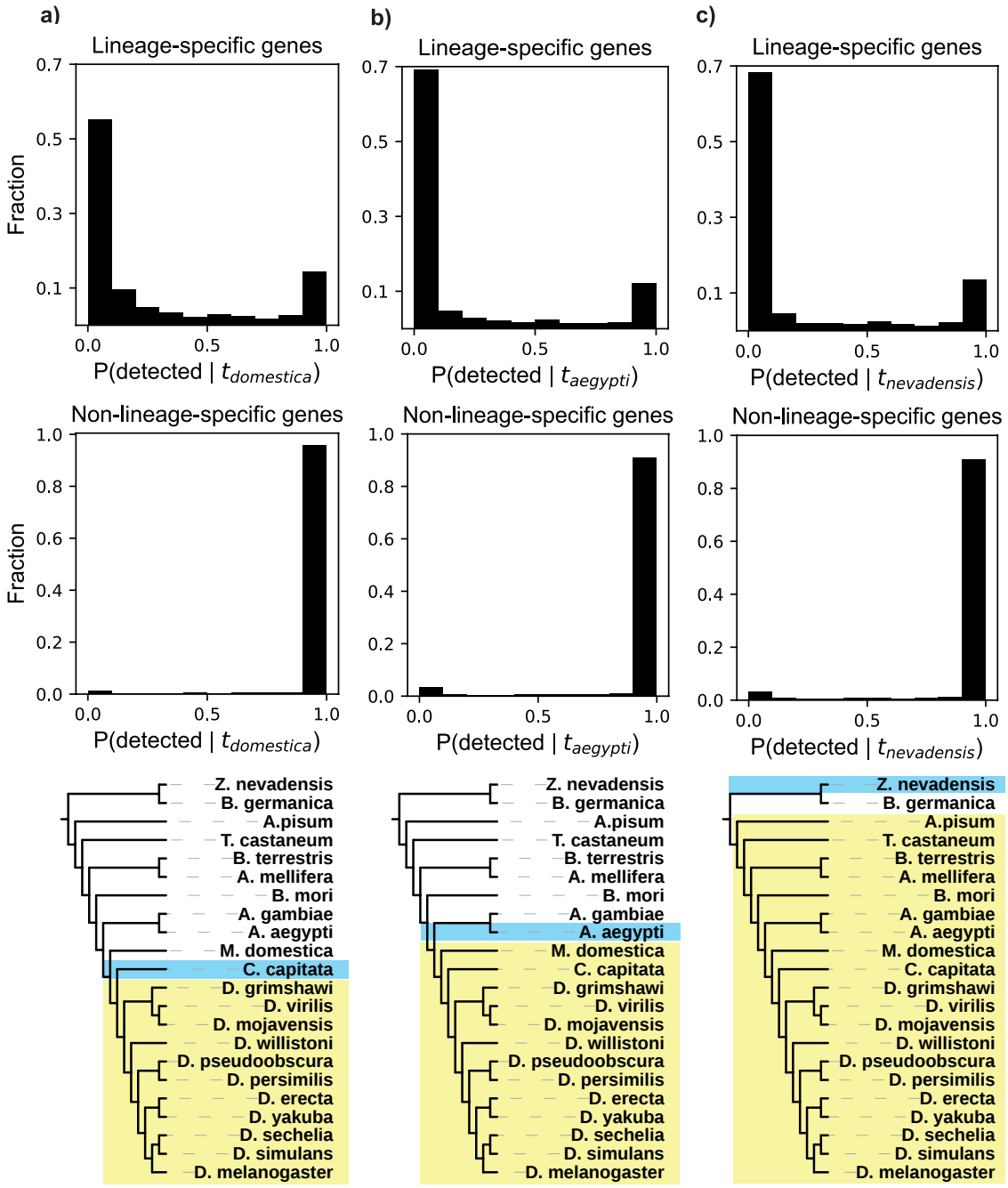

Figure 2.9: Distributions of detectability prediction results for three insect lineages (a, b, c). Top: results for all lineage specific genes. Middle: results of the same analysis for all non-lineage specific genes, which serve as a positive control. These genes, which have detectable orthologs outside of the lineage, should be predicted to be detected, which they largely are. Bottom:

(Continued) Depiction of the lineage (yellow) and closest outgroup (blue) considered in the analyses in the corresponding column. In a), note that *C. capitata* is the topological outgroup to the shaded lineage, but is not the closest species by evolutionary distance (branch lengths are not to scale).

As both the *sensu stricto* yeasts and the drosophilid flies are relatively young lineages, I asked whether these results generalize to older lineages. In fungi, I tested two additional lineages with approximate divergence times of ~70 Mya (Figure 2.8b) and ~250 Mya [109] (Figure 2.8c). In insects, I also tested two additional lineages, with approximate divergence times of ~150 Mya (Figure 9b) and ~350 Mya [110] (Figure 2.9c). I identified all genes specific to each of these four additional lineages, and then calculated P(detected | null model, $t_{outgroup}$) for all genes with the required 2 identified orthologs, exactly as described for the two lineages above. Results in all of these comparisons are very similar to those in the younger lineages tested above: I predict that a large number of lineage-specific genes have very low probabilities of being detected, with a majority more likely to be undetected than detected (Figures 2.8b,c, 2.9b, c). Homology detection failure is thus sufficient to explain a large number of lineage-specific genes in these older lineages as well.

As a control, I asked my model to predict the probability of detecting homologs of genes that are *not* lineage-specific, meaning that these genes have homologs that are detected both inside and outside of the lineage. I repeated the same procedure on all non-lineage-specific genes in the six lineages tested above. As I did for the lineage-restricted genes, I used only similarity scores from orthologs within the given lineage to calculate the probability of detecting homologs in the nearest outgroup to the lineage, P(detected | null model, $t_{outgroup}$). If my model operates correctly, it should predict high values of P(detected | null model, $t_{outgroup}$) for these genes, since their homologs are in fact detected. In accordance with this expectation, my model predicts that the vast majority (>97% in all lineages) of these genes have a very high probability of being detected, P(detected) > 0.95 (Figures 2.8, 2.9).

To assess whether this central analysis was robust to the use of different sets of genes for calculating the underlying evolutionary distances, I computed the correlation between the values of P(detected) for each gene in the three analyses in each lineage for these different gene sets. Results were highly similar across gene sets, showing that this analysis, like earlier analyses, was robust to the use of different sets of genes for calculating evolutionary distances (Table 2).

Table 2.2: Correlation coefficients for gene detectability prediction results based on evolutionary distance estimates derived from the 3 different sets of genes.

S. castellii

	102 BUSCOS	15 BUSCO subset 1	15 BUSCO subset 2
102 BUSCOS			
15 BUSCO subset 1	0.995		
15 BUSCO subset 2	0.923	0.901	

K. waltii

	102 BUSCOS	15 BUSCO subset 1	15 BUSCO subset 2
102 BUSCOS			
15 BUSCO subset 1	0.985		
15 BUSCO subset 2	0.984	0.989	

A. nidulans

	102 BUSCOS	15 BUSCO subset 1	15 BUSCO subset 2
102 BUSCOS			
15 BUSCO subset 1	0.962		
15 BUSCO subset 2	0.929	0.962	

M. domestica

Table 2.2 (Continued)

	125 BUSCOs	15 BUSCO subset 1	15 BUSCO subset 2
125 BUSCOs			
15 BUSCO subset 1	0.994		
15 BUSCO subset 2	0.984	0.971	

A. aegypti

	125 BUSCOs	15 BUSCO subset 1	15 BUSCO subset 2
125 BUSCOs			
15 BUSCO subset 1	0.996		
15 BUSCO subset 2	0.972	0.958	

Z. nevadensis

	125 BUSCOs	15 BUSCO subset 1	15 BUSCO subset 2
125 BUSCOs			
15 BUSCO subset 1	0.992		
15 BUSCO subset 2	0.969	0.958	

We separately considered the set of 784 *S. cerevisiae* genes marked as 'dubious' in the Saccharomyces Genome Database [117], which I excluded from the analyses described above. Although they have been deemed unlikely to encode functional proteins, many of them are lineage-specific and so have been included in previous studies as potentially novel genes [23, 51]. I analyzed the 167 of these dubious genes that met my analysis requirement of having detected orthologs in at least two other species (many are unique to *S. cerevisiae*). I find that homologs of these genes would be undetected at an even higher rate than for validated genes; in all three fungal lineages, at least 99% of these dubious ORFs have P(detected | null model, $t_{outgroup}$) below 0.5, and at least 80% below 0.05.

When controlling for differences in dataset, results are similar to seemingly divergent conclusions from another recent study using an orthogonal method

Another recent paper used a different approach to estimate the fraction of lineage-specific genes that are attributable to homology detection failure [97]. Vakirlis et al. used a small set of genes in "microsyntenic blocks" to count how often a gene is not recognizably similar to its presumptive homolog in the syntenic position in a comparative genome. Assuming that this sample is representative, and so approximates the frequency at which homologous genes in general diverge beyond recognition, they conclude that 20-45% of lineage-specific genes in fungal, insect, and vertebrate phylogenies are attributable to homology detection failure. I consider this result to be in broad qualitative agreement with ours: both values indicate that homology detection failure generates a substantial fraction of lineage-specific genes. However, I do find somewhat larger estimates of the rate of detection failure.

We investigated the cause of this discrepancy. I hypothesized that genes in microsyntenic regions may diverge more quickly than those outside, which would lead to an overestimation of the rate of detection failure when using a dataset comprised exclusively of such genes. Vakirlis et al. did check whether genes in microsyntenic regions had higher values of dN (the inferred number of amino acid substitutions), and so are faster-evolving, than those outside of those regions. They found that this effect was present and statistically significant (Mann-Whitney p-value=$8*10^{-5}$), but hypothesized that the effect size was likely too small (we compute from their data that the median dN of genes in microsyntenic regions was 9% slower than those without) to affect their results.

This analysis was based on dN values computed from a set of previously-produced alignments that include only 5261 *S. cerevisiae* genes [118], compared to the 6002 used here. This difference in number is due to two factors: a different underlying *S. cerevisiae* protein annotation, and the inclusion only of genes for which orthologs were identified in all five of the *sensu stricto* yeast species used here. I speculated that this latter factor in particular may impose a bias against faster-evolving genes, for which orthologs are less likely to be detected in all five species. I also speculated that considering dN alone may not comprehensively enough capture the effect of evolutionary rate on homolog detectability, as insertions and deletions are also known to have a large effect.

As a result, I produced my own alignments of all *S. cerevisiae* genes with an ortholog present in at least *S. bayanus* and used them to perform a similar analysis (Methods). These include 400 *S. cerevisiae* genes that are present in my alignments but absent from the alignments used by Vakirlis et al [118]. I used my alignments to compute substitutions/site between *S. cerevisiae* and *S. bayanus* (Methods). For genes in common between the two datasets, the r^2

between these substitutions/site values and the dN values provided by Vakirlis et al. was 0.75. I found that dN values of genes missing from the alignments used in Vakirlis et al were indeed higher than the genes that were included (Mann-Whitney p-value=$3*10^{-5}$), with a clear tail of very fast-evolving genes unique to the missing genes (Figure 2.10).

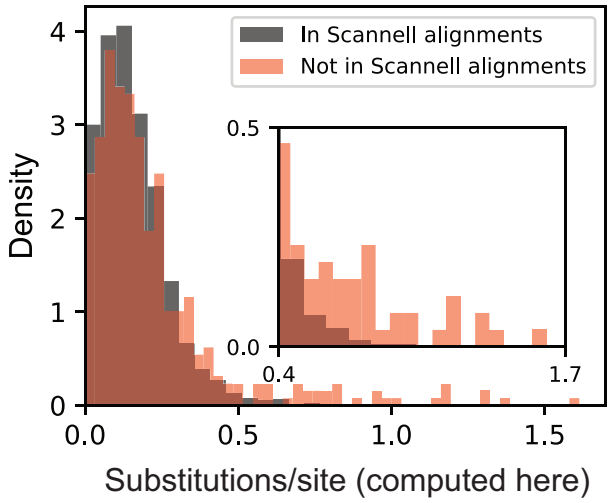

Figure 2.10: dN values for the full dataset of 6002 S. cerevisiae genes used here, colored by their presence or absence from the dataset of Vakirlis et al.

However, this effect is driven by a small number of genes. Accordingly, I find that the overall differences in substitutions/site between genes in syntenic regions and those outside computed from my alignments are similar to those found by Vakirlis et al., excepting the small tail of very fast-evolving genes enriched in the non-microsyntenic distribution not present in their analysis, with a difference in medians of about 9% (Mann-Whitney p-value=$9*10^{-5}$) (Figure 2.11).

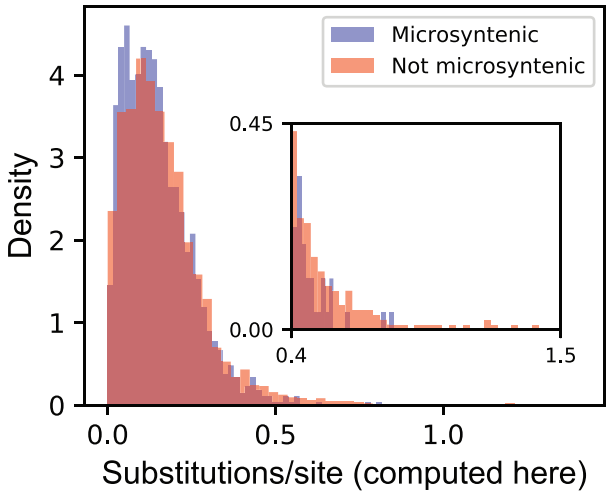

Figure 2.11: dN values for the full dataset of 6002 S. cerevisiae genes used here, colored by their presence or absence from microsyntenic regions according to Vakirlis et al., and so included or not in their analysis.

In addition to point substitutions, homolog detectability over evolutionary time is influenced by the number of insertions and deletions that have occurred; whether this number is comparable between genes within and outside microsyntenic regions was not considered by Vakirlis et al. I find that the number of gaps in my alignments (normalized for number of positions and number of sequences), representing the number of amino acids that have been inserted and deleted, is on average twice as high for genes in microsyntenic regions than in those outside (Mann-Whitney p-value=$3*10^{-24}$) (Figure 2.12).

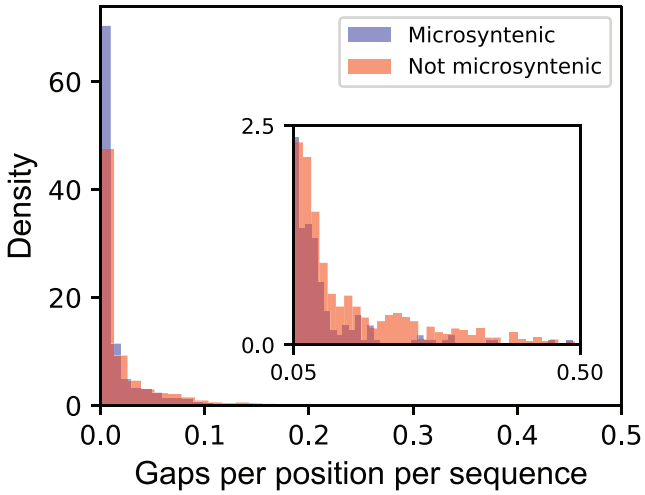

Figure 2.12: Frequency of gaps for the of 6002 *S. cerevisiae* genes used here, colored by their presence or absence from microsyntenic regions according to Vakirlis et al., and so included or not in their analysis.

Consistent with these two observations, genes in microsyntenic regions have higher values of my R parameter, which reflects the overall rate of homology detectability decline and incorporates both substitution rate and insertion/deletion rate (Mann-Whitney p-value = $4*10^{-12}$) (Figure 2.13).

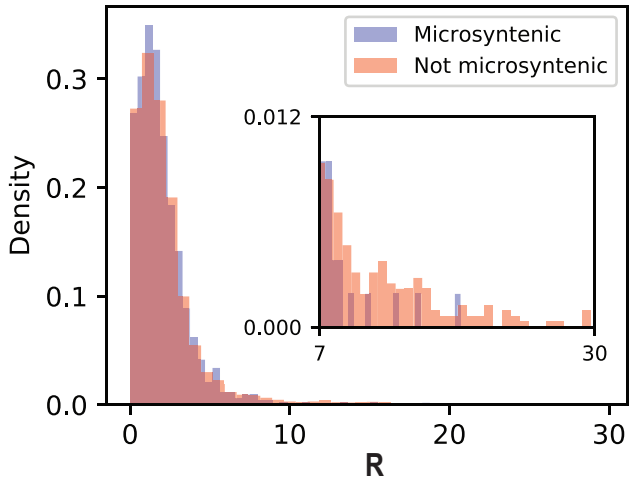

Figure 2.13: Best-fit value of *R* parameter of 6002 *S. cerevisiae* genes used here, colored by their presence or absence from microsyntenic regions according to Vakirlis et al., and so included or not in their analysis.

Based on my analyses, this difference in *R* value seems to be driven largely by a differential rate of insertions and deletions, perhaps with some smaller contribution from differential substitution rate. I note that insertions and deletions are commonly not considered in evolutionary rate calculations, though they have a clear effect on homolog detectability that is at least partially accounted for by my method, which can include their effects on score decline in the best-fit *R* parameter.

Having noted differences in substitution rate and insertion/deletion rate between genes within and outside microsyntenic regions, I asked whether these differences could cause differences in the inferred rate of homology detection failure between these classes of genes. I find that genes in microsyntenic regions are significantly less likely to be predicted to be undetectable in *S. castellii* by my analysis than genes outside syntenic regions (Chi-square p-value=3×10^{-13}*),* suggesting that this is the case (Figure 2.14).

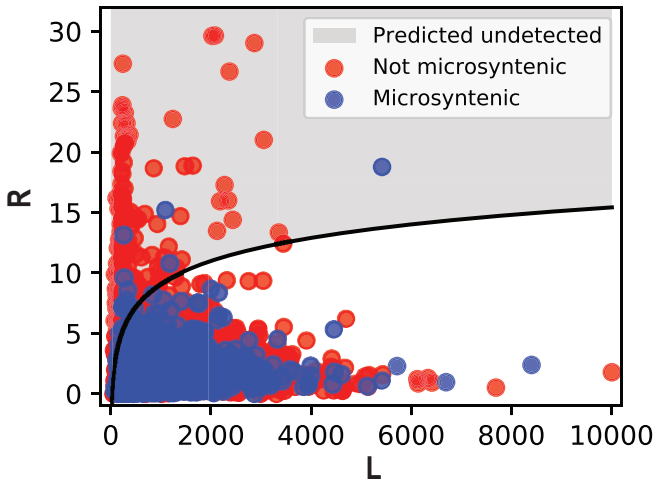

Figure 2.14: Best-fit values of L and R for all genes in my analysis. These two parameters, plus the evolutionary distance between S. cerevisiae and S. castellii (t), determine whether or not an ortholog is expected to be detected at an E-value of 0.001. The gray area corresponds to this prediction, indicating values of *L* and *R* that result in the predicted bitscore being below the detectability threshold (corresponding to a bitscore of 37), defined by the curve $37 < Le^{-Rt}$, where t is evolutionary distance of *S. castellii* from *S. cerevisiae*.

If such a bias were operating, I would expect that, when I restrict my analysis in fungi to consider only lineage-specific genes in microsyntenic regions, I would find a much lower rate of homology detection failure compared to the rate that I infer from all genes (Figure 2.8). I find that this is the case when I perform my detectability prediction analysis only on the subset of lineage-specific genes within microsyntenic regions in *S. castellii* and *K. waltii* (Methods). With the caveat that the total number of genes in microsyntenic regions is small (25 and 77 in the two

lineages analyzed), around 40% of these genes have values of P(detected | null model) less than 0.05, and around 60% less than 0.5 (Figure 2.15).

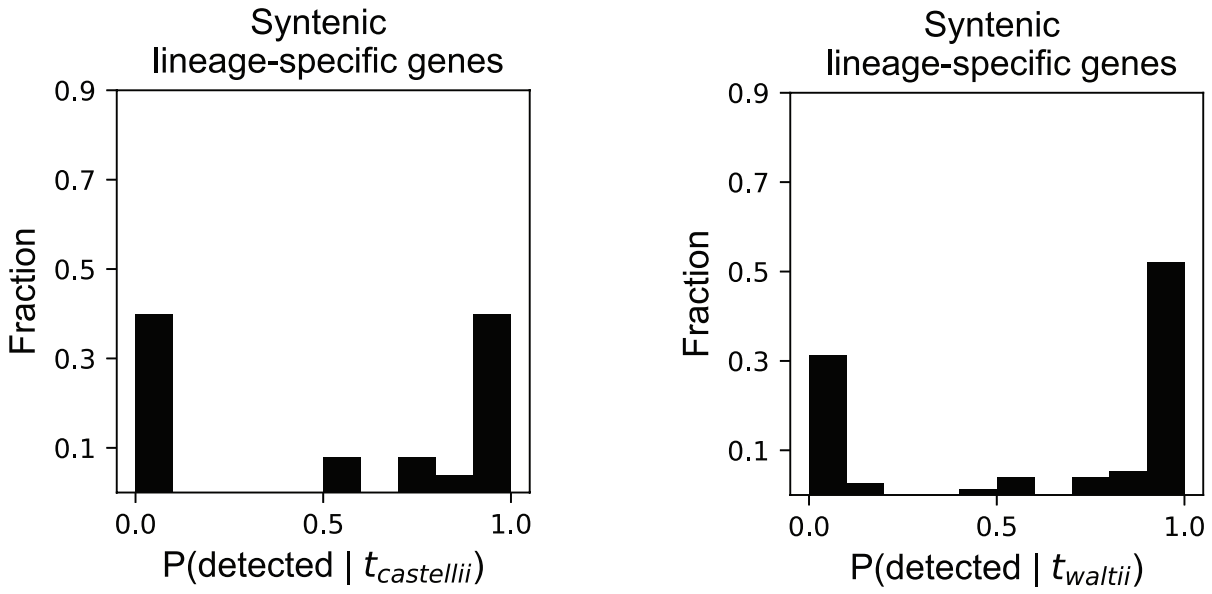

Figure 2.15: Distribution of predicted detectability restricted to lineage-specific genes within microsyntenic regions in the indicated lineages.

This is lower than my findings of around 55% and 75% for all lineage-specific genes. It is also consistent with the estimate in yeast of around 40% from Vakirlis et al.

Based on this analysis, I conclude that genes in microsyntenic blocks evolve more quickly than those outside of such regions, such that considering only this subset of genes leads to a lower inferred rate of detection failure. These two studies roughly agree when I correct for differences in the set of genes used in the underlying analyses.

More sensitive homology searches detect beyond-lineage homologs for many lineage-specific genes well-explained by homology detection failure

If a gene being lineage-specific is due to the failure of BLASTP to detect homologs that are in fact present, I would expect that a more sensitive search will sometimes succeed in finding homologs where BLASTP did not. I asked whether this was the case for genes whose lineage-specificity was consistent with the hypothesis of detection failure: can I use a more sensitive method to find previously undetected homologs for these genes? I refer to such homologs, detected using a different method in species outside of the originally defined lineage, as "beyond-lineage homologs."

We used *sensu stricto* yeast-specific genes as a case study to ask this question. These yeasts and several of their nearest outgroups have a high degree of conservation of chromosomal gene order (synteny), presenting the opportunity for a more sensitive search. A standard similarity search tests all proteins in a large database of sequences, such as a complete proteome. The resulting multiple testing burden requires a higher score to achieve statistical significance than would be required for a search over a smaller number of sequences. In these yeasts, synteny allows us to restrict a similarity search to one candidate gene at the orthologous chromosomal locus, reducing the multiple testing burden and enabling ortholog identification with a lower score. For the fungal species used here, a proteome-wide search would need a BLASTP score of ~37 to achieve an E-value of 0.001, but a single-protein search would only require a score of ~24. Orthologs with scores between these two values would be missed in my initial search but successfully detected with synteny-guided similarity searches.

We used this synteny strategy to search for beyond-lineage orthologs for all *sensu stricto*-specific genes for which the null model of detection failure is a reasonable explanation. I use a threshold of P(detected | null model) < 0.95 to define these genes. This choice is a conservative threshold that corresponds to genes that are insignificant according to a traditional significance test threshold of P(undetected | null model) = 1 – P(detected | null model) > 0.05. There are 126 *sensu stricto*-specific genes that pass this threshold.

To identify the orthologous locus in outgroup yeasts for these 126 *S. cerevisiae* genes, I used the Yeast Gene Order Browser (YGOB), an online resource that curates the chromosomal orthology relationships between species including the *sensu stricto* yeasts, *S. castellii*, *K. waltii, A. gossypii,* and *K. lactis* [119]. 19 of these 126 *sensu stricto*-specific genes are included in YGOB and have an orthologous locus in at least one of these outgroup yeasts. For all of these genes, the upper bound of the 99% prediction interval for the similiarity score predicted by my model is above the detectability threshold of 24 bits, indicating that they are potentially detectable by this analysis. Of these 19 genes, 17 had an annotated gene at the orthologous locus in at least one outgroup species. For 11 of these, at least one of these genes at an outgroup orthologous locus had significant detectable similarity (E<0.001) to the *S. cerevisiae* gene. In all but 2 of these cases, the similarity score fell within my prediction interval (in those 2 cases, the similarity score was slightly higher than predicted). These 11 genes and their proposed orthologs are listed in Table 3.

Table 2.3: List of 11 *S. cerevisiae* genes for which synteny-based searches in YGOB revealed candidate out-of-lineage orthologs, the YGOB IDs of those orthologs, and their synteny search E-values.

Out-of-lineage ortholog YGOB ID (species; syntenic E-value)	SGD ID	Refseq accession
AAR181W (A. gossypii, 9E-05)	YBR230W-A	NP_001018029.1
NCAS0F03520 (S. castellii, 1E-06)	YPR145C-A	NP_001032572.1
NCAS0B08990 (S. castelliii, 2E-04)	YCL048W-A	NP_001032573.1
NCAS0A07920 (S. castellii, 8E-06)	YDR461C-A	NP_001032577.1
Kwal_YGOB_YGL230C (K. waltii; 5e-06)	YGL230C	NP_011284.1
NCAS0J00480 (S. castellii; 2E-05)	YLR053C	NP_013154.1
NCAS0G03380 (S. castellii; 2E-06)	YML053C	NP_013659.1
NCAS0A00920 (S. castellii; 5E-06)	YMR175W	NP_013900.1
KLLA0B03927g (K. lactis; 2E-04)	YHR199C-A	NP_976247.1
NCAS0B02110 (S. castellii; 1E-06)	YBR188C	NP_009747.3
Kwal_27.12107 (K. waltii; 6E-07)	YPL192C	NP_015132.1

Out-of-lineage ortholog YGOB ID (species; syntenic E-value)				KLLA0C14916g (K. lactis, 2E-05)		Kwal_47.18002 (K. waltii; 3E-06)		KLLA0C16225g (K. lactis; 7E-06)			KLLA0E23739g (K. lactis; 1E-04)

In total, I found beyond-lineage homologs for 46% of genes for which I were able to perform a synteny analysis. I note that this is a conservative estimate. I only considered ORFs that are already annotated in outgroup species, although unannotated orthologs may be present. Additionally, the lower bound of the 99% similarity score prediction interval for all remaining 54% of these genes is lower than the threshold required for detection via synteny, so that all have some probability of orthologs still being missed in this analysis.

Some lineage-specific genes are poorly explained by homology detection failure

In all lineages studied here, there are also lineage-specific genes that are poorly explained by the null hypothesis: their similarity score declines too slowly to make homology detection failure alone a good explanation for their lineage-specificity. These are the genes with high values of P(detected | null model). In all six lineages I studied, 10-20% of lineage-specific genes have detection probabilities of 0.95 or greater (Figs 2.8, 2.9).

This result is illustrated by one such *sensu stricto*-specific protein, Spo13, in Fig 2.16.

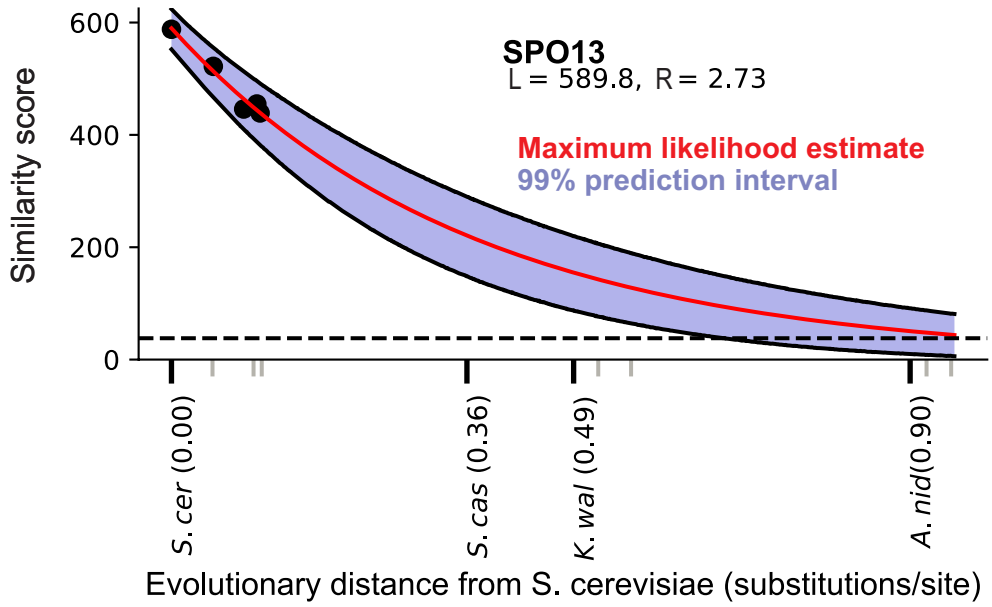

Figure 2.16: Detectability prediction results for the *S. cerevisiae* protein Spo13. At the evolutionary distance of the nearest outgroup *S. castellii*, the entire prediction interval lies well above the detectability threshold, indicating an approximately 100% probability that an ortholog should be detected in this species under the null model.

Spo13 has been proposed as a candidate de novo gene [51] by virtue of its lineage-specificity, and this analysis highlights it as a particularly promising novel gene candidate amongst the large number of other lineage-specific genes in the *sensu stricto* lineage.

The existence of lineage-specific genes like Spo13, which my null model predicts should have detectable homologs outside of the lineage, indicates that evolutionary mechanisms beyond those included in the null model may be operating. Among such mechanisms are those postulated by the novelty hypothesis, like de novo origination and duplication-induced neofunctionalization. However, other known mechanisms could also explain such genes. These include processes that cause the gene tree to deviate from the species tree, like horizontal gene transfer and any mechanisms that change the evolutionary rate of a protein on a restricted part of the tree.

Characterization of yeast lineage-specific genes that are poorly explained by homology detection failure

We next aimed to characterize genes whose lineage-specificity is poorly explained by homology detection failure. I again used *sensu stricto*-specific genes as a case study, allowing for synteny analysis and the biological insight provided by many genes in *S. cerevisiae* being comparatively well-studied. I selected the subset of *sensu stricto*-specific genes, including Spo13, whose lineage-specificity is poorly explained by homology detection failure, i.e. for which P(detected | null model) > 0.95. These are genes on the other side of the threshold applied above: the null hypothesis strongly predicts that homologs should be detected, making their lineage-specificity incompatible with the null hypothesis. There are 25 *sensu stricto*-specific

genes that satisfy this threshold. While a thorough study of these genes is beyond my scope, I report a few initial observations.

"De novo origination," the process of a new gene emerging from previously non-coding sequence, is a commonly proposed origin of lineage-specific genes [34]. I asked how many of these 25 lineage-specific genes might be such de novo genes. By definition, genes that have emerged de novo in the *sensu stricto* lineage should have no out-of-lineage homologs, and so the more sensitive synteny-based homology search strategy used above should fail to find such homologs. I performed a synteny-based search for out-of-lineage homologs for these 25 genes in the same way as above. For 20 of these 25 genes, an orthologous locus is listed in YGOB. Of these 20, 12 have annotated genes with significant similarity ($E<0.001$) at the orthologous locus in at least one outgroup species. Thus, 12 of 25 genes, or just under half, of genes that are not well-explained by homology detection failure did not originate de novo in the *sensu stricto* lineage. This is a conservative estimate of the total number of genes that have out-of-lineage homologs, since, as described above, even this synteny-based homology search has finite sensitivity. Spo13, the gene shown in Fig 5, is one example of these lineage-specific genes that nonetheless are not de novo originated: it has out-of-lineage orthologs identifiable by synteny in *S. castellii*, *K. waltii*, *K. lactis*, and *A. gossypii*.

Genes that acquire a new function following duplication and divergence ("neofunctionalization") are another proposed source of lineage-specific genes [34]. I therefore asked how many of my *sensu stricto*-specific genes have a paralog, consistent with the hypothesis that they emerged through duplication and divergence. Based on BLASTP searches within the *S. cerevisiae* genome, I find that 4 of the 25 lineage-specific genes have annotated paralogs specific to some subset of the *sensu stricto* yeasts, which therefore likely emerged after

their divergence from *S. castellii*. I also find using YGOB that another 4 of these 25 genes have annotated paralogs resulting from the yeast whole genome duplication, which occurred before the divergence of *S. castellii* from the *sensu stricto* yeasts. In total, 8/25, or fewer than one-third, of these genes show evidence of having been the result of duplication events. However, I note that this estimate for the number of genes with paralogs is again conservative due to the finite sensitivity of the homology searches.

We performed a gene ontology enrichment test (Methods) to determine if certain biological processes were statistically overrepresented among these 25 genes. I find significant enrichment of genes involved in several GO categories relating to spore formation and meiosis, including "ascospore-type prospore membrane assembly" ($p = 7*10^{-5}$; 3 observed vs 0.7 expected) and "meiotic cell cycle process" ($p = 5*10^{-5}$; 7 observed vs 1 expected). Spo13, involved in meiotic cell cycle regulation through its roles in maintaining sister chromatid cohesion during meiosis I and promoting kinetochore attachment [120], is one such example. By contrast, no biological processes were overrepresented among lineage-specific genes that are consistent with homology detection failure (although these genes are much less likely to have GO annotations at all: 92% have no annotation, compared to 12% of all *cerevisiae* genes and 36% of lineage-specific genes that are inconsistent with homology detection failure).

Finally, I assessed the length and rates of these genes. Analyses of lineage-specific genes regularly find that they are shorter and faster-evolving than conserved genes [23, 32, 49, 121] There has been controversy about how to interpret this finding. On one hand, it is tempting to conclude that these features are indicative of genetic novelty, given that they are natural consequences of a compelling conceptual model for how novel genes might be born and evolve: emerging from noncoding DNA and the short random ORFs that it encodes, at first under no

selective constraint and so free to evolve quite rapidly [23]. On the other hand, as is formalized in my model but has been appreciated widely for some time [4, 21, 98-100], shorter and faster-evolving genes are more prone to detection failure; this bias is therefore also expected under the hypothesis that lineage-specific genes are largely due to detection failure [98, 99, 103-105, 107]. Do lineage-specific genes appear short and fast-evolving because they are these technical artifacts, or is this observation a biological reality reflecting their novelty?

In my model, the parameters L and R correspond roughly to length and evolutionary rate respectively, and jointly determine whether homologs of a gene are predicted to be detectable at a given evolutionary distance. This dependence is given by the equation $S = Le^{-Rt}$, where S is the bitscore corresponding to the chosen significance threshold and t is the evolutionary distance in question. All lineage-specific genes within the portion of L-R parameter space given by $S < Le^{-Rt}$ are therefore genes for which detection failure is a sufficient explanation. Similarly, all genes for which $S > Le^{-Rt}$ are genes for which detection failure is not a sufficient explanation – the genes under consideration in this section, which I consider to be interesting candidates for genetic novelty.

We can ask how these genes are distributed within L-R parameter space. I take for granted that they reject my null model and so satisfy $S > Le^{-Rt}$. Beyond this, they might be distributed similarly to conserved genes within this portion of parameter space: that is, there might be no way in which they look particularly special. Alternatively, their values of L and R might be biased in some fashion, such that they look noticeably different from conserved genes.

We find that these genes are noticeably different from other, non-lineage-specific genes. Specifically, their lengths and rates are highly biased toward being shorter and faster-evolving

(lower L and higher R) than is true of conserved genes falling below the curve $S = Le^{-Rt}$ (Figure 2.17).

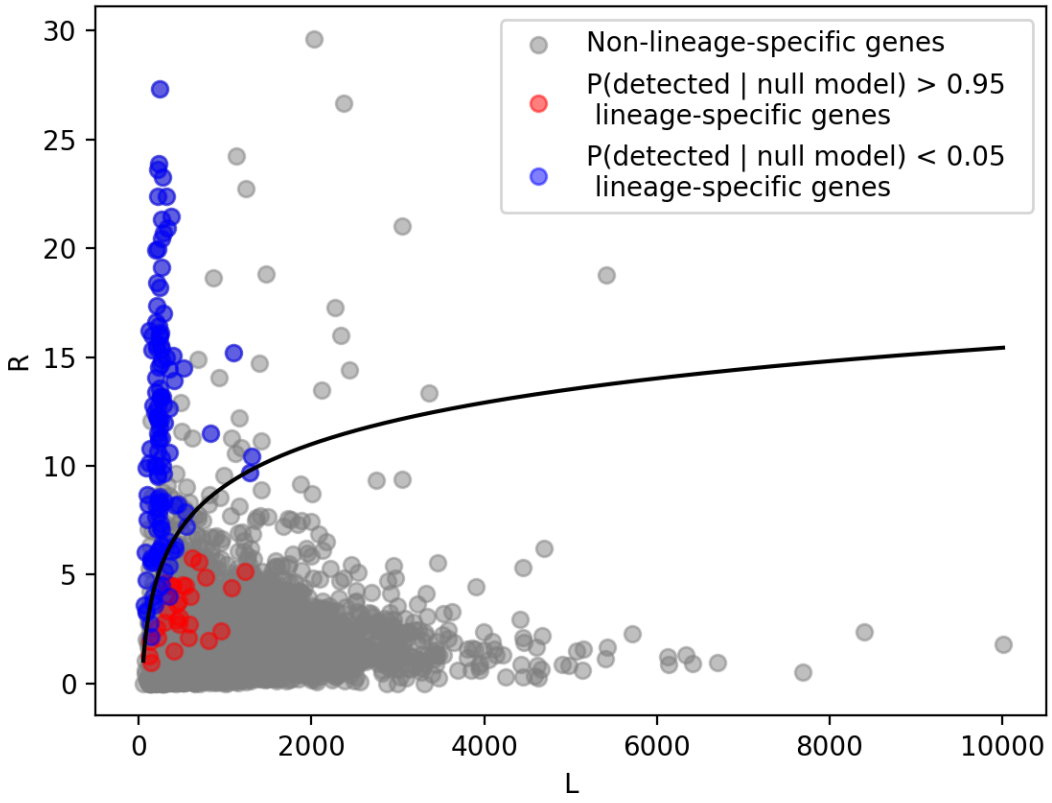

Figure 2.17: Distribution of L and R parameters as inferred from the *sensu strictu* yeast clade for all *S. cerevisiae* genes. Genes specific to that clade, and so absent from the nearest outgroup *S. castellii* and all other fungi, are indicated in color: red genes are those that reject the hypothesis of detection failure at p = 0.05, and blue are those that do not.

By considering only genes that reject the hypothesis of rejection failure, I minimize the contribution of technical artifacts to the lengths and rates of lineage-specific genes. This result suggests that their short lengths and fast evolutionary rates may be biologically meaningful.

An alternative explanation for this pattern is that these genes are merely those that were fairly short and fast in the ingroup where the parameter fitting was conducted: they lay close to, but just on the predicted-undetectable side of, the curve $S = Le^{-Rt}$. If these parameters happened to significantly change, such a change would be likelier than for genes elsewhere in parameter space to bump them over the curve into the space that I predict to be detected. These changes could happen for reasons other than novelty and would explain the observed bias in lengths and rates. Because my approach only detects deviations from my evolutionary null model and does not attempt to pinpoint the particular cause of the deviation, this possibility cannot be excluded.

abSENSE is a simple method to interpret undetected homologs and is freely available as downloadable code and a graphical user interface-based web server

We consider one of the major advantages of the method for predicting homolog detectability, and therefore interpreting the absence of detected homologs, developed here, to be that it should be easily applicable to new genes and new taxa, requiring only two sets of relatively straightforward input parameters: a) the bitscores of detected homologs in at least three species, including the focal species, and b) the pairwise evolutionary distances between the focal species and the other species under consideration. This is especially true compared to alternative simulation-based approaches described in the Introduction.

We have named my method abSENSE, and have made it available to the research community. A fully documented software package is available (https://github.com/caraweisman/abSENSE), allowing abSENSE to be downloaded and run locally. The package includes all required input parameters for running abSENSE on all genes considered here. It also contains instructions for running abSENSE on genes in new taxa.

We also created a web server (http://eddylab.org/abSENSE/), lowering the barrier to use of a command line interface by offering a graphical user interface. Like the downloadable code, the site is preloaded with all data required to analyze all *D. melanogaster* and *S. cerevisiae* genes considered here, and can also be used to analyze user-provided data from genes in new taxa.

Conclusions

The widespread interpretation of lineage-specific genes as evolutionarily novel assumes that absence of evidence for detectable homologs in outgroups is evidence that homologs are absent. The model I have presented here allows us to formally test the alternative, null hypothesis: homologs do exist outside the specified lineage, but they have diverged, at a constant novelty-free evolutionary rate, beyond the ability of a similarity search program to detect them. I find that this hypothesis is sufficient to explain a large number of lineage-specific genes in two taxa where lineage-specific genes have been interpreted as exhibiting some kind of evolutionary novelty. These results caution against automatically assuming that lineage-specific genes are novel.

Two important caveats should be kept in mind. First, this method cannot exclude the possibility that a gene is truly novel, but *also* short enough and evolving fast enough that its ortholog would not be detected if present, such that homology detection failure can also explain

its lineage-specificity. For this reason, it may be difficult for de novo genes in particular, which have been hypothesized to be short and fast-evolving, to reject the null model. However, in this case, where two hypotheses can both explain a lineage-specific gene, I argue that additional evidence should be required to prefer the comparatively exotic hypothesis of novelty to the more conservative one of detection failure. my case study in the *sensu stricto* yeasts finds that more sensitive synteny-based homology searches successfully find previously undetected homologs for many lineage-specific genes, supporting this argument. Second, these results may or may not generalize to classes of lineage-specific genes that I have not considered here. Because my method requires at least two observed orthologs, I have only applied it here to genes found in at least two species in the lineages in question. Moreover, like other studies, I focused on genes in existing annotations, which are prone to biases that may exclude novel genes. However, when I analyze *cerevisiae* ORFs with the requisite 2 orthologs that are marked as of "dubious" coding status in the Saccharomyces Genome Database, I find that an even larger proportion – nearly all – are unlikely to be detected. Although I have not done so here, I note that this method could be extended to these classes of genes. In principle, individual conspecifics with sufficient genetic differentiation could be used as discrete taxa in my method to analyze genes found in only one or two species. Additionally, the method is readily applicable to any protein annotation, which could be custom-made as desired.

Although I find that many lineage-specific genes can be adequately explained by homology detection failure, I also find a minority of lineage-specific genes in fungi and insects that cannot. This leaves open the possibility that these genes are biologically novel. However, the reason that these genes reject my null model is not addressed by my present work. my initial analyses do show that many of these genes are neither de novo genes nor have detectable

paralogs, suggesting that processes other than the commonly proposed hypotheses of de novo origination and duplication-divergence may be at play. There are many possible processes that could cause genes to deviate from my null model, but one speculative example lies in the observed enrichment in yeast of genes involved in meiotic processes, exemplified by Spo13. This strikes us as suggestive of meiotic drive phenomena, which have been observed in yeast [122] and have been shown to cause rapid protein divergence [123], producing clade-specific rate accelerations leading to lineage-specific genes. More detailed characterization of these genes is required to understand if and in what way they are evolutionarily novel.

There is increasing consensus that homology detection failure is frequent [97]. It should be taken into account in studies that aim to use lineage-specific genes to identify candidates for genetic novelty. To this end, my approach allows us to determine whether a *particular* lineage-specific gene is attributable to homology detection failure. Synteny analyses of the kind used here can sometimes be used to determine whether out-of-lineage orthologs are present [46, 124] and can provide strong evidence of de novo origination [125], but syntenic analyses are only possible in the limited taxa where sequenced species are related closely enough that synteny is conserved. By contrast, my method can be used in any set of species for which relative evolutionary distances are known. I hope it to be useful in the wide variety of studies that aim to identify novel genes that may underlie the evolution of morphological, behavioral, and other novel traits [11, 20, 25, 71, 126, 127].

Methods

Identification of *S. cerevisiae* and *D. melanogaster* orthologs

We downloaded previously annotated proteomes of all species used here from several sources, largely Refseq and GenBank. Accession IDs for Refseq and GenBank proteomes and download links for those from other sources are listed in S2_Table. I performed a BLASTP (version 2.8.0) search [128] with an E-value threshold of 0.001 using the *S. cerevisiae* proteome as the query against each of the 11 other yeast proteomes independently. I also performed the reciprocal of each of these searches, using each of the 11 other yeast proteomes as the query against the *S. cerevisiae* proteome. I used a custom Python script to identify reciprocal best BLAST hits for each *S. cerevisiae* protein in each of the other yeast proteomes. A protein in another yeast's proteome was considered a reciprocal best hit to the *S. cerevisiae* protein if a) the E-value of the *S. cerevisiae* protein against that protein was the lowest of any in that species' proteome and b) the E-value of that protein against the *S. cerevisiae* protein was the lowest of any protein in the *S. cerevisiae* proteome. Proteins in the other yeast species satisfying this reciprocal best hit criterion were considered orthologs of the *S. cerevisiae* protein. When no significant homology to a *S. cerevisiae* protein was detected in another species, or when the reciprocal best hit criterion was not met by any protein in that species, no ortholog was assigned in that species. To identify orthologs for *D. melanogaster* proteins, I repeated this same procedure for all *D. melanogaster* proteins and each of the 21 other insect species' proteomes.

Calculation of evolutionary distances

Because evolutionary distance t only appears in my model as a product with the gene-specific rate parameter R, I can use a subset of genes in the species group to infer these relative distances. Each gene's value of R will scale these relative distances appropriately when fit to the model: genes that evolve faster than these relative distances will have values of R above 1, and

slower, below. I used BUSCO genes as the subset of genes from which to estimate distances, as they are generally well-conserved, facilitating ortholog identification and alignment. This enables my desired result of a species tree with correct relative evolutionary distances (in substitutions/site across aligned BUSCO genes), which is the only feature needed by my downstream inference. I downloaded a list of eukaryotic BUSCO genes [112] from the BUSCO web server (https://busco.ezlab.org/) and identified all of these genes for which I were able to identify an ortholog of the corresponding *S. cerevisiae* gene in all 11 other yeast species ("Identification of orthologs" above). I found 102 such BUSCO genes. I used the alignment software MUSCLE (version 3.8.31) [129] with default parameters to create a multiple sequence alignment of the orthologs from all 12 yeast species for of each of these 102 genes. I then concatenated these alignments and used the Protdist program from the PHYLIP software package (version 3.696) [130] with default parameters to find pairwise evolutionary distances for all 12 yeast species in substitutions per site. To test the effect of using a smaller number of genes to infer these distances, I then randomly and independently selected two subsets of 15 of these 102 genes, and performed the same alignment and distance calculation procedure on each of these two subsets. I then performed the same procedure using *D. melanogaster* genes and the 21 other insect species. Here, there were 125 BUSCOs for which I were able to identify orthologs in all species, and the two random subsets of 15 genes were selected from among these 125. Refseq accessions for genes in the three sets of BUSCOs in both taxa are listed in S7_Table.

Correlation of *R parameter* with evolutionary rate

To determine the correlation between each gene's best-fit value of the *R* parameter in my model and the substitution rate, I used alignments of 5261 *S. cerevisiae* genes and their orthologs

in all four other *sensu stricto* yeast species generated by a previous study [118]. I opted not to include more distantly related species in these alignments for the sake of more reliable ortholog identification and alignment construction. I used the protdist function of the PHYLIP package (version 3.696) [130] on these alignments to infer the number of substitutions per site between the *S. cerevisiae* gene and its ortholog in the most distant *sensu stricto* yeast *S. kudriavzevii* (we chose a fairly distant representative of these species to minimize sampling error from low substitution counts), and correlated this value with the *R* parameter inferred from the regression analysis.

Identification of lineage-specific genes

To identify *S. cerevisiae* genes specific to the three yeast lineages tested here, I performed a BLASTP search [128] with an E-value threshold of 0.001 for each gene in the *S. cerevisiae* proteome as the query against each of the 11 other yeast proteomes independently. If the BLASTP search detected no homologs of the *S. cerevisiae* gene in the proteomes of any of these species outside of the specified lineage, I considered it lineage-specific. I applied the same criterion using the 21 other insect proteomes to identify *D. melanogaster* genes specific to the three insect lineages tested here.

Synteny-based homology searches

We used version 7 of the Yeast Gene Order Browser's online web tool (http://ygob.ucd.ie/) [119]. For tested *S. cerevisiae* genes, if the gene was included in this YGOB version, I determined whether an orthologous chromosomal region in any of the outgroup yeast species used here had been identified in the browser. If so, I searched for any genes in these

outgroup species at the locus that were annotated in the browser. I considered genes to be within the outgroup orthologous locus if they were between the outgroup's orthologs of the closest *S. cerevisiae* genes up- and downstream of the query gene. If annotated genes existed at the orthologous locus, I performed a BLASTP search of the *S. cerevisiae* sequence against the sequences of all outgroup genes at that locus as listed in YGOB, and called orthology in cases where this single-search E-value was <0.001.

Gene ontology analysis

We used the Gene Ontology Consortium's online web server (http://geneontology.org/) [131] to test whether or not certain biological functions were enriched in the set of *sensu stricto*-specific genes that I found to be poorly explained by detection failure. I performed a Fisher's exact test using the "GO biological process complete" annotation data set for all *S. cerevisiae* genes.

Calculation of substitutions/site and gaps for *sensu strictu* alignments

Although Vakirlis et al. (2020) [97] provide dN values computed from alignments of *S. cerevisiae* genes and their *sensu strictu* orthologs, these underlying alignments are restricted to the 5261 genes for which orthologs were identified in all five *sensu strictu* species in a previous study [118]. Because I worried that this might introduce a bias against quickly-evolving genes, for which orthologs are less readily identifiable, I opted to make my own alignments, including all 5586 genes for which I could identify an ortholog in at least one of the four other *sensu strictu* species. I used both MUSCLE [55] and ClustalOmega [132] with default settings to produce a multiple alignment of each *cerevisiae* gene and its orthologs in at least one other *sensu strictu*

species. I then used these alignments to compute substitutions/site between *S. cerevisiae* and *S. bayanus* with the Phylip ProtDist program as in my other evolutionary distance calculations (above). I chose *S. bayanus* because it is the most distant species from *cerevisiae* according to my analysis. I then used all genes with an ortholog present in *S. bayanus*, regardless of its status in the three other yeast species, in the subsequent analysis. Results from the two alignment programs were extremely similar, as were results using distances to the slightly closer *S. kudriavzevii.* I then used these same alignments to count the total number of gaps in each alignment and divide by the number of columns and number of sequences in the alignment to calculate the gaps per column per sequence.

Detectability prediction analysis of microsyntenic lineage-specific genes

We repeated my original analysis of fungal lineage-specific genes, but restricted to genes determined to be in microsyntenic regions by Vakiriis et al (2020) [97]. I included genes in this analysis as follows. I started with the same list of genes specific to the lineages for which *S. castellii* and *K. waltii* are the closest outgroups (Fig 4) as in my original analysis. From these, I selected genes that Vakirlis et al. determined to be in a microsyntenic region in at least one of a) a species within that lineage; b) that species itself (*S. castellii* or *K. waltii*); or c) another outgroup species to the lineage of very similar divergence time to b). I chose to include genes in microsyntenic regions in species within the lineage and not just in its closest outgroup to be maximally conservative, and because the number of genes in microsyntenic regions only in the outgroup species was low. I chose to allow for another outgroup species of similar divergence time to be substituted for the species that I used as outgroup because the set of species for which Vakirlis et al. performed a synteny analysis did not overlap exactly with the set of species used

here (for example, *K. waltii* was not included), such that this was the closest approximation possible using those data. In the case of *S. castellii*, these species included *S. arboricola, S. kudriavzeviii,* and *S. castellii* itself. In the case of *K. waltii*, these species included *K. lactis, A. gossypii, L. thermotolerrans, E. cymbalariae, A. aceri, S. arboricola, S. kudriavzeviii,* and *S. castellii.*

Chapter 3: Why correct interpretation of lineage-specific genes matters: the evolutionary origin of mesoderm

Background

So far, I have described and quantified the contribution of processes that generate lineage-specific genes which have nothing to do with evolutionary or biological novelty. The motivation for this enterprise has been to assess the widespread interpretation of lineage-specific genes as largely or entirely novel. On the basis of results so far, I propose that this assumption is not generally reliable. Here, I explore how making this assumption could confuse inquiry into biological questions beyond those regarding lineage-specific genes *per se*. Does erroneously attributing novelty to lineage-specific genes actually muddy understanding of more fundamental or wide-ranging biology? Or is the impact of the findings in the previous chapters limited to the relatively specialized niche of lineage-specific and novel genes?

I take up what I consider to be an evolutionarily impactful question to which the current consensus has been partially based on interpreting lineage-specific genes as novel. I reinterpret the data on which this consensus was based using analysis methods developed in previous chapters, consider the effects of this new interpretation on our understanding of the biology at issue, and describe a new avenue of experimental work opened by this analysis.

The evolutionary origin of mesoderm: one or multiple?

Mesoderm is a primary layer of cells that emerges early in the development of some animals, one of three so-called “germ layers” that are present in some animals [133]. In the

species in which mesoderm is present, it gives rise to tissues including muscle, connective tissue, blood, and kidneys [133]. As such, mesoderm is an important evolutionary innovation that enabled the novel structures and functions made possible by these derivative tissues.

Mesoderm was characterized and defined in Bilaterians [133]. The phylum Cnidaria appears to lack a cell lineage with extremely strong morphological or ontological similarity during development to Bilaterian mesodermal lineages, and also lacks the mesodermal derivatives characteristic of Bilaterian mesoderm (e.g. fully-fledged muscles) [134]. The same is true of the early-branching animal phyla Porifera and Placozoa [135]. The situation is somewhat more complicated for the phylum Ctenophora. Ctenophores have contractile muscle cells that are molecularly and functionally similar to those of Bilateria [136, 137], and at least one order also has striated muscle [138]. They also have micromeres that derive from larger macromeres before migrating toward the center of the embryo during gastrulation and settling between clearly endodermal and ectodermal lineages, then giving rise to muscle cells in the adult [139]. They share these morphological and developmental features with Bilaterian mesoderm cells, but differ in others: for example, the mesodermal lineage originates from the animal pole, rather than the vegetal pole, as in Bilaterians (and as in the so-called "endomesodermal" lineage of Cnidaria) [140]; the micromere mesoderm progenitors lack organization into a clear embryonic germ layer [141]; and striated muscle, where present, has significant cellular differences, including the absence of nuclei and most organelles [138]. Different authors hold different standards for what features are necessary for a lineage to be defined as mesoderm, resulting in Ctenophores sometimes being described as possessing mesoderm [139, 142, 143], and sometimes as lacking it [142] [134].

In discussing whether another taxon "has mesoderm," one might mean to comment on whether that taxon has a structure that is homologous to Bilaterian mesoderm. Alternatively, one may mean merely to comment on the presence of a structure that meets the criterion of a set of shared morphological or functional features. Here, based on the similarities described above, I will refer to the Ctenophore structure in question, which bears the described similarities to Bilaterian mesoderm, *as* "mesoderm," but in doing so, I will *not* be asserting the homology of the two. I will merely be using the term to refer to their shared developmental characteristics. Accordingly, so far, one can summarize the situation thus: Bilateria and Ctenophores both have mesoderm; Cnidaria, Placozoa, and Porifera lack it.

Phylogenetically, the current consensus is that Cnidaria are the sister taxon to Bilateria, with Placozoa, Ctenophora, and Porifera each branching independently at some earlier point [144, 145]. The relative placement of these three taxa, and in particular whether Ctenophora or Porifera is sister to the rest of the animals, remains unclear and is at present the subject of controversy [146]. There is nonetheless significant consensus on two points: that Cnidaria are sister to Bilateria, and that Ctenophores are not the next-latest branching taxon, i.e. that either Porifera, Placozoa, or both branched between Ctenophores and the common ancestor of Bilateria and Ctenophores [145, 146]. A consequence of these possible phylogenies, if true, is that mesoderm is paraphyletic: it is absent in either one or two descendants of the common ancestor of Ctenophores and Bilateria, the two taxa possessing mesoderm.

These two observations – the commonality of some but not all developmental and morphological features, and paraphyly -- jointly raise the question: is Ctenophore mesoderm homologous to Bilaterian mesoderm? There are two maximally simple and distinct possibilities. The first, the "homology hypothesis," is that they are indeed homologous: that mesoderm

evolved once in the common ancestor of Ctenophores and Bilateria, and was lost either two or three times (depending on the correct animal tree topology), in Cnidaria, Placozoa, and perhaps Porifera. The second, the "convergence hypothesis," is that they are not: that Ctenophores and Bilaterians each evolved their respective mesoderms totally independently. Of course, between these two possibilities are a slew of subtler combinations: for example, that pieces of the mesodermal program were present in the common ancestor, with independent elaborations occurring later; or that mesoderm was present in the common ancestor, but that one or both lineages have independently switched out or reinvented components of the ancestral program.

How can one test these possibilities? Decades of experimental work have generated substantial knowledge of the genetics of mesoderm generation in Bilaterians. The strictest version of the homology hypothesis holds that these same genes should then be involved in mesoderm generation in Ctenophores. At a bare minimum, then, one would expect that these Bilaterian mesoderm genes have homologs in Ctenophores.

The publication of two Ctenophore genomes, from *Mnemniopsis leidyi* [136] and *Pleurobrachia bachei* [147], saw this question asked and apparently answered. The authors selected Bilaterian genes known to be involved in mesoderm generation and asked whether they had detectable homologs in these two Ctenophore species. Although BLASTP detected clear homologs for some of the Bilaterian mesoderm genes that the authors considered, a majority of these genes had no apparent homologs: they appeared to be specific only to the Bilaterian lineage. On this basis, the authors concluded that Bilaterian and Ctenophore mesoderm likely evolved convergently [136, 147], a view that has since been endorsed by others [139, 144].

The evolutionary distance between Ctenophores and Bilaterians is large (~800 million years [148]). As evolutionary distances increase, as demonstrated in the previous chapter, the

probability of homology detection failure increases. I therefore wondered whether Ctenophores may actually have homologs of these Bilaterian genes that BLASTP has merely failed to detect. If true, this would have clear implications for our current understanding of the relationship between Bilaterian and Ctenophore mesoderm: if Ctenophores do have these homologs, the homology hypothesis, currently considered falsified, becomes viable again.

Here, I use my method abSENSE, developed in the previous chapter, to assess whether Ctenophores may have undetected homologs of Bilaterian mesoderm genes. I find that this is the case for many of these genes; use results from abSENSE and more sophisticated homology search methods to identify genes in *M. leidyi* that may be these homologs; and describe ongoing work to experimentally test whether these Ctenophore genes do indeed have mesodermal functions.

Application of abSENSE to Metazoa

We first applied abSENSE to Metazoa, repeating the methods used for the fungal and insect lineages as described in Chapter 2. I used *H. sapiens* as my focal species due to its high-quality genome and due to the comparatively short branch length of the vertebrate lineage among Bilateria, such that distance to outgroups would be minimized. I included 18 other target species, selected to roughly uniformly cover the evolutionary distances between *H. sapiens* and the most distant outgroup, the choanoflagellate *Monosiga brevicolis*, and on the basis of genome quality. I also included the Ctenophore *M. leidyi*. I did not include the second Ctenophore with a sequenced genome, *P. bachei*, because its genome was not publicly available at the time of analysis. These species and a consensus topology are shown in Figure 3.1.

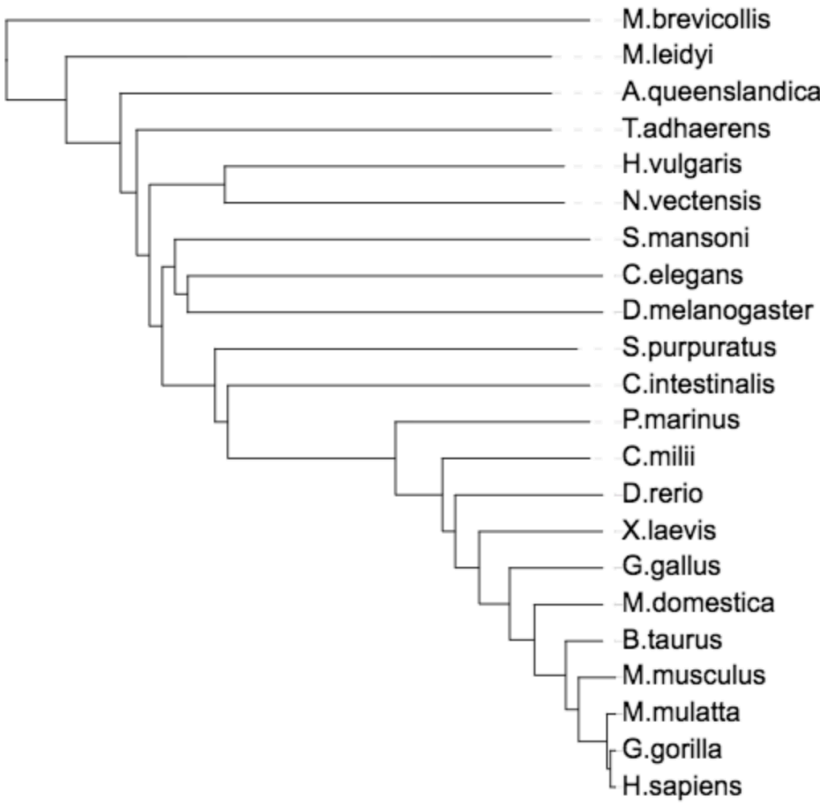

Figure 3.1: Species used in the abSENSE analysis performed in this chapter. Tree inferred using an alignment constructed according to the methods for BUSCO gene selection and alignment described in Chapter 2, which was then analyzed by the ProML program in Phylip using default parameters.

We computed the pairwise evolutionary distances between *H. sapiens* and each of the target species as described in Chapter 2. I also identified orthologs of all *H. sapiens* genes in each of the target species as described in Chapter 2.

We then sought to confirm the accuracy of my mathematical model of homology detectability and the accuracy of the computed evolutionary distances. As in Chapter 2, I

performed a regression analysis comparing the bitscore predicted by my model and the actual bitscore of each *H. sapiens* ortholog in each target species. The distribution of resulting r^2 values for all *H. sapiens* genes are shown in Figure 3.2.

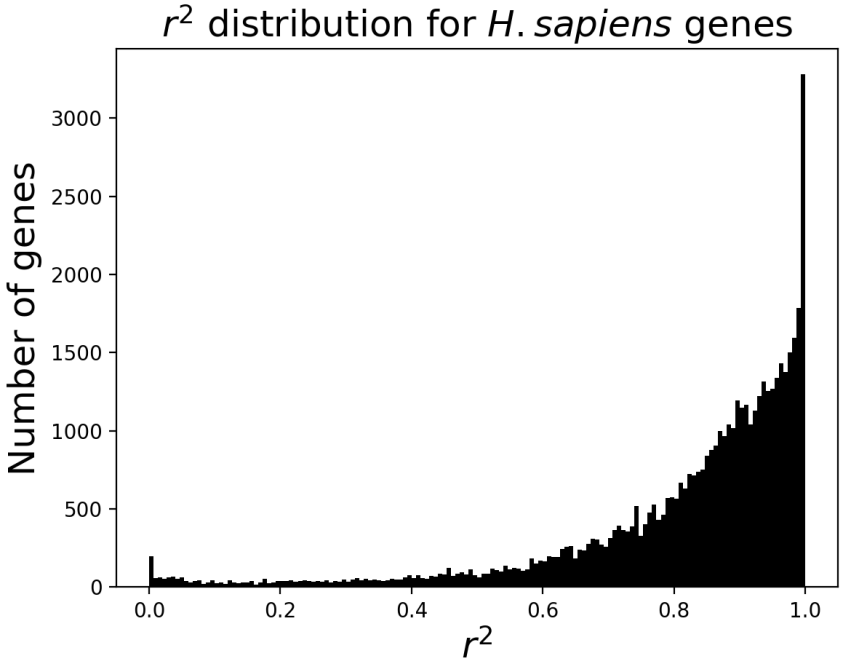

Figure 3.2: r^2 for regression analysis assessing the goodness of fit of the model to the bitscores of human orthologs in the animals considered here.

The mean and median r^2 values were 0.82 and 0.88 respectively, demonstrating a good fit to the model.

We aim to use this model to predict the detectability of orthologs of human genes in *M. leidyi.* To be confident that the model makes predictions in this species, I repeated the positive control described in Chapter 2, computing the probability of being detected for genes that do in fact have detectable homologs in *M. leidyi*. As in Chapter 2, I computed this probability without

using the bitscores of thee genes in *M. leidyi* or any more distant species, mirroring the amount of data that available to fit the model parameters for the desired experimental case of lineage-specific genes. The results of this analysis are shown in Figure 3.3.

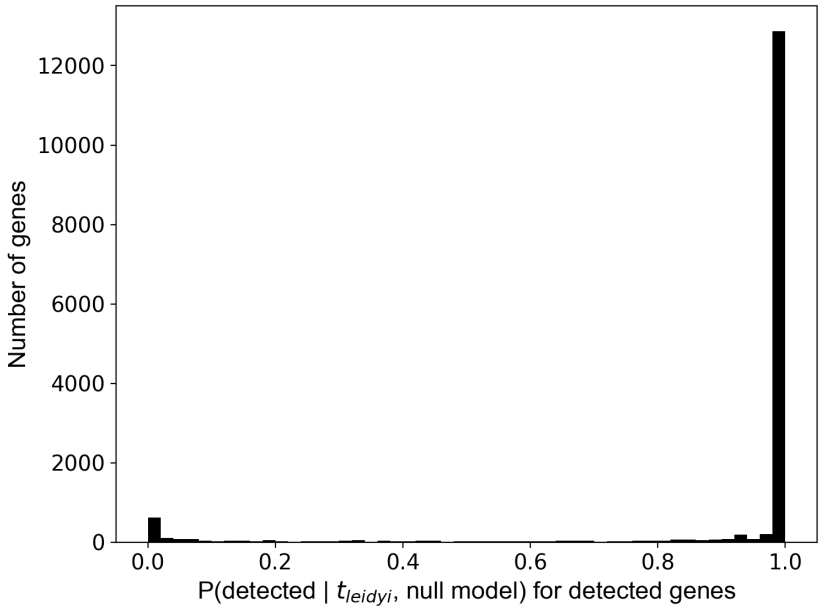

Figure 3.3: Distribution of P(detected | null model) in target species *M. leidyi* for all human genes that do have detected homologs in *M. leidyi.*

As in the fungal and insect clades, I correctly assign a very high probability of being detected to the vast majority (>97%) of these genes, indicating the reliability of the model's predictions.

Homologs of many Bilaterian mesoderm genes are not expected to be detectable in *M. leidyi*

With a reliable method to predict whether orthologs of *H. sapiens* genes would be detected in *M. leidyi* if evolving according to my null model, I then used this method to consider the Bilaterian mesoderm genes considered in previous analyses. These genes were found to lack homologs in *M. leidyi*, on the basis of which it was proposed that Ctenophore and Bilaterian mesoderm likely evolved independently [136, 147].

We used abSENSE to predict the probability of an ortholog of the *H. sapiens* gene being detected in *M. leidyi* for 12 genes considered in these previous analyses and found to lack apparent homologs by BLASTP. I selected these genes according to the following criteria. First, I considered all genes that were considered in either of the two published analyses that considered the presence and absence of Bilaterian mesoderm genes in Ctenophores [136, 147] for which a homolog was not detected. I note that these genes may or may not ultimately be the most informative, or not all equally so, for assessing the homology or convergence of mesoderm; for example, one might place more emphasis on genes controlling earlier stages of mesoderm development and less on genes involved in later stages of differentiation that are unique to Bilateria, for example those involved in heart morphogenesis. I chose to use these same genes to reassess the claims previously made, such as they are, with respect to how homology bears on the question of mesoderm's evolutionary origins, independent of my own views about whether these genes are optimally informative, or all equally so, for the question of the origin of mesoderm. For this reason, here I only describe the results of this analysis, and do not discuss the functions of these genes in mesoderm. Further work in collaboration with experts in mesoderm development and evolution may consider if other genes would be more informative, or if some genes among these should be weighted more highly.

We then excluded genes for which members of the same gene family were clearly present in Ctenophores, but for which there was ambiguity about identity within the family that could not be resolved on the basis of the presence of a distinguishing protein domain. For example, many proteins with a single homeobox domain are clearly present in Ctenophores, but it is not clear on the basis of homology searches alone, and is therefore not answerable by my method, if any of these is an ortholog of the human mesoderm gene goosecoid; determining this would require creating a gene tree, which is difficult and uncertain at evolutionary distances this large. I retained human mesoderm genes that share domains with other members of a large family, but for which the presence of a distinguishing domain is unique to the gene with mesodermal function. For example, HNF1a/Forkhead shares a Forkhead domain with a large family of genes, including other non-mesodermal genes, but has a characteristic HNF-C domain that is therefore diagnostic of identity within this family. In these cases, my method can be used to predict the detectability of the diagnostic domain and therefore the gene.

As a positive control for the possibility that such mesoderm genes are in some way different than the majority of other genes, such that my prediction method is more prone to err in analyzing them, I also considered 7 genes included in these previous analyses for which a Ctenophore homolog was successfully detected by BLASTP. As above, these 7 genes included all such genes included in these previous studies for which assigning family membership was unproblematic.

Consistent with my analysis in Chapter 2, I considered a human gene as being predicted to be detected in *M. leidyi* if it had a 95% or greater value of P(detected | null model, t_{leidyi}). The results of these predictions are shown in Figure 3.4.

Human mesoderm gene	Ctenophore homolog detected (BLAST)?
Myf5	No
Sonic hedgehog	No
Chordin	No
Eomesoderm	No
Forkhead	No
MyoD	No
Myogenin	No
FGF	No
Follistatin	No
Cerberus/DAN	No
Mrf4	No
Noggin	No
Troponin	No
Snail	Yes
Brachyury	Yes
GATA	Yes
TGF-B	Yes
Smad	Yes
Myosin	Yes
Tropomyosin	Yes

Ctenophore homolog detected (BLAST)?	Ctenophore homolog predicted detectable (BLAST)?
No	Yes
No	Yes
No	Yes
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
Yes	Yes
Yes	Yes
Yes	Yes
Yes	Yes
Yes	Yes
Yes	Yes
Yes	Yes

Figure 3.4: Results of abSENSE predictions for Bilaterian mesoderm genes in Ctenophores analyzed in previous studies to assess the likelihood of the homology vs. convergence of Bilaterian and Ctenophore mesoderm. The left table depicts the results of BLASTP searches in the *M. leidyi* genome. Red shading indicates genes taken to be inconsistent with the homology hypothesis, on the basis of lacking apparent homologs in *M. leidyi;* blue shading indicates genes taken to be consistent, based on having apparent homologs. The right table depicts the results of both BLASTP searches and my abSENSE analysis. Red shading indicates genes inconsistent with the homology hypothesis, which both lack detected homologs and should have detectable homologs according to abSENSE analysis; blue shading indicates genes consistent with the homology hypothesis, which lack detected homologs but for which abSENSE predicts homologs are not expected to be detectable. (The bottom seven genes, in blue on the left, serve as a control for the reliability of abSENSE predictions.)

A majority (75%) of the genes for which no homolog was detected in Ctenophore are not predicted to be detectable. The apparent absence of these genes from Ctenophores is therefore strictly consistent with the homology hypothesis: if they were in fact present and evolving the same way in Ctenophores as they are in Bilaterians, they would potentially not be detected by BLASTP, and so their absence is not indicative of processes beyond those included in my evolutionary null model, i.e. presence and uniform evolutionary pattern in all species.

Previous analyses concluded that convergence of Bilaterian and Ctenophore mesoderm was likelier on the basis of the large number (12) of Bilaterian mesoderm genes lacking apparent homologs in Ctenophores [136, 147]. This analysis shows that, when the possibility of detection failure is controlled for, the available data is much more consistent with the homology of mesoderm than previously believed. Only 3 genes both lack apparent homologs and are expected to be detected, and so are inconsistent with the homology hypothesis. The change in the balance of the evidence against the homology hypothesis before and after my analysis is heuristically depicted in Figure 3.4 as the number of genes shaded red (inconsistent with homology) vs. blue (consistent with homology).

More sensitive homology searches reveal potential homologs of Bilaterian mesoderm genes initially undetected in *M. leidyi*

The available data and abSENSE analysis for the nine Bilaterian mesoderm genes above predicted to be undetectable in *M. leidyi* if present are consistent with the hypothesis that homologs for these genes are actually present in *M. leidyi*. If this is the case, although BLAST, as expected, has failed to detect them, as in the previous chapter, a more sensitive homology

search method may yet succeed. Indeed, the 99% prediction interval for the bitscore of all nine of these genes has some probability mass above the detection threshold; although the bitscore from one homology search method does not translate directly into that of another, this suggests roughly that the sequences of homologs may not be so diverged as to be impossible to detect by any method.

Accordingly, I aimed to use more sensitive homology search techniques to identify potential *M. leidyi* homologs of these Bilaterian mesoderm genes. I first turned to HMMER, which converts an alignment of homologs of a gene into a hidden Markov model (HMM) that is used to search for homologs, thereby capturing the information about site-specific substitution patterns that is contained in the underlying alignment [149]. Because protein domains, rather than full-length proteins, display characteristic substitution preferences, it is common to use an HMM of a domain in a homology search. The Pfam database is a repository of pre-computed alignments and HMMs for protein domains [150], and so I began by using these HMMs in a HMMER search to identify homologs.

The FGF ligand is short and has only one eponymously-named domain. A search with the existing Pfam FGF domain profile against the annotated *M. leidyi* proteome immediately revealed one strong candidate homolog (ML04673a at $E=5*10^{-6}$) and another less statistically significant candidate (ML10595a at E=0.03). If these are true FGF homologs, I would expect them to be expressed during embryonic development in *M. leidyi.* I examined the expression profiles of these genes during *M. leidyi* using the Mnemiopsis genome portal (https://research.nhgri.nih.gov/mnemiopsis/) and found that this was the case.

Bilaterian Follistatin is comprised of a Follistatin EGF-like domain followed by three Kazal-type serine protease inhibitor domains. These Kazal-type domains are readily identifiable

in Ctenophores by BLASTP, but are present in many other proteins and so not diagnostic of a Follistatin homolog. A HMMER search with the Pfam profile for the Follistatin EGF-like domain identified a significant hit in the *M. leidyi* proteome: ML00905a at $E=4*10^{-3}$). This protein also has three Kazal-like domains, although the ordering of these four domains occurs in a slightly different order than in the Bilaterian Follistatin. I confirmed that this protein is expressed during development as above.

Bilaterians have three types of Troponin proteins, each playing a different role within the sarcomere: Troponin-I, Tropinin-T, and Troponin-C. Each of these contains a so-called Troponin domain, and are distinguished on the basis of additional domains. A search with the existing Pfam profile for a Troponin domain against the annotated *M. leidyi* proteome immediately identified three hits with middling statistical significance (E~0.1). However, a search with this profile against a custom database of six-frame translations of transcriptomes of other Ctenophore species, assembled from a variety of sources, revealed strong hits ($E < 10^{-40}$) in *Pulkia falcata* and *Hormiphora californensis,* suggesting that troponin was present in the common ancestor of the Ctenophores. I still sought to identify a candidate homolog in the model species *M. leidyi* in particular. Profile HMMs are built from underlying alignments of selected homologs; because the sequence properties and site-specific substitution preferences of homologs may change throughout evolution, the taxonomic source of these homologs may influence the ability of a profile to successfully identify homologs in a target taxon. I hypothesized that the including the candidate troponin homologs from the two other Ctenophore species, phylogenetically closer to the target species *M. leidyi*, could enhance the ability of the profile HMM to identify candidate homologs there. A modified HMM that included these sequences in addition to those in the existing Pfam alignment found two of the same *M. leidyi* genes as my initial search, but with

improved E-values: ML08948a at $E=10^{-5}$ and ML45846a at E=0.03. If these genes are true troponin homologs, I would expect them to be expressed in striated muscle cells in adult *M. leidyi.* I consulted an existing *M. leidyi* single-cell RNAseq dataset [151] in which striated muscle cells had been previously identified on the basis of expression of other conserved muscle components (eg tropomyosin), and found that one of these genes, ML45846a, is indeed expressed in all, and exclusively in, striated muscle cells. I were unable to identify any of the additional domains diagnostic of identity within the three Troponin proteins (e.g. the Troponin-I domain) in *M. leidyi.*

The Bilaterian Forkhead protein has three domains: a Forkhead domain, a Forkhead-N domain and an HNF-C domain. The Forkhead domain is readily identifiable in many *M. leidyi* proteins, but is contained in many members of the family that do not have Bilaterian mesodermal functions; to be more confident in protein identity within this family, I sought to find a gene that contains one of the two additional domains characteristic of Forkhead. I therefore searched with Pfam HMMs of these two domains. The existing Pfam HMM of the HNF-C domain did not find significant hits in *M. leidyi. I* then used that HMM to search against the large and taxonomically diverse UniprotKB database, which identified additional potential homologs, including many from earlier-diverging animal species not represented in the Pfam HMM, and created a new HMM from these hits as above. I then used this new HMM to search within *M. leidyi* proteins that had an identifiable Forkhead domain; I restricted the search to these because I would only consider a gene possessing both domains to be a candidate homolog, and this restriction reduces the search space, increasing search sensitivity. This search resulted in one significant hit: **ML154527a** at $E=10^{-4}$. I confirmed using the NHGRI database that this gene is expressed during

Ctenophore development. I were, however, unable to identify any genes in *M. leidyi* with the Forkhead-N domain.

The Bilaterian Noggin protein contains only a single Noggin domain, which is unique to the Noggin family and so diagnostic of membership. A search with the existing Pfam profile for the Noggin domain in the *M. leidyi* proteome immediately identified one significant hit: ML093034a, at $E=4*10^{-3}$. I expanded the alignment from which the Pfam profile was created by using it as the query in a search against the UniprotKB database, which found additional hits, including early-branching animal species not represented in the existing profile. I added these sequences to the profile and re- searched the *M. leidyi* proteome, which lowered the E-value of the previous hit to 10^{-8}. I confirmed that this gene is expressed during *M. leidyi* development.

The Bilaterian Cerberus/DAN protein family is defined by a single "DAN" domain, which is largely (though not entirely) unique to the family and thereby informative for defining family membership. A search with the existing Pfam domain against the *M. leidyi* proteome found no significant hits, but did find a significant hit in the Ctenophore *P. falcata*, as well as *Trichoplax adhaerens* and several choanoflagellate species. Inclusion of these sequences in the existing Pfam HMM found several significant hits in the *M. leidyi* proteome, with E-values ranging from 10^{-23} to 10^{-6}: ML03229a, ML03224a, ML018012a, and ML231814a. I confirmed that all of these genes are expressed during *M. leidyi* development.

For all other genes in Figure 3.4 above lacking apparent *M. leidyi* homologs, including those for which abSENSE did not predict that homologs would be undetectable by BLAST (red shading), I were unable to identify candidate homologs. I only attempted to use HMM-based methods here; further work may explore more sensitive search strategies.

The results of my attempts to find candidate homologs for these Bilaterian mesoderm genes and their consequences on the question of the homology of Ctenophore mesoderm are summarized in Figure 3.5.

Human mesoderm gene	Ctenophore homolog detected (BLAST)?		Ctenophore homolog detected (BLAST or HMMER?)
Eomesoderm	No		No
MyoD	No		No
Myogenin	No		No
Mrf4	No		No
Myf5	No		No
Sonic hedgehog	No	abSENSE-inspired homology search	No
Chordin	No		No
Forkhead	No		Yes
FGF	No		Yes
Follistatin	No		Yes
Cerberus/DAN	No		Yes
Noggin	No		Yes
Troponin	No		Yes
Snail	Yes		Yes
Brachyury	Yes		Yes
GATA	Yes		Yes
TGF-B	Yes		Yes
Smad	Yes		Yes
Myosin	Yes		Yes
Tropomyosin	Yes		Yes

Figure 3.5: Results of HMM-based homology searches aiming to find previously undetected *M. leidyi* homologs of Bilaterian mesoderm genes. The left depicts genes for which homologs were (blue) or were not (red) detected by BLAST; the right depicts genes for which homologs were (blue) or were not (red) detected either by BLAST or by my HMMER-based searches.

The number of Bilaterian mesoderm genes used in previous analyses to assess the probability of the homology of Ctenophore mesoderm with putative Ctenophore homologs thereby increased from seven (left) to thirteen (right), changing from a minority to a majority of these genes.

Ongoing experimental work aims to test role of putative *M. leidyi* homologs in mesoderm development

Ctenophores having homologs of Bilaterian mesoderm genes is a minimal prediction of the homology hypothesis, but is also compatible with a convergence scenario. The common ancestor of Bilaterians and Ctenophores may have had these genes, and both taxa could have gone on to evolve mesoderm independently, such that although the same genes are present in their genomes, different genes are involved in mesoderm generation in the two groups. Tests of whether the Ctenophore homologs of Bilaterian mesoderm genes actually play analogous roles in Ctenophore mesoderm development are therefore essential for assessing the likelihood of homology.

We are currently working to perform such tests using two experimental approaches. The first is to perform scRNAseq on *M. leidyi* embryos to determine whether these genes are expressed in the mesoderm lineage during development; a positive result would support the hypothesis that they have mesoderm-related functions. I plan to identify the presumptive mesodermal lineage based on its expression of genes known from previous RNAseq data to be expressed in adult muscle tissue [151]. Data from whole-embryo RNAseq developmental timecourse experiments have shown that some of these, like Tropomyosin, are expressed as early in development as 2 hours post fertilization [152], raising the possibility that they can be used as markers for the early mesodermal lineage, even before gastrulation occurs (around 4 hours post fertilization), when many crucial developmental genes involved in mesoderm specification, including many in the list above, would be expected to be expressed. These experiments are underway in the lab of Mark Martindale (Whitney Lab for Marine Biosciences), which has substantial expertise in Ctenophore experimental techniques.

The second is to determine whether phenotypes resulting from knockouts of Bilaterian mesoderm genes in *D. melanogaster* can be rescued by expression of the corresponding *M. leidyi* homolog; a positive result would strongly suggest a shared function of these homologs. These experiments are underway in the lab of Kate Beckingham (Rice University).

There are clear downsides and limitations to these approaches. The first approach runs the risk of being unable to confidently identify the mesodermal lineage. scRNAseq alone cannot unambiguously identify the mesodermal lineage at early developmental timepoints in the absence of unambiguous cell type markers, which do not yet exist in *M. leidyi.* It is my hope that I can use markers from adult mesoderm derivative tissues, like muscle, to identify their progenitor cells at an earlier stage, but it is possible that different lineages express these markers at earlier developmental times, such that this approach will identify the incorrect lineage. I hope to perform sufficiently dense temporal sampling across development (every ~1-2 hours for the 24 hours of development required for the adult body plan to emerge) that I can overcome this problem, using a transitive approach: using the adult markers to identify the corresponding cell type at an earlier stage of development, identifying a new set of markers in that cell type, and then repeating the process to reach the earliest desired time point. I hope that the short time intervals between the sampled time points will minimize risk of the markers identified in the mesodermal lineage at one time point being expressed in a different cell type at the next type point.

Pending a pilot experiment assessing practical difficulty, I are also considering a labeling experiment that would overcome this difficulty. A fate map for *M. leidyi* has been published [142], allowing the cell that gives rise to the mesodermal lineage to be identified in early embryos and injected with fluorescent dye acting as a lineage tracer. Performing this injection,

FACS sorting labeled cells at the desired later points in development, and sequencing only these labeled cells would allow independent identification of the appropriate lineage, which could then be transcriptionally profiled. The main barrier to this approach is that the large number of embryos required for such a scRNAseq experiment is practically difficult to reconcile with time-consuming manual injection of the lineage tracer into the correct macromere of each embryo.

The second approach is unlikely to yield a positive result due to the large phylogenetic distances involved in the attempted rescue experiments. There is some precedent for successful compensation at such timescales using similar types of mesodermal and developmental genes; for example, a dominant negative brachyury mutant from *M. leidyi* was shown to display the same phenotype when injected *into Xenopus laevis* embryos as the corresponding dominant negative mutant in the endogenous protein [153]. I therefore consider this approach worthwhile enough to attempt. However, on the whole, it seems relatively unlikely that even true homologs would successfully functionally compensate for one another when taken from organisms as distant as Ctenophores and Bilaterians. While a positive result would be strongly suggestive of homology, a negative result would therefore be minimally informative. I have opted to take this approach because of its relative practical ease: experiments in *M. leidyi* are difficult, while the genetics of *D. melanogaster* are comparatively quite straightforward.

There are two vastly more conceptually straightforward experiments, neither of which I have so far undertaken. The first is an *in situ* hybridization in *M. leidyi* embryos to assess the expression of the putative mesodermal homologs at the developmental stage corresponding to the developmental stage at which they are known to be expressed in Bilaterians. The second is direct disruption of these putative homologs in *M. leidyi* to assess their role in producing *M. leidyi* mesoderm. The difficulty of working with *M. leidyi* has made me disinclined to attempt either of

these myself. I have identified collaborators with expertise in the organism, but their preferred experimental approach has been the scRNAseq strategy described above, and so I have proceeded accordingly.

Of note is a very recent report of successful gene knockouts and knockdowns in *M. leidyi* using CRISPR and splice-blocking morpholinos in *M. leidyi* [154]. Future work may explore the possibility of using these techniques to directly test the functions of these putative mesoderm homologs in *M. leidyi*.

Conclusions

Here, I consider a question of broad biological interest that does not at first glance have obvious relation to the subject of lineage-specific genes: what is the evolutionary relationship between the mesoderm of two animal taxa? Prior to the availability of relevant molecular data, intuitive arguments could be marshaled for each of the two broad possibilities. A parsimony calculation that merely counts the number of events seems to favor convergence, which requires two events (two origins of mesoderm), compared to homology, which requires at least three (one origin and at least two losses). A clear counterargument contests the comparable ease of these two events: it might seem more difficult to create structures than to lose them, pointing to the higher likelihood of homology. The middling morphological and developmental similarity of Bilaterian and Ctenophore mesoderm – a level of resemblance that is suggestive but not so striking as to seem conclusive – adds to the ambiguity of the situation, and has been interpreted variously as evidence in favor of both possibilities.

Molecular data is widely regarded as uniquely informative in resolving these sorts of questions about homology. The publications introducing sequenced Ctenophore genomes

accordingly took up the question, seeking the molecular evidence that was taken to be necessary to support homology: successful detection of Bilaterian mesoderm gene homologs in Ctenophores. The lack of such apparent homologs was accordingly taken as strong support for the conclusion that convergence of mesoderm in these two taxa was the likelier scenario [136, 147].

Here, I apply my analysis method, developed in a previous chapter, to the explore an alternative possibility also consistent with the observed data: that Btenophore and Bilaterian mesoderm are homologous, and that Bilaterian mesoderm genes accordingly have Ctenophore homologs, but that these homologs are merely undetectable by homology search due to the large evolutionary distance between the taxa. I find that for the majority of genes lacking detected Ctenophore homologs, such homologs are not expected to be detectable if present.

Motivated by these results, which suggest that Ctenophore homologs could yet be lurking undetected, I undertake more sensitive homology searches in an effort to identify them. I find putative homologs for 6 new Bilaterian mesoderm genes. Some of these were found with easy, off-the-shelf homology search methods (e.g. using an existing Pfam profile in a HMMER search), but others required time and manual intervention (e.g. construction of Ctenophore transcriptome databases not available in standard sequence databases like RefSeq and so requiring manual literature searches and third-party websites to identify). The increased difficulty of these approaches is likely why they are not regularly performed by the average user of programs like BLAST and HMMER. abSENSE analysis results were key to success in these cases: knowledge that the apparent absence of a homolog is consistent with it yet being present decreased the probability that continued efforts would be wasted and so was a strong motivator to undertake them.

We argue that this new analysis changes the fairest conclusion regarding whether Ctenophore and Bilaterian mesoderm are homologous. Contrary to previous analyses proposing that homology is unlikely, I argue that the results of previous homology searches are consistent with both homology and convergence, and so both hypotheses remain viable. The discovery of six new additional putative mesoderm genes in Ctenophores supports this position.

We do find a small number (3) of genes absent in Ctenophores that are predicted to have detectable homologs. This inconsistent with the strongest form of the homology hypothesis, which holds that the full Bilaterian mesoderm program was present in the common ancestor and remained as such, unchanged and unelaborated, in all descendants. I find this strongest formulation fairly unlikely on face: over large expanses of evolutionary time, elaboration, replacement, or loss of parts of the mesoderm program seems likely. This is especially plausible given that Bilateria have mesoderm-derived tissues totally absent in Ctenophores (e.g. heart, blood), for which additional genetic programs may have evolved following their divergence from Ctenophores.

A priori, the likeliest form of the homology hypothesis seems to be an intermediate one: that some components of the mesoderm program existed in the common ancestor, while others were evolved or elaborated independently in descendant taxa. Rather than a yes/no answer arrived at by way of counting how many or which fraction of ‘mesoderm genes’ are conserved in Ctenophores, I feel that a subtler analysis taking into account the functions of conserved vs. likely actually absent genes and their roles in the overall mesoderm program will be most revealing as to the origins of mesoderm. To this end, it remains important not to conflate genes with undetected homologs as those that are definitively absent: such an error could muddy these types of functional evolutionary analyses.

Our results do not provide positive evidence of homology. They merely revive its status as a viable hypothesis consistent with the available data. Mesoderm could be convergently evolved even if the same genes are present in all species; they may have evolved mesodermal roles only in Bilateria, retaining the ancestral function or evolving another one in Ctenophores. I feel that the strongest evidence for homology is experimental evidence of shared function of these Bilaterian and Ctenophore genes. We are currently undertaking such experiments.

Chapter 4: Against all odds? The origins and functions of de novo genes

Introduction

The work described in the preceding chapters was motivated by a desire to understand novel genes, with a particular eye toward what I consider the most dramatic form of molecular novelty: de novo genes. There is much work claiming to reveal features of the origin and functions of de novo genes: the bulk of these were studies of *lineage-specific genes*, which, by way of the novelty hypothesis, equated the two. As the preceding chapters detail, inquiry into the matter has convinced me that the novelty hypothesis is not reliable, and that lineage-specific genes cannot be reliably interpreted as de novo genes, or indeed as any type of novel genes. The substantial body of work on lineage-specific genes thereby becomes uninformative for understanding de novo genes.

So: *what do we actually know* about de novo genes? In lieu of my own experimental inquiry into this question, I have here tried to systematically survey published results that bear on this question, and especially on a particular conundrum inspired by Francois Jacob's famed argument for the implausibility of these genes. What follows is a review that reflects my current understanding of the question, with an emphasis toward including only results that strike me as reliable, based on the lessons learned in the first two chapters.

Jacob's unanswered argument for the implausibility of de novo genes

As is almost proudly noted in introductions of most papers on the subject, the birth of new genes out of non-genic sequence – the "de novo" birth of genes -- was at first largely regarded to be essentially impossible. A canonical example comes from Francois Jacob's famous discussion of evolutionary novelty, "Evolution and Tinkering," in which he argues broadly that new functions in evolution come about through tinkering with components of existing functions, rather than emerging intact from wholly new cloth. Amongst his discussion of this general stance on novel functions and structures writ large, there is a brief but specific claim about the prospect of de novo genes in particular. As he put it in his now oft-cited discussion: "the probability that a functional protein would appear de novo by random association of amino acids is practically zero" [3].

Since then, documented cases of de novo genes have provided direct evidence that this probability is *not* zero. Jacob's claim about the implausibility of de novo genes has turned out to be wrong. Nonetheless, it remains compelling, voicing an intuition that I still share: that a protein comprised of largely unselected sequence is enormously unlikely to be biologically functional. What drives this intuition? I propose the following formalization of Jacob's argument. He did not put it this way, but it seems to make precise what I have found compelling in it.

Premise 1 ("Sparsity"): A very small fraction of all possible sequences has biological function.

Premise 2 ("Fair play"): The sequences available for de novo gene birth are a random and unbiased sample of all possible sequences.

Premise 3 ("Limited trials"): The number of sequences assessed for function during evolution is modest.

Conclusion: A functional gene evolving from unselected sequence is very unlikely.

Where does this argument go wrong? Its constituent premises reveal at least three (not mutually exclusive) possibilities.

The first: my intuition about the rareness of biological function in among random sequences is wrong. Function may be much easier to get – more abundant in sequence space -- than we think.

The second: de novo genes emerge from sequences that are *not* a random sample of sequence space, and as a result are in some way enriched for function.

The third: although the first two premises hold, and so the *fraction* of unselected sequences that are functional is extremely low, the *number* of such sequences tested by evolution is so high that de novo gene birth is nonetheless appreciable.

What do we know about which of these possibilities surmounts Jacob's argument?

Try, try again: ample evolutionary trials enable de novo gene birth

Part of the answer is already clear: recent work has demonstrated that the limited trials premise is false. Evolution has available to it a huge number of unselected sequences, each representing a trial from which a de novo gene could spring. Given their sheer number, even if only an extremely small fraction are successful, de novo genes should still emerge appreciably.

This abundance of unselected sequence is found in the large amount of noncoding DNA present in many genomes, especially those of "higher" eukaryotes. New technologies like RNA-seq and ribosome profiling have revealed that a surprisingly large amount of this sequence, despite lacking obvious functionality and not being under detectable selective constraint,

undergoes appreciable “pervasive” transcription and translation [155, 156]. Recent comparative work additionally suggests that which such sequences are expressed shifts fairly quickly across evolution [90, 91, 157]. These observations suggest strongly that the total number of unselected sequences tested for function is vast, overwhelming Jacob’s problem.

What of the other two possibilities?

Where do de novo genes come from, and what do they do?

The first and second possibilities concern the functions that de novo genes perform and the sequences from which they emerge.

Is biological function extremely rare in sequence space? Or is it more abundant than we think? If the latter, how and why? Are one or a few quite specific functions much more common than we appreciated? Or are we more uniformly wrong, such that all sorts of functions are actually fairly common? In either case, can we offer any molecular explanation for the true map between sequence space and function space?

Do de novo genes emerge from a representative sample of unselected sequence? Or do they preferentially emerge from a biased subset of sequence space that may be enriched for function? If the latter, what is this bias, and for what types of biological function is it enriched?

One way to answer these questions is to look to known examples of de novo genes, and ask where they came from and what they do. Enrichment in these genes for having emerged from particular kinds of sequences, or for having particular functions, would be strongly suggestive.

A case study approach

Much work has aimed to systematically identify and characterize de novo genes, which other reviews have comprehensively summarized [34, 46]. There are two substantial limitations to these results. The first is that most published studies use methods that lineage-specific genes as synonymous with de novo genes, an approach that I believe to be unreliable because it includes a significant fraction – potentially even a large majority -- of genes that are *not* de novo [97] [82]. The second is that even studies that avoid this first issue very rarely study the actual *functions* of de novo genes. Instead, they loosely infer broad features, often from accessible but weakly informative proxies like expression pattern. This type of information is not well suited to our question, which asks, in molecular and cellular detail, what these genes do.

Here, I instead take a case study approach to understand the origin and functions of de novo genes that differs from most studies in two main aspects. I consider only genes whose de novo status is fairly clear-cut, and for which there is direct experimental data regarding their function. The number of such genes is very small. Nonetheless, as of the last few years, I feel that there is a critical mass such that an accounting may suggest the outlines of answers to Jacob's unresolved problem.

While this review aims to be in part an objective summary of these case studies, the subsequent interpretation is speculative, as any analysis based on so few cases must be. My primary aim is to draw attention to this interesting and unanswered problem, to encourage further work, and to offer a few hypotheses.

Criteria for inclusion of a gene in this review

There has been a lack of consensus about the burden of proof that should be required to demonstrate that a sequence is truly a de novo gene, as opposed to, for example, a gene with

homologs that are present but undetected in other species, or a sequence that is not actually functional in the organism under consideration (and therefore not, at least by my definition, a gene). Beyond making clear the criteria that I use to consider a sequence likely enough to be a true de novo gene to include it here, I will not consider this issue. For the interested reader, the question has been reviewed elsewhere [46, 158].

First, there must be direct evidence of the gene's absence in outgroup species, as should be the case if the gene truly emerged de novo in the focal species. For RNA genes, the orthologous genome sequence must be identified in outgroups and shown experimentally or inferred computationally (e.g. on the basis of absent promoter elements) not to be transcribed there, or to likely have an extremely dissimilar transcript structure (e.g. on the basis of absent splice sites), making shared function unlikely. For protein-coding genes, the orthologous sequence must be identified in outgroups and shown experimentally not to be translated there, or shown computationally to lack an ORF of comparable length or coding sequence.

Second, there must be at least two outgroups for which this first criterion is true. This is the minimum number of such outgroups required to make de novo origination a likelier scenario than the alternative of gene loss, assuming -- generously -- that the two are equally likely.

Finally, the gene must be directly shown to be functional. This excludes the many cases of computationally-identified genes for which there has been no experimental follow-up. It also excludes genes for which *inferences* about function have been made on the basis of various features, like expression pattern, but for which there is no *direct evidence* of function, like a knockout phenotype. Moreover, if it is a protein, as opposed to the transcript potentially encoding that protein, that has been proposed to be de novo, there must be some evidence that the functional effect indeed occurs at the protein level.

By standards of the field, these criteria are conservative. They are also still imperfect. For example, the possibility of noncanonical, 'leaky' translation beyond the bounds of conventional ORFs leaves open the possibility that an ORF whose sequence has a requisite early termination codon in an outgroup may yet be translated and functional, despite appearing otherwise on the basis of sequence analysis. Ideally, the absence of the protein product or of translation would be experimentally shown. But such tests are usually difficult and so are rarely done.

On the flip side, the stringency of these criteria likely lead to biases. One example is that a large proportion of these genes come from humans and have roles in cancer. This is unlikely to be due in its entirety to the special enrichment of de novo genes for roles in human cancer (the contribution of this effect will be discussed later), but at least in part to the increased probability that genes with these properties have been functionally characterized. Another is that the taxonomic source of the genes considered here is restricted to vertebrates and yeast. This is not an intentional selection criterion on my part; it just so happens that the small number of genes that meet these criteria do not completely fill out the tree of life, and so are absent from, for example, bacteria.

I have chosen to these criteria not because they are bulletproof or without drawback, but because I believe they constitute a balance of the evidence making such cases likelier than not to be true de novo genes, and to carry an acceptable degree of functional and taxonomic bias in doing so.

Case studies: origins and functions of known de novo genes

RNA genes

H. sapiens ELFN1-AS1

ELFN1-AS1 is an RNA gene purportedly unique to humans, found antisense to and contained entirely within an intron of the conserved gene ELFN1. I include it here as I find the claim of de novo origination plausible on balance [159], but flag it as the most tenuous of these cases, as the evidence for the absence of a functional homolog in outgroups is somewhat weak [159].

ELFN1-AS1 is expressed in a variety of tumor types, and only at low levels in normal tissue [159]. In colon cancer, knockdown has been shown in vitro to decrease cell proliferation and migration. Possible high-level mechanisms underlying this effect include its observed reduction of Erk activation, decrease of vimentin expression, and increase of E-cadherin expression [160]. These may in turn be due, at least in part, to ELFN-AS1 acting as a competitive sponge for miRNAs targeting TRIM44 [161], a gene shown to be capable of inducing the EMT to facilitate cancer progression [162]. ELFN-AS1 knockdown was also shown in *in vitro* esophageal cancer cells to decrease cell proliferation and migration, again likely by competitively sponging miRNAs. Here, these seem to be miRNAs targeting GFTP1, whose functional role in downstream oncogenesis is not known [163]. A similar story was again found in ovarian cancer, where ELFN-AS1 seems to sponge miRNAs targeting CLDN4 [164], a gene known to be involved in cancer cell migration and invasion [165].

M. musculus Poldi

Poldi is a noncoding RNA present in several *Mus* species, including *musculus*, and likely emerged in the lineage around 3Mya. In *M. musculus*, it is expressed specifically in the postmeiotic round spermatids of the seminiferous tubules. Knockout mice exhibit somewhat

decreased sperm motility and testis weight. The mechanisms underlying this phenotype were not investigated [166].

Protein-coding genes

S. cerevisiae MDF1

MDF1, one of the first functionally characterized de novo genes [115], is a 153 amino acid protein found only in the budding yeast *S. cerevisiae,* in which it likely originated de novo within the last few million years.

The overall biological role of MDF1 is to increase vegetative growth in haploid cells [115], which it does via at least two mechanisms affecting two different conserved pathways.

The first is repression of the mating pathway, which in turn increases vegetative growth due to downstream interactions between these pathways. MDF1 binds to the transcription factor MATalpha2, which, in diploids, dimerizes with another transcription factor MATa1 to cooperatively repress (unnecessary, in diploids) haploid-specific mating pathway genes. In alpha-type haploids, MATa1 is not expressed, seemingly leaving MATalpha2 without the necessary binding partner to perform this repression of the haploid mating pathway. MDF1 fills this void: in alpha haploids, it dimerizes with MATalpha2, forming a complex that binds to the same mating pathway promoters as the MATa1-MATalpha2 dimer does in diploids and repressing them in like fashion. Overall, MDF1 seems to function by acting as a stand-in for an absent MATa1, supplying the interaction partner necessary for MATalpha2, which binds DNA only weakly and so requires a binding partner [167], to perform an established regulatory role but in a novel condition.

The second route by which MDF1 increases vegetative growth is by shortening lag phase. This may occur through MDF1 binding to and repressing the kinase SNF1, preventing its glucose repression of fermentation genes and thereby speeding fermentative growth [168].

The regulation of MDF1 has an interesting property. MDF1 is antisense to and partially overlaps the conserved gene ADF1. The ADF1 protein, a transcriptional repressor, binds to the promoter of MDF1 to suppress its transcription. MDF1 is therefore regulated by the protein product of the gene that it happens to lie beneath [115]. This strikes one as a bizarre coincidence: such regulation would be natural if mediated by the complementarity present at the RNA level, but evidence that it is at the protein level is strong.

H. sapiens NYCM

NCYM is a 109 amino acid human protein, named for its location antisense to and overlapping the oncogene MYCN. It likely emerged either uniquely along the human lineage or prior to the split with chimpanzees. It is arguably the best-studied of all de novo genes; a much more thorough review of its regulation and function can be found in [169].

Among other cancers, MYCN is amplified in human neuroblastoma, where like other myc oncogenes it enhances cell proliferation and survival [170]. NCYM is found to be co-amplified with MYCN in neuroblastoma, where it too has been shown to enhance proliferation and survival [171].

This effect is at least partly via NYCM acting on MYCN. Knockdown experiments show that NCYM stabilizes the MYCN protein in a proteasome-dependent manner. Data suggest that NYCM may physically bind to one or both of MYCN and GSK3B, a kinase that phosphorylates

and thereby targets MYCN for proteasome degradation, to inhibit this phosphorylation step and prevent degradation.

NYCM may also stabilize MYCN levels by globally repressing GSK3B: through unidentified mechanisms, it seems to promote inhibitory phosphorylation of GSK3B by the mTOR pathway [171]. This repression of general GSK3B function is thought to stabilize other substrates destabilized by GSK3B that promote oncogenesis, including beta-catenin. NYCM may therefore also have oncogenic activity through this MYCN-independent mechanism.

Finally, NYCM may enhance MYCN activity by increasing the amount of a cleaved form of MYCN ("Myc-nick"), which localizes to the cytoplasm and promotes tubulin acetylation. Data again suggest that this may occur through NYCM's direct binding to the complex between MYCN and calpain, the protease that performs the cleavage [172].

The regulatory interactions between NYCM and MCYN are complex and intriguing. NYCM expression is positively regulated by the MYCN *protein* via an E-box promoter element within the NYCM gene body [171]. This site is the same one used by MYCN to positively regulate its own expression [173], its dual use enabled by the palindromicity of the E-box and the overlap between the two genes. That MYCN protein is in turn stabilized by NYCM makes this a feed-forward regulatory loop. Moreover, this interacts with another regulatory loop in which MYCN and the neuroblastoma-expressed reprogramming factor Oct4 promote each other's expression. And MYCN also promotes the expression of the other reprogramming factors Nanog and Sox2, which together increase the multipotency of stem-like cancer cells like neuroblastoma. NYCM, embedded in these networks through its effects on and from MYCN, seems to increase the levels of all of these factors [174].

NCYM also seems to have noncoding functions. The NCYM *transcript*, in addition to the protein, also seems to increase transcription of MYCN, potentially through several mechanisms. The NYCM transcript has been shown to increase CTCF-mediated transcription of MYCN, possibly through direct CTCF binding and recruitment to the MYCN promoter [175]. In virtue of its complementarity, the NYCM transcript also binds directly to the MYCN locus and to MYCN RNA, potentially regulating transcription levels, promoter choice, and alternative splicing or other forms of RNA processing [176-178]. Although the NCYM peptide emerged in a subset of the primate lineage, the locus itself, and potentially the transcript and these noncoding functions, is common to mammals [171].

H. sapiens PBOV1

PBOV1 is a 135 amino acid human protein first characterized as being overexpressed in prostate and breast cancer [179]. It is antisense to and lies entirely within the intron of the conserved gene BIG3. It likely originated de novo in humans or possibly hominid primates.

In prostate cancer cells [180], PBOV1 overexpression and knockdown lead to increased and decreased, respectively, cell proliferation and anchorage-independent growth. Overexpression also increases progression through the G1-S checkpoint, the proposed cause of this proliferative effect. This may occur via the reduction in levels of G1-S transition inhibitors p21 and p27, and increase in levels of activators cyclin d1 [180, 181] and phosphorylated Rb, that PBOV overexpression causes though undetermined mechanisms.

Similar results were found in hepatocellular carcinoma cells [181], where in addition PBOV overexpression and knockdown was found to respectively increase and decrease tumor

cell migration in vitro and metastasis and vascularization in vivo. This may occur through PBOV's effects on expression of the wnt/beta-catenin regulator HIF1A, epithelial markers (alpha-cadherin and E-cahderin), and mesenchymal markers (N-cadherin and vimentin), which resemble those in found an EMT transition. PBOV was also shown to positively regulate expression of pluripotency factors OCT4, Nanog and c-Myc. Mechanisms for these effects are unclear. Finally, PBOV was also found to increase beta-catenin signaling, possibly in part by direct binding that prevents inhibitory phosphorylation by GSK3B and allows more efficient nuclear localization.

In ovarian cancer in vitro experiments, somewhat opposite results, in which PBOV *negatively* regulates cancer cell proliferation and tumorigenesis, were observed [182]. Another dissimilar role was found in monocytes, where PBOV overexpression was found to *decrease* progression through the G1-S checkpoint. It was also found here to increase monocyte differentiation into macrophages [183]. Mechanisms were not explored in these studies.

H. sapiens MYEOV

MYEOV is a 313-amino acid protein that emerged in the human lineage [184, 185]. It was first identified in a tumorigenicity screen in gastric carcinoma, and then found to be upregulated in some multiple myelomas [186]. RNA knockdown of MYEOV in gastric [187], colon [188], and neuroblastoma [189] cancer cells in vitro has since been shown to reduce proliferation and invasion. In colon cancer, MYEOV levels were also shown to be positively regulated by prostaglandin E2 [188]. Mechanisms underlying these results are unknown. Although endogenous MYEOV protein has been detected in non-cancerous tissue [184] and shown to be translated from transfected plasmids in cell lines [190], these experiments did not

assay for the protein product or attempt to determine whether the protein or the transcript was responsible for the observed effects.

In pancreatic cancer, increasing MYEOV expression was shown in vivo to increase metastasis and tumorigenesis. Here, there is strong evidence that the MYEOV protein binds to the transcription factor SOX9, which among many other regulatory roles has been shown to form a homodimer to promote expression of HES1 in breast cancer [191], at a regulatory element similar to those used in its more standard biological role in chondrogenesis [192]. The resulting MYEOV-SOX9 heterodimer cooperatively binds the HES1 enhancer to increase its expression, resulting in the observed oncogenic phenotypes. Although the MYEOV protein was shown to bind to the HES1 enhancer on its own, binding was enhanced by SOX9, and increases in HES1 expression were shown to depend on SOX9 [193]. This suggests that MYEOV serves as a necessary interaction partner, akin to the second molecule of SOX9 that forms the dimer in other physiological contexts, that enhances insufficient baseline levels of SOX9 binding to drive effective transcription factor activity.

In non-small cell lung cancer, however, the MYEOV protein was shown not to be expressed, and the MYEOV transcript shown in vivo to enhance tumor invasion and metastasis in a protein-independent manner. There is strong evidence that the MYEOV transcript acts as a sponge for one or more miRNAs that target two components of the TGF-B/SMAD pathway, USP15 and TGBR2, thereby increasing TGF-B signaling activity and promoting an epithelial-to-mesenchymal transition [194].

MYEOV thus also seems to have both coding and noncoding function. Although the protein ORF is unique to humans, the locus itself is present in most primates [185], and transcription of the region has been detected in chimpanzee. [184]

Northern gadid *AFGP*

The antifreeze glycoprotein *AFGP* is essential to the ability of Arctic codfishes (gadids) to survive the cold temperatures of their native environment: it is secreted into the blood, where it inhibits the formation of ice crystals [195]. AFGP evolved de novo in the gadid lineage about 3 Mya, presumably in response to the Pliocene glaciation event that occurred at that time [125].

The protein consists of a large and variable (20-500) number of Thr-(Ala/Pro)-Ala repeats, preceded by a signal peptide allowing for its secretion and a small glutamine-rich propeptide that seems to be removed post-translationally [125]. The threonine in each repeat is glycosylated with an O-linked N-acetyl-D-galactosamine [195], which presumably occurs via the standard O-glycosylation pathway, taking place in the Golgi and specified by the amino acid sequence [196].

AFGP is currently the only de novo gene for which there is detailed experimental data regarding both protein structure and biological function. A recent 2D infrared spectroscopy-based study indicates that AFGP seems to exist in an ensemble of conformations: mostly random coils, polyproline-II helices, and alpha helices [197]. The same study found that application of compounds that inhibit or enhance the antifreeze effects of AFGP do not significantly alter protein structure, implying that, consistent with previous results [198], its antifreeze activity is due largely to ice crystals interacting with hydroxyl groups on the disaccharide additions, rather than with the peptide backbone or side chains themselves.

While there have been many independent origins of proteins with antifreeze function [195], AFGPs in polar fish are an example of particularly precise evolutionary convergence. Antarctic notothenioid fish independently evolved an antifreeze glycoprotein with essentially the

same sequence as the arctic gadid AFGP. The notothenioid protein is, however, not de novo in origin: its 5' end, including the secretory signal, and part of its 3'UTR -- though not its repetitive coding sequence -- were derived from a trypsinogen-like protease [199].

Where did Jacob's argument go wrong?

I now speculate on what these case studies tell us about how de novo genes solve Jacob's problem: which of its premises have errors, what these errors are, and why I fell into them.

Some functions require only small regions of Watson-Crick complementarity and so are common in sequence space

Two examples, MYEOV and ELFN1-AS1, work by reducing miRNA suppression of existing cellular pathways. They do this by competitively binding to these miRNAs, "sponging" them up and thereby depleting the amount of free miRNA available for suppression.

The sequence requirements to act as a miRNA sponge are minimal. Generally, an exact match of only 6-8 nucleotides is sufficient complementarity for an miRNA to bind a target [200]. miRNA sponges are thus an excellent example of a function for which the sparsity premise fails. Sequence space should be rife with them.

A similar argument should hold for other biological functions for which small numbers of complementary nucleotides are sufficient. These should include miRNAs themselves. That no miRNA genes have been found to be de novo originated likely reflects ascertainment bias more than biological reality.

Overlap with a conserved gene lowers the barrier to expression

The functionality of a gene depends not just on the sequence of its product, but also on how it is regulated. The overall abundance of function in sequence space depends on both of these features: regulatory elements enabling the gene to be transcribed and translated in a context in which it is functional must also be hit upon. While recent experimental work suggests that at least minimal regulatory sequences are not terribly uncommon [201, 202], proper regulation nonetheless remains a hurdle that must be surmounted en route to de novo birth.

Four of the genes described here (NYCM, MDF1, PBOV1, and ELFN1-AS1) overlap conserved genes, consistent with previous reports about de novo genes [203, 204]. As noted before [203], this likely at least in part reflects ascertainment bias, as it is easier to demonstrate de novo emergence in the manner that I have required here for genes that overlap such an "anchor gene." But this may also be a genuine feature reflecting a solution to Jacob's problem. Overlapping an existing gene provides an easy route to expression. It allows the co-option of nearby features already in place to drive transcription of the conserved gene, like regulatory elements and open chromatin, removing the need for the gene to find its own such elements in sequence space [203].

Two examples described here, NCYM and MDF1, are also transcriptionally regulated by the *protein products* of these overlapping genes. While at first this seems an enormous coincidence -- why, *of all the genes in a genome*, should a protein happen to regulate its new downstairs neighbor? – we have seen in one case a parsimonious explanation. The same palindromic regulatory sequence used by MYCN in autoregulation, which presumably predates the emergence of NYCM, falls conveniently within this new gene, driving its regulation. In the case of MDF1 and ADF1, there is no evidence for this: ADF1 has not previously been noted to regulate its own expression. But this could nonetheless be the case, as ADF1 is poorly-studied;

alternatively, such regulation may have been ancestrally present and since lost. I speculate that this striking observation of protein-based regulatory interactions between de novo genes and their conserved neighbors may just be another form of new genes taking advantage of regulatory elements already established in and around existing genes.

Here is a clear example of the failure of Jacob's premise of fair play. Many de novo genes do not merely happen to hit upon regulatory elements as they traverse a random sample of sequence space. Instead, they avail themselves of perhaps the most decidedly nonrandom sequences in the genome: existing genes and the regulatory structures that drive them.

Noncoding function lowers the barrier to coding function

Two examples, MYEOV and NCYM, function both as proteins and as noncoding RNAs. In both cases, the protein has a recent de novo origin, but the locus itself is largely conserved in outgroups. MYEOV in fact seems to be transcribed in chimpanzee. This raises the possibility that functionality as a noncoding RNA predates the emergence of the protein.

That protein-coding genes may emerge within noncoding RNAs has been previously proposed [91, 92, 184, 205], and represents another violation of Jacob's fair play premise. Rather than emerging from unselected sequence, and so needing to find the means of their own expression and regulation in sequence space, de novo proteins get these for free by emerging from within noncoding RNAs.

Many non-genic, non-functional ORFs lying within transcripts are translated at low levels [91, 155]. This exposes their sequences to selection, and so may cause selection to alter them in ways that, while not immediately productive of what I would call function, nonetheless bias them toward areas of sequence space in which function is more abundant. The best-supported such

example is the finding from Joanna Masel and colleagues that selection seems to act on such lowly-translated ORFs to "preadaptively" reduce harmful aggregation propensity [53]. Selection on sequences resulting from low-level pervasive expression, insofar as it alters these sequences in these or other currently unappreciated ways, could bias the region of sequence space from which they emerge toward those enriched for biological function, again violating Jacob's premise of fair play.

Though I am not sure what to make of the observation, I would be remiss to end a discussion of the potential of noncoding function to facilitate the emergence of coding function without noting what feels like another striking coincidence. For both MYEOV and NCYM, the functions of the protein and the noncoding transcript are very similar -- despite being performed by entirely distinct molecules. In the case of NCYM, these two functions have identical molecular-level effects: through totally orthogonal mechanisms, both the protein (itself apparently acting in at least two independent ways) and the noncoding RNA increase cellular levels of MYCN. In the case of MYEOV, the similarity is at a higher level: acting via distinct signaling pathways, both the noncoding RNA and the protein promote the epithelial-to-mesenchymal transition [193, 194].

Assuming that this observation is not an experimental error resulting from failed controls of some kind, why these proteins should have the same effect as the noncoding RNAs whose locus they coinhabit is unclear. It does not seem that this must be so. While the locus as a whole clearly receives a larger selective benefit from housing two beneficial functions than just one, there is no obvious need for those functions to be functionally similar. To speculate, one possibility is that some feature of being encoded at the same locus biases the two molecules to evolve similar functions: for example, shared expression timing, transcript localization, RNA

binding proteins, or other interaction partners. Another is that using the locus to produce a new protein reduces the amount of noncoding transcript that is available to perform its function – its transcripts soaked up by translating ribosomes, or the locus occupied by polymerases making a different isoform – thereby putting pressure on the new protein to functionally compensate for this loss. Yet another is that there *is* a selective advantage to functional similarity: a positive epistasis in the pathways in which these genes act, in which the benefit of the functions from both molecules is greater than their sum.

We might fruitfully think of all of these examples as altering the map between sequence space and function space depending on one's physical location in the genome. Certain sequences may confer a higher selective benefit when found at one locus than at another: for example, loci currently occupied by noncoding RNAs may be particularly advantageous. Considered this way, the availability of noncoding transcripts as a substrate for de novo protein birth is another failure of Jacob's fair play premise. De novo genes can emerge on sequences enriched in function to avoid the difficulty of an unbiased search of sequence space.

Abundant "attractors" in sequence space increase the probability of function

In using the metaphor of sequence space to discuss de novo gene birth, I have implicitly considered a new gene being, all at once, assigned to a point in that space. If this point corresponds to a functional sequence, gene birth is successful; if not, it fails. In reality, neutral sequences perform a random walk through sequence space as they mutate over evolutionary time. Mutation thus defines a relationship between points in sequence space that has implications for would-be genes occupying them: a nascent gene occupies its current point in sequence space

and also has some probability of transitioning to every other point, defined by the mutation spectrum.

A gene need not hit upon a functional sequence 'immediately': having a sufficiently high probability of being mutated into a functional sequence will suffice. I can think loosely of the points in sequence space that are not themselves functional, but that have a comparatively high probability of being mutated into functional points, as "attractors": if a gene hits upon one, it is likelier to be "drawn into" the function.

The probability of a new gene being functional is not just the fraction of sequence space that is itself functional, but should also include the fraction of sequence space that are such attractors, weighted appropriately by the probability of successful attraction. For some functions, while abundance per se in sequence space may be quite low, the abundance of its attractors could be high – and so it is not actually so unlikely to be found by a new gene.

One example of such a function is miRNA sponging. I omitted a wrinkle in the above argument about their abundance in sequence space. To meaningfully reduce the effects of a miRNA, a sponge must offer enough binding sites to appreciably deplete the cellular miRNA pool, which is often appreciable. In addition to a high expression level, this can be achieved by having multiple binding sites per transcript, as in the case of MYEOV [194]. The existence of a *single* binding site *is,* as discussed above, highly abundant in sequence space. And while not sufficient for function in a lowly-expressed transcript, it is likely an attractor to it: mutational mechanisms like tandem duplication and unequal crossing over at repeat elements make large numbers of repeats fairly easily accessible. So, even if sequence space is not quite rife with fully-operational miRNA sponges, it is nonetheless likely full of "attractors" to them.

Another example is AFGP, in which many repeats of the Thr-Ala/Pro-Ala motif are essential for function. A single repeat is unlikely to act as an effective antifreeze (otherwise, many proteins would have such activity), but a sequence containing one such repeat is, as above, an attractor to it. Indeed, analysis of outgroup genomes allowing reconstruction of the evolutionary history of the AFGP locus reveals that this is precisely what happened: a single ancestral repeat element was extensively duplicated to produce the functional protein [125]. Though not de novo originated, the convergently evolved notothenoid AFGP protein also produced its many repeats by duplicating a single ancestral element. This shared history suggests that this sequence is indeed a strong attractor for antifreeze function.

The abundance of attractors in sequence space for these, and likely other, functions represents a failure of Jacob's sparsity premise: when considered, function becomes more abundant in sequence space.

The cellular context offers many "freeloader functions" that require little more than binding and are abundant in sequence space

Certain types of biological function are notably absent from the examples here. These includes enzymes, motors, and any function primarily effected by the gene *itself.* Instead, all of these examples produce their effects solely by way of interacting with existing cellular components to modulate or co-opt their established functions.

Moreover, although data are limited, the physical sophistication of these interactions seems to be quite low. In contrast to the sorts of finely-tuned molecular motions performed by many genes – which are precise, regulated, dynamic – the molecular mechanism of these de novo genes seems to be *merely to bind* to their interaction partners. Sponges *bind to* miRNAs,

perturbing their existing effects on other genes. The NCYM transcript *binds to* complementary RNA, altering the output of existing transcription and translation. The NCYM protein *binds to* MYCN and other proteins, modulating existing interactions to alter their effects. MDF1 and MYEOV *bind to* transcription factors and to their DNA binding sites, facilitating their partners' existing function at the locus. AFGP *is bound by* existing glycosylation and export machinery, giving it the sugar additions and localization necessary to *bind to* ice crystals.

These genes function only a) by modulating existing functions of other cellular components and b) doing so though molecularly unsophisticated 'mere binding.' I refer to such functions as "freeloader" functions. I propose that freeloader functions are dominant in sequence space, and so are preferentially found by de novo genes. This is supported anecdotally by the examples here, but can I offer an explanation of why it should be so?

The distribution of function in sequence space cannot be defined in a vacuum. It depends on the cellular environment in which the sequences exist. Some functions are comparatively 'self-powered' and independent of this environment, like many catalytic reactions: with the substrate present, many enzymes can do their jobs entirely on their own. Other functions are much more highly dependent on the details of the cellular environment, requiring the presence of many interaction partners, cofactors, local physiological contexts, and so on.

For each existing functional component in the cellular environment, there is a corresponding repertoire of dependent functions: those that can be achieved merely by *altering that existing function.* The enormous complexity of modern cells, brimming with thousands of components carrying out countless functions, thereby offers a correspondingly huge number of *dependent* functions.

How abundant are these dependent functions in sequence space? Are they particularly easy to come by, or are they as rare as more canonical independent functions? It is easy to imagine that many of these dependent functions can be achieved by *mere binding*: interacting with existing cellular components seems like an easy route to impede, alter, or enhance their native functions, resulting in a downstream effect. Mere binding, in turn, seems likely to be quite abundant in sequence space. All that must be done is sticking. And, arguably, existing proteins, in virtue of often having large numbers of interaction partners, have been selected to interact with other proteins, priming them for being clung to by new, coarse, unselected partners.

If these premises are true – that the complicated cellular environment offers many dependent functions; that many dependent functions can be achieved by mere binding; and that mere binding is common in sequence space – then functions with these properties should be abundant in sequence space. That is, *freeloader functions* should be very abundant. Moreover, I argue that they are much more abundant than the independent functions with which I are more familiar, and are in fact the *most abundant* type of function in sequence space. This would explain why de novo genes are so enriched for these functions: they are closer to their birth, and so a better (though, as I argue here, not perfect) representation of the unbiased distribution of function in sequence space, than are older genes, which have had longer to slowly traverse sequence space and be led by selection to the rare patches of highly functional space within it.

Although the aforementioned discovery bias may well be operating, with conservation of a gene being arguably a less important criterion in being considered important enough to study in cancer biology than other fields, and human cancer genes comparatively likely to be studied conservation notwithstanding, the particular abundance of freeloading may also help to explain why so many de novo genes function in cancer. To be selected, a gene must exert a *beneficial*

effect. Since existing organisms are fairly optimized, any new interaction is on average likely to *decrease* fitness. This seems especially true in complex multicellular organisms: what could be beneficial for a single cell may be detrimental to the organism as a whole: beneficial effects are especially rare and hard-fought, needing to perform a tenuous balancing act. Especially in these cases, freeloader functions are likely to be harmful, and the corresponding sequences to be selected against. In cancer, of course, this breaks down: the whole process is one of functions that are beneficial for the cell but detrimental to the organism. It should not be a surprise, then, that if freeloader functions are dominant in sequence space, and if they are mostly deleterious, new genes to preferentially emerge here.

The abundance of freeloader functions represents a failure of Jacob's sparsity premise. Function is indeed much more common in sequence space than one might imagine. I at first found sparsity intuitive because when I imagine function, what springs to mind is a *very particular* kind of function: the shining examples of molecular sophistication by which we are most awestruck, and which we therefore spend most of our time studying. This bias has led us to overlook the less glamorous and less impressive, but much more abundant, repertoire of freeloader functions. In conceiving of sequence space, I did not properly appreciate that genes find themselves not in the near-vacuum in which my beloved 'self-powered' genes can freely act, but in the bustling milieu of a cell. And so I underestimated how far the coarse role of mere binding might take us.

Final thoughts and outlook

Here I have reviewed what is known about the origins and functions of the very small number of de novo genes for which such information is available. Absent any broader conclusions, I hope such a survey is interesting and informative in its own right.

I have also used these examples to speculate on the question with which I opened: where and why did Jacob's argument for the improbability of de novo gene birth -- and those, like myself, who nodded along -- go wrong? A proposal of particular interest came from considering what these de novo genes suggest about the distribution in sequence space of function in general, and of different types of functions in particular. I hypothesize that so-called "freeloader functions" are extremely common in sequence space, and that de novo genes therefore preferentially access them as a result.

This hypothesis can be empirically tested. Increasing the number of de novo genes with characterized functions would support or undermine it. Experiments testing binding activity of random sequences would do the same, as this hypothesis requires that binding is fairly common in sequence space. Indeed, random sequence experiments in general are extremely relevant to the study of de novo gene birth, independent of any particular hypothesis. There are to date relatively few informative studies on the subject; much could be gained by pursuing it.

With enough such work characterizing de novo genes born at different times, I could consider 'freeloaderness' as a function of gene age. While freeloader functions seem to be enriched in de novo genes, most conserved genes seem to have more sophisticated roles. One way to reconcile these observations is to posit a Hertzsprung-Russell diagram of sorts describing the time evolution of genes: new genes are likely to be freeloaders, but, as they age and are shaped by selection, their functions are gradually refined from this crude starting point into ones requiring more molecular specificity and that are more under their own power.

Freeloader functions prompt a consideration relevant to the source of much of the interest in de novo genes: the hypothesis that, being molecularly novel, they are likely to be involved in the evolution of functional novelty, underlying the kinds of striking evolutionary innovations that so often capture my attention and imaginations [8, 24]. That de novo genes largely have freeloader functions would be pointedly ironic: at least at the molecular level, these functions are *minimally* novel. Of course, molecular novelty and novelty at higher scales need not be related. For example, morphological innovations like new appendages in animals are now well known to have evolved from redeployment of existing developmental pathways. Nonetheless, the abundance of freeloader functions among de novo genes might at least prompt a pause in which I reevaluate the widespread assumption that new genes are likely to underlie evolutionary innovation. Even in inventing, evolution may yet be a tinkerer.

Again, this is mostly wild speculation. The available data are too sparse, and from too narrow a taxonomic sample (happening to come from only vertebrates and yeast), to amount to anything else. Detailed studies of individual genes that confirm de novo birth and elucidate function are few and far between. If we have any hope of understanding the biology of de novo genes, we have much farther to go.

References

1. Keese PK, Gibbs A. Origins of genes:" big bang" or continuous creation? Proceedings of the National Academy of Sciences. 1992;89:9489-93.

2. Chothia C. One thousand families for the molecular biologist. Nature. 1992;357:543-4.

3. Jacob F. Evolution and tinkering. Science. 1977;196:1161-6.

4. Dujon B. The yeast genome project: what did we learn? Trends in Genetics. 1996;12:263-70.

5. Fischer D, Eisenberg D. Finding families for genomic ORFans. Bioinformatics (Oxford, England). 1999;15:759-62.

6. Rubin GM, Yandell MD, Wortman JR, Gabor GL, Nelson CR, Hariharan IK, et al. Comparative genomics of the eukaryotes. Science. 2000;287:2204-15.

7. Casari G, Daruvar D, Sander C, Schneider R. Bioinformatics and the discovery of gene function. Trends in genetics. 1996;128:244-5.

8. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC. More than just orphans: are taxonomically-restricted genes important in evolution? Trends in Genetics. 2009;25:404-13.

9. Nelson PA, Buggs RJ. Next generation apomorphy: the ubiquity of taxonomically restricted genes. Next generation systematics. 2016;85:237.

10. Zhou K, Huang B, Zou M, Lu D, He S, Wang G. Genome-wide identification of lineage-specific genes within Caenorhabditis elegans. Genomics. 2015;106:242-8.

11. Johnson BR, Tsutsui ND. Taxonomically restricted genes are associated with the evolution of sociality in the honey bee. BMC Genomics. 2011;12:164.

12. Sollars ES, Harper AL, Kelly LJ, Sambles CM, Ramirez-Gonzalez RH, Swarbreck D, et al. Genome sequence and genetic diversity of European ash trees. Nature. 2017;541:212.

13. Toll-Riera M, Castelo R, Bellora N, Albà M. Evolution of primate orphan proteins. Biochemical Society Transactions. 2009;37:778-82.

14. Paps J, Holland PW. Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. Nature communications. 2018;9:1-8.

15. Bowles AM, Bechtold U, Paps J. The Origin of Land Plants Is Rooted in Two Bursts of Genomic Novelty. Current Biology. 2020.

16. Hilgers L, Hartmann S, Hofreiter M, von Rintelen T. Novel genes, ancient genes, and gene co-option contributed to the genetic basis of the radula, a molluscan innovation. Molecular Biology and Evolution. 2018;35:1638-52.

17. Khalturin K, Anton-Erxleben F, Sassmann S, Wittlieb J, Hemmrich G, Bosch TC. A novel gene family controls species-specific morphological traits in Hydra. PLOS Biology. 2008;6:e278.

18. Schmitz JF, Chain FJ, Bornberg-Bauer E. Evolution of novel genes in three-spined stickleback populations. Heredity. 2020;125:50-9.

19. Babonis LS, Martindale MQ, Ryan JF. Do novel genes drive morphological novelty? An investigation of the nematosomes in the sea anemone Nematostella vectensis. BMC evolutionary biology. 2016;16:114.

20. Shigenobu S, Stern DL. Aphids evolved novel secreted proteins for symbiosis with bacterial endosymbiont. Proceedings of the Royal Society B: Biological Sciences. 2013;280:20121952.

21. Domazet-Loso T, Tautz D. An evolutionary analysis of orphan genes in Drosophila. Genome Research. 2003;13:2213-9.

22. Tautz D. The discovery of de novo gene evolution. Perspectives in biology and medicine. 2014;57:149-61.

23. Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, et al. Proto-genes and de novo gene birth. Nature. 2012;487:370.

24. Johnson BR. Taxonomically restricted genes are fundamental to biology and evolution. Frontiers in genetics. 2018;9:407.

25. Behl S, Wu T, Chernyshova A, Thompson G. Caste-biased genes in a subterranean termite are taxonomically restricted: implications for novel gene recruitment during termite caste evolution. Insectes Sociaux. 2018;65:593-9.

26. Zelhof AC, Mahato S, Liang X, Rylee J, Bergh E, Feder LE, et al. The brachyceran de novo gene PIP82, a phosphorylation target of aPKC, is essential for proper formation and maintenance of the rhabdomeric photoreceptor apical domain in Drosophila. PLoS genetics. 2020;16:e1008890.

27. Wang Y-W, Hess J, Slot JC, Pringle A. De Novo Gene Birth, Horizontal Gene Transfer, and Gene Duplication as Sources of New Gene Families Associated with the Origin of Symbiosis in Amanita. Genome biology and evolution. 2020;12:2168-82.

28. Warner MR, Qiu L, Holmes MJ, Mikheyev AS, Linksvayer TA. Convergent eusocial evolution is based on a shared reproductive groundplan plus lineage-specific plastic genes. Nature communications. 2019;10:1-11.

29. Nishida H. Detection and characterization of fungal-specific proteins in Saccharomyces cerevisiae. Bioscience, biotechnology, and biochemistry. 2006:0610050137-.

30. Ma D, Ding Q, Guo Z, Zhao Z, Wei L, Li Y, et al. Identification, characterization and expression analysis of lineage-specific genes within mangrove species Aegiceras corniculatum. Molecular Genetics and Genomics. 2021:1-13.

31. Šestak MS, Domazet-Lošo T. Phylostratigraphic profiles in zebrafish uncover chordate origins of the vertebrate brain. Molecular Biology and Evolution. 2015;32:299-312.

32. Domazet-Lošo T, Brajković J, Tautz D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends in Genetics. 2007;23:533-9.

33. Neme R, Tautz D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. BMC Genomics. 2013;14:117.

34. Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. Nature Reviews Genetics. 2011;12:692.

35. Schmitz JF, Ullrich KK, Bornberg-Bauer E. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. Nature ecology & evolution. 2018;2:1626-32.

36. Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E. Mechanisms and dynamics of orphan gene emergence in insect genomes. Genome Biology and Evolution. 2013;5:439-55.

37. Sun W, Zhao X-W, Zhang Z. Identification and evolution of the orphan genes in the domestic silkworm, Bombyx mori. FEBS letters. 2015;589:2731-8.

38. Palmieri N, Kosiol C, Schlötterer C. The life cycle of Drosophila orphan genes. eLife. 2014;3:e01311.

39. Arendsee ZW, Li L, Wurtele ES. Coming of age: orphan genes in plants. Trends in Plant Science. 2014;19:698-708.

40. Campbell MA, Zhu W, Jiang N, Lin H, Ouyang S, Childs KL, et al. Identification and characterization of lineage-specific genes within the Poaceae. Plant physiology. 2007;145:1311-22.

41. Yang L, Zou M, Fu B, He S. Genome-wide identification, characterization, and expression analysis of lineage-specific genes within zebrafish. Bmc Genomics. 2013;14:1-15.

42. Lin H, Moghe G, Ouyang S, Iezzoni A, Shiu S-H, Gu X, et al. Comparative analyses reveal distinct sets of lineage-specific genes within Arabidopsis thaliana. BMC evolutionary biology. 2010;10:1-14.

43. Yang X, Jawdy S, Tschaplinski TJ, Tuskan GA. Genome-wide identification of lineage-specific genes in Arabidopsis, Oryza and Populus. Genomics. 2009;93:473-80.

44. Mukherjee S, Panda A, Ghosh TC. Elucidating evolutionary features and functional implications of orphan genes in Leishmania major. Infection, Genetics and Evolution. 2015;32:330-7.

45. Li G, Wu X, Hu Y, Muñoz-Amatriaín M, Luo J, Zhou W, et al. Orphan genes are involved in drought adaptations and ecoclimatic-oriented selections in domesticated cowpea. Journal of experimental botany. 2019;70:3101-10.

46. McLysaght A, Hurst LD. Open questions in the study of de novo genes: what, how and why. Nature Reviews Genetics. 2016;17:567.

47. Van Oss SB, Carvunis A-R. De novo gene birth. PLoS genetics. 2019;15:e1008160.

48. Albà MM, Castresana J. On homology searches by protein BLAST and the characterization of the age of genes. BMC Evolutionary Biology. 2007;7:53.

49. Blevins WR, Ruiz-Orera J, Messeguer X, Blasco-Moreno B, Villanueva-Cañas JL, Espinar L, et al. Uncovering de novo gene birth in yeast using deep transcriptomics. Nature Communications. 2021;12:604. doi: 10.1038/s41467-021-20911-3.

50. Li Z-W, Chen X, Wu Q, Hagmann J, Han T-S, Zou Y-P, et al. On the origin of de novo genes in Arabidopsis thaliana populations. Genome biology and evolution. 2016;8:2190-202.

51. Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, et al. A molecular portrait of de novo genes in yeasts. Molecular Biology and Evolution. 2017;35:631-45.

52. Heames B, Schmitz J, Bornberg-Bauer E. A continuum of evolving de novo genes drives protein-coding novelty in Drosophila. Journal of molecular evolution. 2020;88:382-98.

53. Wilson BA, Foy SG, Neme R, Masel J. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. Nature ecology & evolution. 2017;1:1-6.

54. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics. 2012;13:329-42.

55. Salzberg SL. Next-generation genome annotation: we still struggle to get it right. BioMed Central; 2019.

56. Thibaud-Nissen F, DiCuccio M, Hlavina W, Kimchi A, Kitts P, Murphy T, et al. P8008 the NCBI eukaryotic genome annotation pipeline. Journal of Animal Science. 2016;94:184-.

57. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. Nucleic acids research. 2021;49:D884-D91.

58. Drysdale R, Consortium F. FlyBase. Drosophila. 2008:45-59.

59. Howe K, Davis P, Paulini M, Tuli MA, Williams G, Yook K, et al., editors. WormBase: annotating many nematode genomes. Worm; 2012: Taylor & Francis.

60. Giraldo-Calderón GI, Emrich SJ, MacCallum RM, Maslen G, Dialynas E, Topalis P, et al. VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. Nucleic acids research. 2015;43:D707-D13.

61. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome research. 2008;18:188-96.

62. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome biology. 2008;9:1-22.

63. Basile W, Elofsson A. The number of orphans in yeast and fly is drastically reduced by using combining searches in both proteomes and genomes. bioRxiv. 2017:185983.

64. Casola C. From de novo to “de nono”: the majority of novel protein-coding genes identified with phylostratigraphy are old genes or recent duplicates. Genome Biology and Evolution. 2018;10:2906-18.

65. Zile K, Dessimoz C, Wurm Y, Masel J. Only a single taxonomically restricted gene family in the Drosophila melanogaster subgroup can be identified with high confidence. Genome Biology and Evolution. 2020.

66. Lee CM, Barber GP, Casper J, Clawson H, Diekhans M, Gonzalez JN, et al. UCSC Genome Browser enters 20th year. Nucleic acids research. 2020;48:D756-D61.

67. Teeling EC, Vernes SC, Dávalos LM, Ray DA, Gilbert MTP, Myers E, et al. Bat biology, genomes, and the Bat1K project: to generate chromosome-level genomes for all living bat species. Annual review of animal biosciences. 2018;6:23-46.

68. Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, et al. The genomic substrate for adaptive radiation in African cichlid fish. Nature. 2014;513:375-81.

69. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990;215:403-10.

70. Prabh N, Rödelsperger C. De novo, divergence, and mixed origin contribute to the emergence of orphan genes in pristionchus nematodes. G3: Genes, Genomes, Genetics. 2019;9:2277-86.

71. Aguilera F, McDougall C, Degnan BM. Co-option and de novo gene evolution underlie molluscan shell diversity. Molecular Biology and Evolution. 2017;34:779-92.

72. Huang J, Chen J, Fang G, Pang L, Zhou S, Zhou Y, et al. Two novel venom proteins underlie divergent parasitic strategies between a generalist and a specialist parasite. Nature communications. 2021;12:1-16.

73. Papanicolaou A, Schetelig MF, Arensburger P, Atkinson PW, Benoit JB, Bourtzis K, et al. The whole genome sequence of the Mediterranean fruit fly, Ceratitis capitata (Wiedemann), reveals insights into the biology and adaptive evolution of a highly invasive pest species. Genome biology. 2016;17:1-31.

74. Forêt S, Knack B, Houliston E, Momose T, Manuel M, Quéinnec E, et al. New tricks with old genes: the genetic bases of novel cnidarian traits. Trends in Genetics. 2010;26:154-8.

75. Donoghue MT, Keshavaiah C, Swamidatta SH, Spillane C. Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. BMC evolutionary biology. 2011;11:1-23.

76. Wilkinson RD, Steiper ME, Soligo C, Martin RD, Yang Z, Tavaré S. Dating primate divergences through an integrated analysis of palaeontological and molecular data. Systematic biology. 2011;60:16-31.

77. Maestri R, Monteiro LR, Fornel R, Upham NS, Patterson BD, de Freitas TRO. The ecology of a continental evolutionary radiation: Is the radiation of sigmodontine rodents adaptive? Evolution. 2017;71:610-32.

78. Shi JJ, Rabosky DL. Speciation dynamics during the global radiation of extant bats. Evolution. 2015;69:1528-45.

79. Sharp PM, Li W-H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic acids research. 1987;15:1281-95.

80. Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic acids research. 2013;41:e74-e.

81. Dosztányi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics. 2005;21:3433-4.

82. Weisman CM, Murray AW, Eddy SR. Many, but not all, lineage-specific genes can be explained by homology detection failure. PLoS biology. 2020;18:e3000862.

83. Moyers BA, Zhang J. Evaluating phylostratigraphic evidence for widespread de novo gene birth in genome evolution. Molecular biology and evolution. 2016;33:1245-56.

84. Richter DJ, Fozouni P, Eisen MB, King N. Gene family innovation, conservation and loss on the animal stem lineage. eLife. 2018;7:e34226.

85. Mitreva M, Blaxter ML, Bird DM, McCarter JP. Comparative genomics of nematodes. TRENDS in Genetics. 2005;21:573-81.

86. Riley R, Haridas S, Wolfe KH, Lopes MR, Hittinger CT, Göker M, et al. Comparative genomics of biotechnologically important yeasts. Proceedings of the National Academy of Sciences. 2016;113:9882-7.

87. Palmer WJ, Jiggins FM. Comparative genomics reveals the origins and diversity of arthropod immune systems. Molecular biology and evolution. 2015;32:2111-29.

88. Bhattacharya D, Agrawal S, Aranda M, Baumgarten S, Belcaid M, Drake JL, et al. Comparative genomics explains the evolutionary success of reef-forming corals. elife. 2016;5:e13288.

89. Nielly-Thibault L, Landry CR. Differences between the raw material and the products of de novo gene birth can result from mutational biases. Genetics. 2019;212:1353-66.

90. Durand É, Gagnon-Arsenault I, Hallin J, Hatin I, Dubé AK, Nielly-Thibault L, et al. Turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. Genome research. 2019;29:932-43.

91. Ruiz-Orera J, Verdaguer-Grau P, Villanueva-Cañas JL, Messeguer X, Albà MM. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. Nature ecology & evolution. 2018;2:890-6.

92. Wilson BA, Masel J. Putatively noncoding transcripts show extensive association with ribosomes. Genome biology and evolution. 2011;3:1245-52.

93. Söllner JF, Leparc G, Hildebrandt T, Klein H, Thomas L, Stupka E, et al. An RNA-Seq atlas of gene expression in mouse and rat normal tissues. Scientific data. 2017;4:1-11.

94. Langmead B. Aligning short sequencing reads with Bowtie. Current protocols in bioinformatics. 2010;32:11.7. 1-.7. 4.

95. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923-30.

96. Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic acids research. 2013;41:e108-e.

97. Vakirlis N, Carvunis AR, McLysaght A. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. eLife. 2020;9. Epub 2020/02/19. doi: 10.7554/eLife.53500. PubMed PMID: 32066524; PubMed Central PMCID: PMCPMC7028367.

98. Elhaik E, Sabath N, Graur D. The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. Molecular Biology and Evolution. 2005;23:1-3.

99. Moyers BA, Zhang J. Phylostratigraphic bias creates spurious patterns of genome evolution. Molecular Biology and Evolution. 2014;32:258-67.

100. Cai JJ, Woo PC, Lau SK, Smith DK, Yuen K-Y. Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in ascomycota. Journal of Molecular Evolution. 2006;63:1-11.

101. Kuo C-H, Kissinger JC. Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites Plasmodium and Theileria. BMC Evolutionary Biology. 2008;8:1-16.

102. Xu Y, Wu G, Hao B, Chen L, Deng X, Xu Q. Identification, characterization and expression analysis of lineage-specific genes within sweet orange (Citrus sinensis). BMC genomics. 2015;16:1-10.

103. Moyers BA, Zhang J. Toward reducing phylostratigraphic errors and biases. Genome Biology and Evolution. 2018;10:2037-48.

104. Moyers BA, Zhang J. Further simulations and analyses demonstrate open problems of phylostratigraphy. Genome Biology and Evolution. 2017;9:1519-27.

105. Moyers BA, Zhang J. Evaluating phylostratigraphic evidence for widespread de novo gene birth in genome evolution. Molecular Biology and Evolution. 2016;33:1245-56. Epub 2016/01/14. doi: 10.1093/molbev/msw008. PubMed PMID: 26758516; PubMed Central PMCID: PMCPMC5010002.

106. Jain A, Perisa D, Fliedner F, von Haeseler A, Ebersberger I. The evolutionary traceability of a protein. Genome Biology and Evolution. 2019;11:531-45.

107. Domazet-Lošo T, Carvunis A-R, Albà M, Šestak MS, Bakarić R, Neme R, et al. No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. Molecular Biology and Evolution. 2017;34:843-56.

108. Ekman D, Elofsson A. Identifying and quantifying orphan protein sequences in fungi. Journal of Molecular Biology. 2010;396:396-405.

109. Beimforde C, Feldberg K, Nylinder S, Rikkinen J, Tuovila H, Dörfelt H, et al. Estimating the Phanerozoic history of the Ascomycota lineages: combining fossil and molecular data. Molecular Phylogenetics and Evolution. 2014;78:386-98.

110. Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, et al. Phylogenomics resolves the timing and pattern of insect evolution. Science. 2014;346:763-7.

111. Fitzpatrick DA, Logue ME, Stajich JE, Butler G. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evolutionary Biology. 2006;6:99.

112. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. Molecular Biology and Evolution. 2017;35:543-8.

113. Zhang J. Protein-length distributions for the three domains of life. Trends in Genetics. 2000;16:107-9.

114. Koonin EV. Are there laws of genome evolution? PLOS Computational Biology. 2011;7:e1002173.

115. Cai J, Zhao R, Jiang H, Wang W. De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. Genetics. 2008;179:487-96.

116. Metzger MB, Michaelis S. Analysis of quality control substrates in distinct cellular compartments reveals a unique role for Rpn4p in tolerating misfolded membrane proteins. Molecular Biology of the Cell. 2009;20:1006-19.

117. Ng PC, Wong ED, MacPherson KA, Aleksander S, Argasinska J, Dunn B, et al. Transcriptome visualization and data availability at the Saccharomyces Genome Database. Nucleic acids research. 2020;48:D743-D8.

118. Scannell DR, Zill OA, Rokas A, Payen C, Dunham MJ, Eisen MB, et al. The awesome power of yeast evolutionary genetics: new genome sequences and strain resources for the Saccharomyces sensu stricto genus. G3: Genes, Genomes, Genetics. 2011;1:11-25.

119. Byrne KP, Wolfe KH. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Research. 2005;15:1456-61.

120. Shonn MA, McCarroll R, Murray AW. Spo13 protects meiotic cohesin at centromeres in meiosis I. Genes & Development. 2002;16:1659-71.

121. Luis Villanueva-Cañas J, Ruiz-Orera J, Agea MI, Gallo M, Andreu D, Albà MM. New genes and functional innovation in mammals. Genome Biology and Evolution. 2017;9:1886-900.

122. Nuckolls NL, Núñez MAB, Eickbush MT, Young JM, Lange JJ, Jonathan SY, et al. wtf genes are prolific dual poison-antidote meiotic drivers. eLife. 2017;6:e26033.

123. Malik HS, Henikoff S. Adaptive evolution of Cid, a centromere-specific histone in Drosophila. Genetics. 2001;157:1293-8.

124. Guerzoni D, McLysaght A. De novo genes arise at a slow but steady rate along the primate lineage and have been subject to incomplete lineage sorting. Genome Biology and Evolution. 2016;8:1222-32.

125. Zhuang X, Yang C, Murphy KR, Cheng CC. Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. Proceedings of the National Academy of Sciences. 2019;116:4400-5.

126. Surm JM, Stewart ZK, Papanicolaou A, Pavasovic A, Prentis PJ. The draft genome of Actinia tenebrosa reveals insights into toxin evolution. Ecology and Evolution. 2019.

127. Milde S, Hemmrich G, Anton-Erxleben F, Khalturin K, Wittlieb J, Bosch TC. Characterization of taxonomically restricted genes in a phylum-restricted cell type. Genome Biology. 2009;10:R8.

128. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997;25:3389-402.

129. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004;32:1792-7.

130. Felsenstein J. PHYLIP (phylogeny inference package), version 3.5 c: Joseph Felsenstein.; 1993.

131. Ontology CG. The gene ontology resource: 20 years and still GOing strong. Nucleic Acids Research. 2018;47:D330-D8.

132. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology. 2011;7.

133. Gilbert SF. Developmental Biology: Sinauer Associates; 2010.

134. Burton PM. Insights from diploblasts; the evolution of mesoderm and muscle. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution. 2008;310:5-14.

135. Ereskovsky AV, Dondua AK. The problem of germ layers in sponges (Porifera) and some issues concerning early metazoan evolution. Zoologischer Anzeiger-A Journal of Comparative Zoology. 2006;245:65-76.

136. Ryan JF, Pang K, Schnitzler CE, Nguyen A-D, Moreland RT, Simmons DK, et al. The genome of the ctenophore Mnemiopsis leidyi and its implications for cell type evolution. Science. 2013;342:1242592.

137. Hernandez-Nicaise M-L, Nicaise G, Malaval L. Giant smooth muscle fibers of the ctenophore Mnemiopsis leydii: ultrastructural study of in situ and isolated cells. The Biological Bulletin. 1984;167:210-28.

138. Mackie G, Mills C, Singla C. Structure and function of the prehensile tentilla of Euplokamis (Ctenophora, Cydippida). Zoomorphology. 1988;107:319-37.

139. Jager M, Manuel M. Ctenophores: an evolutionary-developmental perspective. Current opinion in genetics & development. 2016;39:85-92.

140. Martindale MQ, Henry JQ. The development of radial and biradial symmetry: The evolution of bilaterality. American Zoologist. 1998;38:672-84.

141. Hymen L. The Invertebrates: Protozoa through Ctenophora. McGraw Hill, New York, xii; 1940.

142. Martindale MQ, Henry JQ. Intracellular fate mapping in a basal metazoan, the ctenophore Mnemiopsis leidyi, reveals the origins of mesoderm and the existence of indeterminate cell lineages. Developmental biology. 1999;214:243-57.

143. Derelle R, Manuel M. Ancient connection between NKL genes and the mesoderm? Insights from Tlx expression in a ctenophore. Development genes and evolution. 2007;217:253-61.

144. Dunn CW, Leys SP, Haddock SH. The hidden biology of sponges and ctenophores. Trends in ecology & evolution. 2015;30:282-91.

145. Giribet G. Genomics and the animal tree of life: conflicts and future prospects. Zoologica Scripta. 2016;45:14-21.

146. Li Y, Shen X-X, Evans B, Dunn CW, Rokas A. Rooting the animal tree of life. bioRxiv. 2020.

147. Moroz LL, Kocot KM, Citarella MR, Dosung S, Norekian TP, Povolotskaya IS, et al. The ctenophore genome and the evolutionary origins of neural systems. Nature. 2014;510:109.

148. Parfrey LW, Lahr DJ, Knoll AH, Katz LA. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proceedings of the National Academy of Sciences. 2011;108:13624-9.

149. Eddy SR. Accelerated profile HMM searches. PLoS computational biology. 2011;7:e1002195.

150. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, et al. The Pfam protein families database. Nucleic acids research. 2004;32:D138-D41.

151. Sebé-Pedrós A, Chomsky E, Pang K, Lara-Astiaso D, Gaiti F, Mukamel Z, et al. Early metazoan cell type diversity and the evolution of multicellular gene regulation. Nature ecology & evolution. 2018;2:1176-88.

152. Levin M, Anavy L, Cole AG, Winter E, Mostov N, Khair S, et al. The mid-developmental transition and the evolution of animal body plans. Nature. 2016;531:637-41.

153. Yamada A, Martindale MQ, Fukui A, Tochinai S. Highly conserved functions of the Brachyury gene on morphogenetic movements: insight from the early-diverging phylum Ctenophora. Developmental biology. 2010;339:212-22.

154. Presnell JS, Browne WE. Krüppel-like factor gene function in the ctenophore Mnemiopsis leidyi assessed by CRISPR/Cas9-mediated genome editing. bioRxiv. 2020:527002.

155. Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJ, Jackson SE, et al. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell reports. 2014;8:1365-79.

156. Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, et al. The reality of pervasive transcription. PLoS Biol. 2011;9:e1000625.

157. Neme R, Tautz D. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. Elife. 2016;5:e09977.

158. McLysaght A, Guerzoni D. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. Philosophical Transactions of the Royal Society B: Biological Sciences. 2015;370:20140332.

159. Polev DE, Karnaukhova IK, Krukovskaya LL, Kozlov AP. ELFN1-AS1: a novel primate gene with possible microRNA function expressed predominantly in human tumors. BioMed research international. 2014;2014.

160. Dong L, Ding C, Zheng T, Pu Y, Liu J, Zhang W, et al. Extracellular vesicles from human umbilical cord mesenchymal stem cells treated with siRNA against ELFN1-AS1 suppress colon adenocarcinoma proliferation and migration. American journal of translational research. 2019;11:6989.

161. Lei R, Feng L, Hong D. ELFN1-AS1 accelerates the proliferation and migration of colorectal cancer via regulation of miR-4644/TRIM44 axis. Cancer Biomarkers. 2020:1-11.

162. Wei C-Y, Wang L, Zhu M-X, Deng X-Y, Wang D-H, Zhang S-M, et al. TRIM44 activates the AKT/mTOR signal pathway to induce melanoma progression by stabilizing TLR4. Journal of Experimental & Clinical Cancer Research. 2019;38:137.

163. Zhang C, Lian H, Xie L, Yin N, Cui Y. LncRNA ELFN1-AS1 promotes esophageal cancer progression by up-regulating GFPT1 via sponging miR-183-3p. Biological chemistry. 2020;1.

164. Jie Y, Ye L, Chen H, Yu X, Cai L, He W, et al. ELFN1-AS1 accelerates cell proliferation, invasion and migration via regulating miR-497-3p/CLDN4 axis in ovarian cancer. Bioengineered. 2020;11:872-82.

165. Wang F, Gao Y, Tang L, Ning K, Geng N, Zhang H, et al. A novel PAK4-CEBPB-CLDN4 axis involving in breast cancer cell migration and invasion. Biochemical and biophysical research communications. 2019;511:404-8.

166. Heinen TJ, Staubach F, Häming D, Tautz D. Emergence of a new gene from an intergenic region. Current Biology. 2009;19:1527-31.

167. Li T, Stark MR, Johnson AD, Wolberger C. Crystal structure of the MATa1/MATα2 homeodomain heterodimer bound to DNA. Science. 1995;270:262-9.

168. Li D, Yan Z, Lu L, Jiang H, Wang W. Pleiotropy of the de novo-originated gene MDF1. Scientific reports. 2014;4:1-4.

169. Suenaga Y, Nakatani K, Nakagawara A. De novo evolved gene product NCYM in the pathogenesis and clinical outcome of human neuroblastomas and other cancers. Japanese Journal of Clinical Oncology. 2020;50:839-46.

170. Kang J-H, Rychahou PG, Ishola TA, Qiao J, Evers BM, Chung DH. MYCN silencing induces differentiation and apoptosis in human neuroblastoma cells. Biochemical and biophysical research communications. 2006;351:192-7.

171. Suenaga Y, Islam SR, Alagu J, Kaneko Y, Kato M, Tanaka Y, et al. NCYM, a Cis-antisense gene of MYCN, encodes a de novo evolved protein that inhibits GSK3β resulting in the stabilization of MYCN in human neuroblastomas. PLoS Genet. 2014;10:e1003996.

172. Shoji W, Suenaga Y, Kaneko Y, Islam SR, Alagu J, Yokoi S, et al. NCYM promotes calpain-mediated Myc-nick production in human MYCN-amplified neuroblastoma cells. Biochemical and biophysical research communications. 2015;461:501-6.

173. Suenaga Y, Kaneko Y, Matsumoto D, Hossain MS, Ozaki T, Nakagawara A. Positive auto-regulation of MYCN in human neuroblastoma. Biochemical and biophysical research communications. 2009;390:21-6.

174. Kaneko Y, Suenaga Y, Islam SR, Matsumoto D, Nakamura Y, Ohira M, et al. Functional interplay between MYCN, NCYM, and OCT 4 promotes aggressiveness of human neuroblastomas. Cancer Science. 2015;106:840-7.

175. Zhao X, Li D, Pu J, Mei H, Yang D, Xiang X, et al. CTCF cooperates with noncoding RNA MYCNOS to promote neuroblastoma progression through facilitating MYCN expression. Oncogene. 2016;35:3565-76.

176. Besançon R, Valsesia-Wittmann S, Locher C, Delloye-Bourgeois C, Furhman L, Tutrone G, et al. Upstream ORF affects MYCN translation depending on exon 1b alternative splicing. BMC cancer. 2009;9:445.

177. Vadie N, Saayman S, Lenox A, Ackley A, Clemson M, Burdach J, et al. MYCNOS functions as an antisense RNA regulating MYCN. RNA biology. 2015;12:893-9.

178. Krystal GW, Armstrong B, Battey J. N-myc mRNA forms an RNA-RNA duplex with endogenous antisense transcripts. Molecular and cellular biology. 1990;10:4180-91.

179. An G, Ng AY, Meka CR, Luo G, Bright SP, Cazares L, et al. Cloning and characterization of UROC28, a novel gene overexpressed in prostate, breast, and bladder cancers. Cancer research. 2000;60:7014-20.

180. Pan T, Wu R, Liu B, Wen H, Tu Z, Guo J, et al. PBOV1 promotes prostate cancer proliferation by promoting G1/S transition. OncoTargets and therapy. 2016;9:787.

181. Guo Y, Wu Z, Shen S, Guo R, Wang J, Wang W, et al. Nanomedicines reveal how PBOV1 promotes hepatocellular carcinoma for effective gene therapy. Nature communications. 2018;9:1-16.

182. Wang L, Niu C-H, Wu S, Wu H-M, Ouyang F, He M, et al. PBOV1 correlates with progression of ovarian cancer and inhibits proliferation of ovarian cancer cells. Oncology Reports. 2016;35:488-96.

183. Yang C-A, Li J-P, Yen J-C, Lai I-L, Ho Y-C, Chen Y-C, et al. lncRNA NTT/PBOV1 axis promotes monocyte differentiation and is elevated in rheumatoid arthritis. International journal of molecular sciences. 2018;19:2806.

184. Chen J-Y, Shen QS, Zhou W-Z, Peng J, He BZ, Li Y, et al. Emergence, retention and selection: a trilogy of origination for functional de novo proteins from ancestral LncRNAs in primates. PLoS genetics. 2015;11:e1005391.

185. Papamichos SI, Margaritis D, Kotsianidis I. Adaptive Evolution Coupled with Retrotransposon Exaptation Allowed for the Generation of a Human-Protein-Specific Coding Gene That Promotes Cancer Cell Proliferation and Metastasis in Both Haematological Malignancies and Solid Tumours: The Extraordinary Case of MYEOV Gene. Scientifica. 2015;2015.

186. Janssen JW, Vaandrager J-W, Heuser T, Jauch A, Kluin PM, Geelen E, et al. Concurrent activation of a novel putative transforming gene, myeov, and cyclin D1 in a subset of multiple myeloma cell lines with t (11; 14)(q13; q32). Blood, The Journal of the American Society of Hematology. 2000;95:2691-8.

187. Leyden J, Murray D, Moss A, Arumuguma M, Doyle E, McEntee G, et al. Net1 and Myeov: computationally identified mediators of gastric cancer. British journal of cancer. 2006;94:1204-12.

188. Moss AC, Lawlor G, Murray D, Tighe D, Madden SF, Mulligan A-M, et al. ETV4 and Myeov knockdown impairs colon cancer cell line proliferation and invasion. Biochemical and biophysical research communications. 2006;345:216-21.

189. Takita J, Chen Y, Okubo J, Sanada M, Adachi M, Ohki K, et al. Aberrations of NEGR1 on 1p31 and MYEOV on 11q13 in neuroblastoma. Cancer science. 2011;102:1645-50.

190. de Almeida RA, Heuser T, Blaschke R, Bartram CR, Janssen JW. Control of MYEOV protein synthesis by upstream open reading frames. Journal of Biological Chemistry. 2006;281:695-704.

191. Müller P, Crofts JD, Newman BS, Bridgewater LC, Lin C-Y, Gustafsson J-Å, et al. SOX9 mediates the retinoic acid-induced HES-1 gene expression in human breast cancer cells. Breast cancer research and treatment. 2010;120:317-26.

192. Bridgewater LC, Walker MD, Miller GC, Ellison TA, Holsinger LD, Potter JL, et al. Adjacent DNA sequences modulate Sox9 transcriptional activation at paired Sox sites in three chondrocyte-specific enhancer elements. Nucleic acids research. 2003;31:1541-53.

193. Liang E, Lu Y, Shi Y, Zhou Q, Zhi F. MYEOV increases HES1 expression and promotes pancreatic cancer progression by enhancing SOX9 transactivity. Oncogene. 2020;39:6437-50.

194. Fang L, Wu S, Zhu X, Cai J, Wu J, He Z, et al. MYEOV functions as an amplified competing endogenous RNA in promoting metastasis by activating TGF-β pathway in NSCLC. Oncogene. 2019;38:896-912.

195. Cheng C-HC. Evolution of the diverse antifreeze proteins. Current opinion in genetics & development. 1998;8:715-20.

196. Gill DJ, Clausen H, Bard F. Location, location, location: new insights into O-GalNAc protein glycosylation. Trends in cell biology. 2011;21:149-58.

197. Giubertoni G, Meister K, DeVries AL, Bakker HJ. Determination of the solution structure of antifreeze glycoproteins using two-dimensional infrared spectroscopy. The journal of physical chemistry letters. 2019;10:352-7.

198. Devries AL. Glycoproteins as biological antifreeze agents in Antarctic fishes. Science. 1971;172:1152-5.

199. Chen L, DeVries AL, Cheng C-HC. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. Proceedings of the National Academy of Sciences. 1997;94:3811-6.

200. Bartel DP. MicroRNAs: target recognition and regulatory functions. cell. 2009;136:215-33.

201. de Boer CG, Vaishnav ED, Sadeh R, Abeyta EL, Friedman N, Regev A. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. Nature biotechnology. 2020;38:56-65.

202. Yona AH, Alm EJ, Gore J. Random sequences rapidly evolve into de novo promoters. Nature communications. 2018;9:1-10.

203. Murphy DN, McLysaght A. De novo origin of protein-coding genes in murine rodents. PLoS One. 2012;7:e48650.

204. Knowles DG, McLysaght A. Recent de novo origin of human protein-coding genes. Genome research. 2009;19:1752-9.

205. Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD. De novo ORFs in Drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences. PLoS Genet. 2013;9:e1003860.

Supplemental Materials

	Oreochromis	*Neolamprologus*	*Astatotilapia*	*Pundamilia*	*Metriaclimia*	**Cichlids**		**Species**
	Broad Institute	Broad Institute	Broad Institute	Broad Institute	Broad Institute			**Annotation 1**
	Custom pipeline	Custom pipeline	Custom pipeline	Custom pipeline	Custom pipeline			**Brief description**
	https://static-	https://static-	https://static-	https://static-	https://static-			**Link to method**
	ftp://ftp.broadinst	ftp://ftp.broadinst	ftp://ftp.broadinst	ftp://ftp.broadinst	ftp://ftp.broadinst			**Download link**
	NCBI	NCBI	NCBI	NCBI	NCBI			**Annotation 2**
	NCBI custom	NCBI custom	NCBI custom	NCBI custom	NCBI custom			**Brief description**
	https://www.ncbi.	https://www.ncbi.	https://www.ncbi.	https://www.ncbi.	https://www.ncbi.			Link to method
	ftp://ftp.ncbi.nlm.	https://ftp.ncbi.nl	https://ftp.ncbi.nl	https://ftp.ncbi.nl	https://ftp.ncbi.nl			**Download link**

Mus caroli	*Mus musculus*	**Rodents**			*Cebus imitator*	*Rhinopithecus*	*Mandrillus*	*Macaca nemestri*	*Macaca fascicula*	**Primates**
UCSC	Ensembl				Ensembl	Ensembl	Ensembl	Ensembl	Ensembl	
UCSC custom	Ensembl custom				Ensembl custom	Ensembl custom	Ensembl custom	Ensembl custom	Ensembl custom	
https://github.com	https://useast.ense				http://useast.ense	http://uswest.ens	http://uswest.ens	http://uswest.ens	http://nov2020.ar	
ftp://ftp.ensembl.o	ftp://ftp.ensembl.o				http://ftp.ensembl.	http://ftp.ensembl.	http://ftp.ensembl.	http://ftp.ensembl.	http://ftp.ensembl.	
NCBI	NCBI				NCBI	NCBI	NCBI	NCBI	NCBI	
NCBI custom	NCBI custom				NCBI custom	NCBI custom	NCBI custom	NCBI custom	NCBI custom	
https://www.ncbi.	https://www.ncbi.				https://www.ncbi.	https://www.ncbi.	https://www.ncbi.	https://www.ncbi.	https://www.ncbi.	
https://ftp.ncbi.nl	https://ftp.ncbi.nl				https://ftp.ncbi.nl	https://ftp.ncbi.nl	https://ftp.ncbi.nl	https://ftp.ncbi.nl	https://ftp.ncbi.nl	

Pteropus alecto	*Molossus molossus*	*Myotis myotis*	*Myotis brandtii*	*Myotis lucifugus*	**Bats**			*Peromyscus*	*Rattus norwegic*	*Mus pahari*
Beijing Genomic	Bat1K, using	Bat1K, using	Beijing Genomic	Ensembl				Ensembl	Ensembl	UCSC
Custom pipeline	Custom pipeline	Custom pipeline	Custom pipeline	Ensembl custom				Ensembl custom	Ensembl custom	UCSC custom
https://science-	https://www.natur	https://www.natur	https://science-	https://uswest.ens				https://jan2019.ar	https://useast.ense	https://github.com
https://ftp.ncbi.nl	https://ftp.ncbi.nl	https://ftp.ncbi.nl	https://ftp.ncbi.nl	http://ftp.ensembl.				ftp://ftp.ensembl.o	ftp://ftp.ensembl.o	ftp://ftp.ensembl.o
NCBI	NCBI	NCBI	NCBI	NCBI				NCBI	NCBI	NCBI
NCBI custom	NCBI custom	NCBI custom	NCBI custom	NCBI custom				NCBI custom	NCBI custom	NCBI custom
https://www.ncbi.	https://www.ncbi.	https://www.ncbi.	https://www.ncbi.	https://www.ncbi.				https://www.ncbi.	https://www.ncbi.	https://www.ncbi.
https://ftp.ncbi.nl	https://ftp.ncbi.nl	https://ftp.ncbi.nl	https://ftp.ncbi.nl	https://ftp.ncbi.nl				https://ftp.ncbi.nl	https://ftp.ncbi.nl	https://ftp.ncbi.nl

Supplemental Table 1: Details of annotation pairs considered for all analyses in Chapter 1.

Fungi	Refseq/Genbank Accession ID (if applicable)	Download link (if source not Refseq/Genbank)
S. cerevisiae	GCF_000146045.2	
S. paradoxus	n/a	http://www.saccharomycessensustricto.org/current/Spar/Spar.aa
S. mikatae	n/a	http://www.saccharomycessensustricto.org/current/Smik/Smik.aa
S. kudriavzevii	n/a	http://www.saccharomycessensustricto.org/current/Skud/Skud.aa
S. bayanus	n/a	http://www.saccharomycessensustricto.org/current/Sbay/Sbay.aa
S. castellii	GCF_000237345.1	
K. waltii	n/a	https://media.nature.com/full/nature-assets/nature/journal/v428/n6983/extref/S2_ORFs/predicted_proteins.fasta
A. gossyppii	GCF_000091025.4	
K. lactis	GCF_000002515.2	
A. nidulans	GCA_000149205.2	

S. pombe	GCA_000002945.2	
Y. lipolytica	GCA_000002525.1	
Insects		
D. melanogaster	GCF_000001215.4	
D. simulans	GCF_000754195.2	
D. sechelia	GCF_000005215.3	
D. erecta	GCF_000005135.1	
D. yakuba	GCF_000005975.2	
D. pseudoobscura	GCF_000001765.3	
D. persimilis	GCF_000005195.1	
D. willistoni	GCF_000005925.1	
D. virilis	GCF_000005245.1	
D. grimshawi	GCF_000005155.2	

D. mojavensis	GCF_000005175.2	
M. domestica	GCF_000371365.1	
C. capitata	GCF_000347755.3	
A. aegypti	GCF_002204515.2	
A. gambiae	GCF_000005575.2	
Z. nevadensis	GCF_000696155.1	
T. castaneum	GCF_000002335.3	
A. mellifera	GCF_000002195.4	
B. germanica	GCA_003018175.1	
B. terrestris	GCF_000214255.1	
B. mori	GCF_000151625.1	
A. pisum	GCF_000142985.2	

Supplemental Table 2: Sources for protein annotations used in all analyses in Chapter 2.