PhD Confirmation Report

Zoe Vance

March 2019

Summary

Gene duplication has long been known to be an important process in the evolution of genome structure and creation of new genetic material. Duplication may occur by one of two major mechanisms; tandem duplication or whole genome duplication. The sets of duplicates retained following these events differ greatly,forming largely non-overlapping sets. Many differences have been noted between these gene sets in function, basic gene features such as genomic length and more complex properties such as how gene expression is regulated. However, the current body of work on this topic lacks a cohesive narrative regarding what features distinguish a gene which is retained following whole genome duplication (ohnolog) from a gene which may duplicate by tandem duplication. Current studies in the literature are inconsistent in study species and definitions of the gene sets to be compared. Here we show, by comparing a large number of features in consistently defined sets of duplicable genes in human, that there is a general trend of higher constraint in ohnologs, and of lower constraint in tandem duplicated genes, relative to singletons. Ohnologs are generally more complex, slower evolving and more heavily regulated with tandem duplicates showing the opposite pattern. Additionally, we confirm previously reported trends for enriched functions in both gene sets; ohnologs are enriched for developmental, nervous system and regulatory functions while tandem duplicates show enrichment for functions associated with environmental triggers such as immune and sensory functions. As these comparisons are performed within a single species, using a consistent definition for the categories compared, they allowed a comparison of the relative importances of each feature in determination of the mode of duplication; this analysis indicates that features related to conservation play a major role in retention of ohnologs while features related to gene length and complexity are more important in successful tandem duplication. The results shown here serve as a useful starting point to answer more complex questions about the nature of gene duplication.

1 Introduction

Gene duplication has long been known to be an important process in the evolution of genome structure and creation of new genetic material. This idea was first proposed by Susumu Ohno, who put forward the idea that duplication creates new genetic material for evolutionary forces to act on and therefore allows for biological innovation and adaptation (Ohno, 1970). Genes may duplicate by several different mechanisms including tandem duplication, whole genome duplication and retrotransposition. The two major mechanisms of these, tandem duplication and whole genome duplication, will be discussed here.

The first, tandem duplication or small scale duplication (SSD), concerns duplication of relatively short stretches of DNA (Bailey et al., 2001, up to the level of 100s of kb;[]) which contain only a small number of genes. This process primarily occurs through non-allelic homologous recombination (NAHR). This process is generally associated with recombination hotspots and repetitive sequences, as it results from misalignment at replication, leading to differences in the likelihood that a gene will be duplicated depending on its sequence context. Genes which duplicate in this manner are generally thought to have low levels of constraint, which allows duplication to occur but also allows new duplicates to be easily lost again, particularly as at the point of creation a new duplicate is completely redundant to its parent gene. There is support for this idea from the observation that most duplicates are lost and that their half-life is generally short (Lynch and Conery, 2003). It is also shown that gene loss is the most likely fate of new duplicates and that it occurs quickly after duplication, which may be a result of lack of pressure to retain the duplicate or, in some cases, pressure for the loss of the duplicate.

Gene loss (non-functionalization) is the most likely fate for new duplicate genes, however other pathways have been proposed that allow for duplicate retention through the gain of necessary functions, consistent with Ohno's theory of evolution through gene duplication. Subfunctionalization is a process where the original function of the parent gene is compartmentalised between duplicates, creating a selective pressure for both copies to be retained. This may occur through passive processes whereby mutation degrades complementary function in each duplicate so that both are required to achieve all previously existing functions (the duplication-degeneration-complementation model) (Force et al., 1999), or through a more active process where the duplication allows adaptation in cases where a gene may have previously had functions which existed in conflict with each other (the escape from adaptive conflict model) (Hittinger and Carroll, 2007; Des Marais and Rausher, 2008). The partitioning of ancestral function may occur through paralogs adopting different functions of the parent gene, as mainly addressed in the above references, but also through partitioning of the parent gene's expression domains (Klüver et al., 2005) or dosage sharing such that both paralogs are necessary to reach the ancestral expression level (Lan and Pritchard, 2016).

A second mechanism which creates pressure for duplicate retention is neofunctionalisation. This occurs when a duplicate gains an entirely new function relative to the parent gene. The most widely known example of this process is the origin of anti-freeze proteins in fish, which have arisen multiple times from neofunctionalisation following duplication of genes with distinct functions such as trypsinogen and lectins (Liu et al., 2007; Chen et al., 1997). This process is not necessarily exclusive from subfunctionalization as a fate for duplicate genes. He and Zhang (2005b) indicate that subfunctionalisation provides an initial pressure for duplicate retention, allowing time for greater divergence and neofunctionalization to occur. Thus there is likely interplay between the processes related to duplication retention and evolution

The other major mechanism by which genes duplicate is whole genome duplication (WGD). As the name suggests, this is a much larger scale event than tandem duplication concerning the entire genome. These events are relatively common in plants, with all angiosperms sharing two ancient duplication events and many more lineages undergoing further duplication (Conant and Wolfe, 2007). WGD events have also occurred in yeast (Wolfe and Shields, 1997) and Paramecium (Aury et al., 2006). These events are comparatively rare in the vertebrate lineage, although it is generally agreed that two rounds of duplication occurred in the ancestor of all vertebrates (the 2R hypothesis) (Dehal and Boore, 2005; Nakatani et al., 2007; Putnam et al., 2008). Additional duplications are also found specific to teleost fish (Jaillon et al., 2004) and salmon (Macqueen and Johnston, 2014).

The tetraploid genome created by WGD is reliably reduced back to a diploid state by a period of genomic rearrangement referred to as re-diploidization (Gerstein et al., 2006). This process is complete when WGD paralogs have diverged sufficiently that they do not pair tetravalently at meiosis. This process is characterised by large genomic rearrangements and gene loss is common (Lien et al., 2016), however a subset of genes, termed ohnologs, are retained following this process. These genes are generally under relative dosage constraint and must be retained in order to maintain dosage balance (Makino and McLysaght, 2010). Balance is preserved following WGD but would be disrupted if individual genes were lost subsequently. Dosage constraint (reviewed in Rice and McLysaght, 2017) may occur in genes which are members of multi-protein complexes, where different members must maintain relative dosage, in cases where certain dosage thresholds must be met as in the case of haploinsufficient genes or many developmental morphogens. Reflecting this constraint, human ohnologs are enriched for functions associated with these groups of genes (Brunet et al., 2006; Blomme et al., 2006). In contrast to tandem duplicates, ohnologs are generally highly constrained in dosage and many do not duplicate outside of the WGD context, but when they are duplicated are retained over long periods of evolutionary time. This difference is reflected in the observation that ohnologs and SSDs form largely non-overlapping gene sets (Makino and McLysaght, 2010)

Ohnologs may also be subject to the processes which lead to retention of tandem duplicates, however their initial retention is not reliant on these processes occurring as there is no redundancy in duplicates in cases where dosage must be maintained - both must be retained despite performing the same function. Additionally, while tandem duplicates may face pressure to be lost as some duplications are actively detrimental, WGD is comparatively benign (Birchler and Veitia, 2012). As described in the context of sub-/neo-functionalization, this dosage retention pressure does not necessarily have to exist exclusively of the other processes, but could delay duplicate loss long enough to allow new functions to develop. In agreement with Ohno's idea of duplication leading to innovation, WGD events often precede large species radiations and novel functional innovation, as is the case with vertebrates but also in angiosperms (Jiao et al., 2011), yeasts (Scannell et al., 2006) and teleosts (Jaillon et al., 2004). Potentially this large scale creation of new genetic material allows for much broader and far-reaching innovations than the smaller scale tandem duplications, which is also supported by reports of higher adaptability in polyploid lineages (van Hoek and Hogeweg, 2009).

There are a number of biases which affect duplicate occurrence and the duplicate's subsequent fate. These can be classified as mutational biases, fixation biases and retention biases. Mutation bias refers to the factors which affect the initial occurrence of a duplication, fixation bias to factors which affect fixation of the duplication event in the population and retention biases to whether the duplication is retained in the long term. In the case of WGD, only the retention step is relevant as by definition all genes in the genome are duplicated, so there is no bias in mutation, and the entire event will be either fixed or not. Tandem duplication, on the other hand, concerns individual genes which will each differ in the factors which create each of these biases. Certain genes, for example, may be largely unable to ever duplicate, regardless of the chances of duplicate retention or fixation, due to their sequence context not being conducive to the mutational step required. Others, although they may be prone to undergoing duplication, may possess features which make them unlikely to ever be retained. Ohnologs, for example, are very rarely successfully duplicated outside of the WGD context regardless of how likely any one ohnolog is to be duplicated in the first place.

Given the differences in the biases which affect duplicate fate in the context of different duplication mechanisms, and the almost completely opposite patterns of retention the genes concerned exhibit, it is reasonable to assume that genes able to duplicate by each mechanism would show substantial differences in a number of genetic features. Tandem duplicates are more than likely shorter as they are genes which facilitate successful duplication via NAHR, ohnologs are likely more involved in interactions with other proteins as they are enriched for complex membership etc. There is a large body of work concerning differences in various features between genes which duplicate by WGD and genes which duplicate by SSD. The features considered are diverse, ranging from basic features such as genomic length and sequence features (Chapman et al., 2006; Jiang et al.; Zhu et al., 2013; Banerjee et al., 2017; Qiao et al., 2018), to various aspects of regulation (He and Zhang, 2005a; Guan et al., 2007; Amoutzias et al., 2010; Keller and Yi, 2014), gene function (Guan et al., 2007; Hakes et al., 2007; Kassahn et al., 2009; Session et al., 2016) and higher level features such as how essential a duplicate is (Gu, 2003; He and Zhang, 2006; Hakes et al., 2007; Guan et al., 2007) and adaptation and divergence following duplication (Gu, 2003; Guan et al., 2007; Qian and Zhang, 2014; Session et al., 2016; Qiao et al., 2018). The broad trend shows that ohnologs are generally more constrained than tandem duplicates in a number of ways. They are usually longer and more complex genes, with tighter regulation and slower rates of evolution.

However results have not necessarily been consistent in cases where there are multiple studies to compare for a given feature. For example, Guan et al. (2007) and Hakes et al. (2007) both find evidence of lower essentiality in yeast ohnologs while Gu (2003) finds no difference in fitness effects of deleting ohnologs versus other paralogous genes. The existence of contradictions such as this even within the same species and using broadly the same duplicate groups, raises questions about the remainder of the literature where there is a severe lack of cohesion between studies. Currently existing work has been conducted across a variety of species, mainly *S. cerevisiae* and a few plant species. There is also considerable variance in how duplicates are defined and which categories are specifically compared with some studies comparing tandem duplicates to ohnologs, others comparing ohnologs to non-ohnologs and some simply comparing older duplicates to younger duplicates. In many cases a given feature will only have one study directly examining it, so it is difficult to tell how this variation in methodology might affect overall interpretations of how duplicates differ and to collate existing work. There is also a lack of work in vertebrates relative to other species with WGD histories, which is unfortunate given that certain feature patterns are problematic to extrapolate from the given data. Yeast lacks multicellularity and a complex body plan, so patterns regarding tissue specificity, developmental functions and other features relevant in the case of vertebrates can not be determined. Plants, the other group with reasonable representation in the current literature, clearly have considerably different pressures governing the prevalence of WGD given its frequency in plant lineages and also differ from vertebrates in terms of developmental plasticity, potentially also raising some differences regarding ohnolog retention patterns. Another issue is that few studies have examined multiple features together, which raises the possibility that certain features may show differences due to their correlation to another feature which is actually causal in the difference in duplication pattern. Similarly, very few features examined currently do not show differences across duplicate categories, which begs the question of which features are actually important for dissecting the differences in duplication mechanisms, and which are merely passengers or irrelevant, perhaps artefacts of examining two highly divergent groups.

This work aims to give a more complete picture of the differences between genes undergoing different duplication mechanisms, by examining multiple features within human duplicates. This should give a more consistent idea of how ohnologs and tandem duplicates differ and, as all comparisons will be carried out in the same species with consistent definitions of duplicate categories, should also allow examination of which features may be the most important in defining these categories and assessment of underlying relationships between features.

2 Methods

Data Sources and processing

Paralog and ohnolog lists

A list of all human protein coding genes, as well as a list of human paralog pairs with their duplication timing (duplication node) was obtained from Ensembl v90 (Zerbino et al., 2018). The paralog pair list was filtered using the obtained duplication nodes to only include duplications which have occurred within the vertebrate lineage as that is the time period this work aims to focus on. A list of ohnolog pairs was obtained from Makino and McLysaght (2010). The genes in a paralogous pair were designated as ohnologs if the pair was present in the list of ohnolog pairs and as small scale duplicates if not. Genes not present in any of the paralogous pairs were designated as singletons.

Pairs designated as small scale duplication were further examined for evidence of retroduplication, as there are likely further differences in the biases affecting these two modes of duplication. Paralog pairs were excluded as potentially retroduplicated if one member of the pair has zero introns while the other has 3 or more, or if at least one member of the pair has zero introns and the other has less than 3 *and* there is no conserved microsyteny between the duplicates. Microsynteny was defined as having at least one other paralagous pair linking the surrounding region (within 5 genes either side of the genes being tested) (see Figure 1), following from methods used by Jun et al. (2009). The additional microsynteny check was used for low intron genes as there is a possibility that the second member of the pair is a tandem duplicate and reached zero introns through intron loss rather than retroduplication or, in the case where both members of a pair have zero introns, that the original parent gene had zero introns prior to duplication. The possibility of intron gain in a retroduplicated pair was not considered as intron gain is rare and losses typically outnumber gains (Babenko, 2004; Roy and Penny, 2007).

Genes contained in multiple pairs that differed in duplication mode were excluded from the analysis.

Duplication classification was repeated in the same manner for other vertebrate species (mouse, rat, pig, dog and chicken). These species were selected as information on ohnolog status for these gene sets was available from Singh et al. (2015); the intermediate stringency sets were used.

Essentiality

A measure of essentiality, the CRISPR score, was taken from Wang et al. (2015). This score is derived from a proliferation screen defined as the average $\log_2$ fold-change in the abundance of all sgRNAs from the library used which target a given gene, i.e. the change in sgRNAs causing disruption to a gene. The greater the decrease in the disruptive sgRNAs, the more essential a

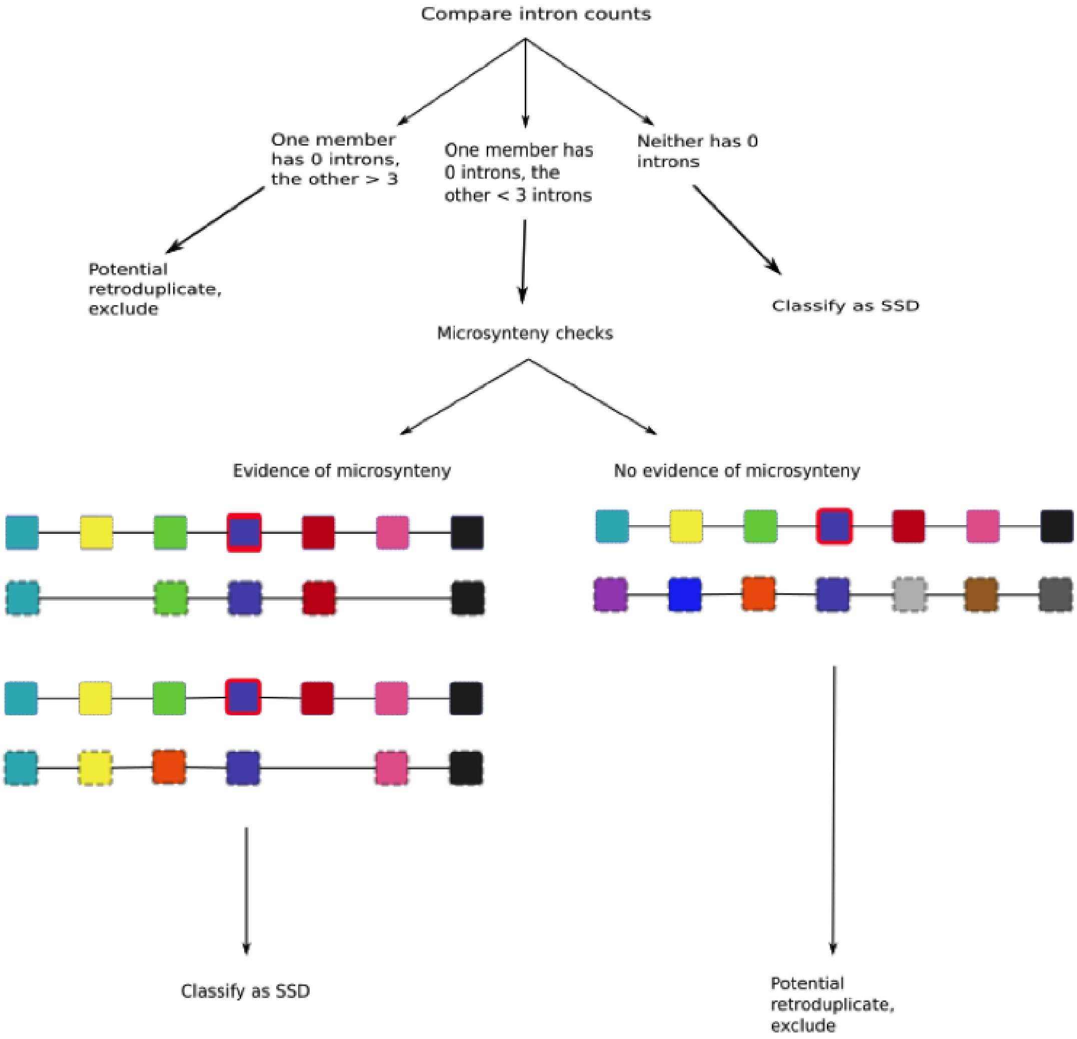
Compare intron counts
One member has 0 introns, the other > 3
One member has 0 introns, the other < 3 introns
Neither has 0 introns
Potential retroduplicate, exclude
Microsynteny checks
Classify as SSD
Evidence of microsynteny
No evidence of microsynteny
Classify as SSD
Potential retroduplicate, exclude

Figure 1: Exclusion of retroduplicates. Gene of interest is highlighted in red with neighbouring genes shown on either side. Paralogs of a given gene are shown in the same colour but with a dashed outline. There is no paralogous relationship between genes shown in different colours.

gene is, so smaller CRISPR scores indicate greater essentiality. The minimum CRISPR score across the 4 cell lines in the dataset (i.e. the maximum essentiality) was taken. The 'degree of essentiality' in figures etc. is the negative of this score.

Expression and expression specificity

Gene expression levels (in TPM) were obtained from GTEx (v7) (GTEx Consortium, 2017). Expression data for differing developmental stages was obtained through Expression Atlas (fetal expression data from FANTOM5 and expression at various stages of prenatal brain development from the Human Developmental Biology Resource, project numbers E-MTAB-3358 and E-MTAB-4840 respectively). Expression data (TPM values from RNA-seq) was available for mouse, rat, pig and chicken (at project numbers E-MTAB-3579, E-GEOD-53960, E-MTAB-5895 and E-MTAB-2797 respectively).

An average expression value across samples was taken for each tissue in the GTEx gene expression dataset. All datasets were combined and maximal expression of each gene across the available tissues/developmental stages was taken. In all cases, a cutoff of 1 TPM was used to defined genes as expressed.

Expression specificity was calculated as the tissue specificity index τ which is given by $\tau = \frac{\Sigma_{i=1}^{N}(1-x_i)}{N-1}$ where N is the number of tissues and x_i is the expression value for the i^{th} tissue scaled by the highest expression value for the gene. τ essentially represents the average difference across tissues from the maximal gene expression, scaled relative to the maximal gene expression; a value of 0 indicates a housekeeping gene while a value of 1 indicates a tissue specific gene. For the purposes of this calculation, different development stages were treated as separate tissues. This measure of tissue specificity was determined to be the most robust when benchmarked by Kryuchkova-Mostacci and Robinson-Rechavi (2016).

Other features

Protein-protein interactions (PPIs) were obtained from BioGRID v3.4.157 (Oughtred et al., 2019). All other features used (genomic length, CDS length, evolution rate, number of introns, average intron length, intron coverage, regulatory motifs, protein domains, unique protein domains, % GC content, % GC3 content) were obtained from Ensembl biomart or through the Ensembl API where available for a given species. In cases where a feature may differ between different transcripts/protein products of a gene, the value for the longest transcript was used.

Statistical methods

The Mann-Whitney U test was used for all direct comparisons of features between duplicate types and between duplicates and singletons. P-values were Bonferroni corrected for multiple testing when applicable. The effect sizes given are Cohen's r, a modified version of Cohen's D (Fritz et al., 2012).

Depletion and enrichment of GO terms in each category was determined based on the hypergeometric distribution using the hypergeom.cdf and hypergeom.sf functions in the scipy python package.

Random forest classifiers for determining feature importance were constructed using the genes which had acceptable values for all features examined with the sklearn Python package. The random forest algorithm is an ensemble machine learning method; methods of this type aim to aggregate results from a set of weaker classifiers in order to create one strong classifier. In the case of random forests, outcomes from a set of decision trees, which are individually prone to overfitting, are averaged through majority voting. Each tree is given a random sub-sample of the dataset and a random sub-set of the features. This method was selected over, for example, a regression model as it is a better choice in cases where there may be complex interactions between features.

Hyperparameters (i.e. parameters of the model set prior to training) for these classifiers were determined using the best values determined from K-folds cross validation (using 10 folds) with a model trained on the remaining features. Following this step, the hyperparameters used were the same as defaults except in the case of max_depth (10 for the WGD model, 30 for the SSD model), max_features ('sqrt' for both), min_samples_split (10 for both) and n_estimators (1200 for WGD model, 800 for the SSD model). The dataset was split 80-20 into training and test sets and the training set was then undersampled (a random selection of cases dropped so that frequency of each category is roughly equal) to account for the fact that singletons are more common than duplicates of either type.

Feature importances were calculated using the rfpimp python package, using permutation importance as the method of calculation. This method was chosen over the default feature importances from sklearn, which are based on mean impurity decrease, as this method can be biased by variable scale or number of categories (Strobl et al., 2007). Permutation importance records the drop in accuracy caused by randomly permuting each feature relative to a base accuracy.

Feature importances from these initial models were checked for negative values, which would indicate that the model is improved by permuting the feature i.e. the feature is making the model worse because it is not predictive and is adding noise. The final models were rebuilt with datasets that excluded features which showed an initial negative value in each case.

Model training and feature importance calculation was repeated ten times using the final hyperparameters and feature set with each pass using different randomly selected training and validation datasets in order to estimate how variable the importance rankings were due to randomness in the model.

Lan and Pritchard reanalysis

Duplicate pairs used in the reanalysis, along with their expression patterns and *dS* values, were obtained from the authors of the original paper (Lan and Pritchard, 2016). Pairs classified as 'unmappable' in the original analysis were excluded, as was one pair (ENSG00000097007 and ENSG00000143322) which had an abnormally high *dS* value (26.3). Gene locations and other information needed for proximity classifications (e.g. intron counts and coordinates) were obtained from Ensembl; determination of retroduplicate pairs was carried out as described in the original work. Expression correlations were calculated using GTEx v7 TPM values (GTEx Consortium, 2017) due to lack of access to the raw expression data used in the original paper.

Ohnolog datasets used were from Makino and McLysaght (2010) and Singh et al. (2015) (intermediate stringency set). The pairs used in the relaxed definition set were allowed to come from either set while strict definition pairs had to be present in both.

3 Results

Duplicate Type Classification

Of 20,277 protein coding genes included, 8,264 do not possess any paralogs from duplications in the vertebrate lineage (5,386 of these duplicated prior to the vertebrate divergence). For the remaining genes, paralog pairs were initially classified into WGD and non-WGD pairs (9,276 and 44,666 pairs respectively), with non-WGD pairs being further tested for the possibility that they arose from retroduplication. This left 39,796 pairs classified as arising from SSD and 4,870 excluded due to the possibility of retroduplication. On classifying individual genes based on the type of duplication they may undergo, the counts in Table 1 were obtained.

Table 1: Counts of genes falling into each duplication category

Duplication mode	**Count**
WGD	3647
SSD	5232
singleton	8264
Retroduplication	144
Mixed duplication mode	1990

Ohnologs generally show the highest level of constraint across features

Comparisons of features across ohnologs, tandem duplicates and singletons are shown in Figure 2. The three categories differ significantly in almost all features. The one exception to this in the case of ohnologs and tandem duplicate comparisons is essentiality, while number of regulatory motifs and %GC content do not differ significantly between ohnologs and singletons. Differences between singletons and SSDs are significant on all counts, however the difference in CDS length shows a relatively high p-value relative to the values for the ohnolog comparison to either group.

Overall, the direction of the differences observed generally indicates higher constraint on ohnologs relative to SSDs in many ways, confirming the pattern broadly shown in existing work. Ohnologs are longer, slower evolving and show greater complexity of structure and regulation with higher expression and broader expression patterns. Singletons generally show either intermediate values between the two duplicate categories or similar values to ohnologs, placing SSD duplicated genes as a relatively unconstrained group. Some exceptions to this pattern for singletons are essentiality, which is higher in singletons than either duplicate category, and a set of features where singletons show lower values than duplicates; specificity, %GC3 content and intron coverage.

The higher essentiality in singletons is in agreement with previous work, including the work which produced the essentiality estimates used here (Gu, 2003; He and Zhang, 2006; Wang et al., 2015), however previous studies have shown ohnologs to be either less essential than other duplicated genes (Guan et al., 2007; Hakes et al., 2007) or similarly essential (as found here) (Gu, 2003). All of this previous work (excluding the data source) has been conducted in yeast, however the values used here were based on cell lines rather than whole organisms so this is less likely to explain the differences in conclusions than differences in methods of estimating essentiality (the two studies which are in agreement that ohnologs are less essential used the same dataset). One point to note is that the Wang et al. (2015) dataset did not include many olfactory genes, which are likely to be tandem duplicated and so we lack estimates for a group of SSDs.

Differences between duplicate groups are generally consistent across vertebrates

To verify that these patterns of genomic feature differences are consistent in different vertebrate lineages, the comparisons were replicated in five other species (mouse, rat, dog, pig and chicken) which were available from Singh et al. (2015). Effect sizes for duplicate category comparisons by species are given in Table 2 for features where values were obtained for a given species.

Effect sizes and directions are generally consistent across the species tested and with the patterns observed in the human comparisons. Any minor differences can likely be attributed to differences in dataset completeness. For example, effect sizes generally seem much lower in the non-human species in the case of expression level comparisons, however the expression datasets for these species incorporate far fewer tissues and there is a lack of expression data for different developmental stages. The same is true of protein interactions, which have much smaller datasets in non-human species, to the point where comparison is likely meaningless (for example, only 36 dog genes have any interactions listed).

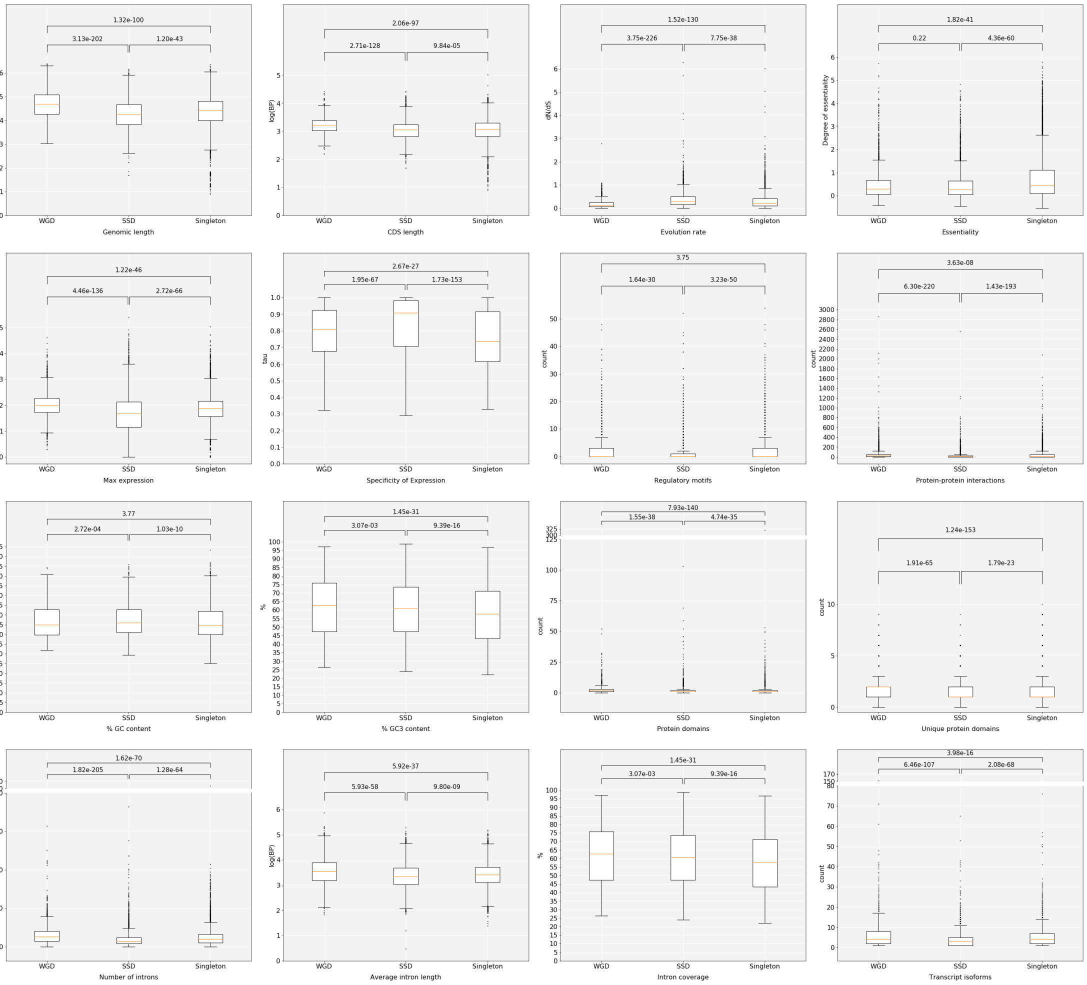

Figure 2: Comparison of gene features across duplication categories. P-values for each pairwise comparison are Bonferroni corrected across feature comparisons with n=16.

Table 2: Duplicate category comparisons by species. Effect sizes given for each comparison. In cases where the difference between the group was not significant (corrected p-value ¿ 0.05), this is indicated.

Feature	**WGD vs Singleton**					**SSD vs Singleton**					**WGD vs SSD**				
	Species					**Species**					**Species**				
	Mouse	Rat	Dog	Pig	Chicken	Mouse	Rat	Dog	Pig	Chicken	Mouse	Rat	Dog	Pig	Chicken
Genomic length	0.284	0.237	0.205	0.174	0.245	-0.049	-0.155	-0.016 (ns)	-0.038	0.095	0.248	0.246	0.182	0.14	0.158
CDS length	0.271	0.223	0.21	0.195	0.237	0.155	0.051	0.108	0.116	0.2	0.148	0.15	0.113	0.085	0.101
Evolution rate	-0.205	-0.213	-0.183	-0.166	-0.144	0.032	-0.002 (ns)	-0.022 (ns)	-0.019 (ns)	-0.099	-0.185	-0.16	-0.145	-0.116	-0.067
Max expression	-0.105	0.075	-	0.085	0.023 (ns)	-0.051	-0.045	-	-0.025 (ns)	-0.072	-0.052	0.088	-	0.081	0.069
Specificity of Expression	0.007 (ns)	0.006 (ns)	-	0.038 (ns)	0.091	0.158	0.186	-	0.167	0.23	-0.141	-0.152	-	-0.086	-0.097
PPIs	0.167	0.067	-0.001 (ns)	-	0.003 (ns)	-0.027	-0.001 (ns)	0.001 (ns)	-	-0.001 (ns)	0.151	0.051	-0.001 (ns)	-	0.003 (ns)
Intron count	0.252	0.198	0.2	0.163	0.209	-0.075	-0.15	0.014 (ns)	-0.04	0.084	0.237	0.224	0.164	0.131	0.132
Average intron length	0.167	0.142	0.114	0.086	0.155	0.05	-0.067	-0.049	-0.003 (ns)	0.055	0.112	0.148	0.129	0.064	0.101
Intron coverage	0.148	0.124	0.127	0.064	0.163	0.006 (ns)	-0.093	-0.07	-0.003 (ns)	0.02 (ns)	0.126	0.155	0.156	0.049	0.132
Transcript isoforms count	0.14	0.033	0.082	0.021 (ns)	0.017	-0.074	0.003 (ns)	0.017 (ns)	0.003 (ns)	0.011	0.165	0.023	0.06	0.013 (ns)	0.008 (ns)
Domains	0.2	-	0.217	0.177	-	0.242	-	0.306	0.231	-	0.009 (ns)	-	-0.044	-0.008 (ns)	-
Unique domains	0.196	-	0.214	0.171	-	0.238	-	0.29	0.226	-	0.011 (ns)	-	-0.028 (ns)	-0.009 (ns)	-

Table 3: GO terms enriched in each duplication category. List of enriched GO terms for each duplication category, ranked by lowest p-value. Only top 25 terms shown in each case.

WGD enriched		SSD enriched		Singleton enriched	
GO:0006468	protein phosphorylation	GO:0005576	extracellular region	GO:0005739	mitochondrion
GO:0004672	protein kinase activity	GO:0003676	nucleic acid binding	GO:0003723	RNA binding
GO:0035556	intracellular signal transduction	GO:0002376	immune system process	GO:0005654	nucleoplasm
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	GO:0006955	immune response	GO:0006364	rRNA processing
GO:0016301	kinase activity	GO:0003823	antigen binding	GO:0006281	DNA repair
GO:0016310	phosphorylation	GO:0005615	extracellular space	GO:0006412	translation
GO:0043565	sequence-specific DNA binding	GO:0004252	serine-type endopeptidase activity	GO:0005743	mitochondrial inner membrane
GO:0004674	protein serine/threonine kinase activity	GO:0002250	adaptive immune response	GO:0006974	cellular response to DNA damage stimulus
GO:0045893	positive regulation of transcription: DNA-templated	GO:0005622	intracellular	GO:0005730	nucleolus

WGD enriched		**SSD enriched**		**Singleton enriched**	
GO:0000978	RNA polymerase II sequence-specific DNA binding	GO:0050776	regulation of immune response	GO:0005840	ribosome
GO:0046777	protein autophosphorylation	GO:0006958	complement activation: classical pathway	GO:0005759	mitochondrial matrix
GO:0005667	transcription factor complex	GO:0046872	metal ion binding	GO:0070125	mitochondrial translational elongation
GO:0001077	transcriptional activator activity: RNA polymerase II	GO:0031424	keratinization	GO:0030529	intracellular ribonucleoprotein complex
GO:0045202	synapse	GO:0006355	regulation of transcription: DNA-templated	GO:0070126	mitochondrial translational termination
GO:0018105	peptidyl-serine phosphorylation	GO:0070268	cornification	GO:0005515	protein binding
GO:0051056	regulation of small GTPase mediated signal transduction	GO:0002377	immunoglobulin production	GO:0003735	structural constituent of ribosome
GO:0060070	canonical Wnt signaling pathway	GO:0030246	carbohydrate binding	GO:0042254	ribosome biogenesis
GO:0008134	transcription factor binding	GO:0005882	intermediate filament	GO:0005829	cytosol
GO:0032587	ruffle membrane	GO:0045087	innate immune response	GO:0005929	cilium

WGD enriched		SSD enriched		Singleton enriched	
GO:0018108	peptidyl-tyrosine phosphorylation	GO:0045095	keratin filament	GO:0006260	DNA replication
GO:0007411	axon guidance	GO:0030449	regulation of complement activation	GO:0019083	viral transcription
GO:0030522	intracellular receptor signaling pathway	GO:0072562	blood microparticle	GO:0006413	translational initiation
GO:0030036	actin cytoskeleton organization	GO:0006956	complement activation	GO:0008033	tRNA processing
GO:0004713	protein tyrosine kinase activity	GO:0016021	integral component of membrane	GO:0000398	mRNA splicing: via spliceosome
GO:0007389	pattern specification process	GO:0006351	transcription: DNA-templated	GO:0005762	mitochondrial large ribosomal subunit

Table 4: GO terms depleted in each duplication category. List of depleted GO terms for each duplication category, ranked by lowest p-value. Only top 25 terms shown in each case.

WGD depleted		SSD depleted		Singleton depleted	
GO:0004984	olfactory receptor activity	GO:0005654	nucleoplasm	GO:0004930	G-protein coupled receptor activity
GO:0050911	detection of chemical stimulus ... sensory perception of smell	GO:0005694	chromosome	GO:0005886	plasma membrane
GO:0003676	nucleic acid binding	GO:0016310	phosphorylation	GO:0004871	signal transducer activity
GO:0007608	sensory perception of smell	GO:0000166	nucleotide binding	GO:0007165	signal transduction
GO:0005576	extracellular region	GO:0016301	kinase activity	GO:0007186	G-protein coupled receptor signaling pathway
GO:0004930	G-protein coupled receptor activity	GO:0046777	protein autophosphorylation	GO:0004984	olfactory receptor activity
GO:0050896	response to stimulus	GO:0006468	protein phosphorylation	GO:0050911	detection of chemical stimulus involved in sensory perception of smell
GO:0003723	RNA binding	GO:0007049	cell cycle	GO:0005887	integral component of plasma membrane

WGD depleted		**SSD depleted**		**Singleton depleted**	
GO:0006955	immune response	GO:0005524	ATP binding	GO:0007608	sensory perception of smell
GO:0007186	G-protein coupled receptor signaling pathway	GO:0000784	nuclear chromosome: telomeric region	GO:0050896	response to stimulus
GO:0002376	immune system process	GO:0004672	protein kinase activity	GO:0003700	transcription factor activity: sequence-specific DNA binding
GO:0016021	integral component of membrane	GO:0030529	intracellular ribonucleoprotein complex	GO:0005622	intracellular
GO:0005634	nucleus	GO:0016740	transferase activity	GO:0016021	integral component of membrane
GO:0042742	defense response to bacterium	GO:0003723	RNA binding	GO:0006958	complement activation: classical pathway
GO:0004871	signal transducer activity	GO:0004674	protein serine/threonine kinase activity	GO:0003823	antigen binding
GO:0006958	complement activation: classical pathway	GO:0051301	cell division	GO:0007155	cell adhesion
GO:0003823	antigen binding	GO:0006364	rRNA processing	GO:0006355	regulation of transcription: DNA-templated
GO:0006364	rRNA processing	GO:0007411	axon guidance	GO:0004252	serine-type endopeptidase activity

WGD depleted		SSD depleted		Singleton depleted	
GO:0005739	mitochondrion	GO:0016567	protein ubiquitination	GO:0002376	immune system process
GO:0031424	keratinization	GO:0046982	protein heterodimerization activity	GO:0003676	nucleic acid binding
GO:0005882	intermediate filament	GO:0070126	mitochondrial translational termination	GO:0006811	ion transport
GO:0005694	chromosome	GO:0005739	mitochondrion	GO:0002250	adaptive immune response
GO:0030529	intracellular ribonucleoprotein complex	GO:0070125	mitochondrial translational elongation	GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules
GO:0005615	extracellular space	GO:0000775	chromosome: centromeric region	GO:0006956	complement activation
GO:0000786	nucleosome	GO:0004527	exonuclease activity	GO:0072562	blood microparticle

Enriched/depleted functions show a clearer complementary pattern between singletons and ohnologs than between SSDs and ohnologs

Results for GO terms significantly enriched in each duplication category are consistent with previous work in both human and other species (Hakes et al., 2007; Makino and McLysaght, 2010; Roux et al., 2017; Qiao et al., 2018). The top 25 terms for each category (determined by p value) are listed in Table 3 and Table 4. Ohnologs generally show enrichment for regulatory functions as well as developmental and nervous system related functions. The top 25 enriched terms shown in Table 3 reflect these trends with each of these categories represented. Phosphorylation seems to be a particularly over-represented regulatory function in this class, which is consistent with previous reports that ohnologs are enriched for kinases (Hakes et al., 2007; Qiao et al., 2018). The developmental aspect is shown in highly significant enrichments for functions related to cell migration and patterning.

SSDs are generally known to show enrichment for functions relating to external factors such as immunity and perception of various stimulus (taste, smell etc.). Results here confirm this pattern, with most significant results for enrichment in this group coming from terms concerning immune or host-pathogen interactions. Also present are several terms relating to keritinization, the process by which the outermost cells of the epidermis have their cytoplasm converted to keratin. Although perhaps not immediately apparent, this is likely related to the enrichment for immune functions, as keratinization is part of innate immunity, with many hyperkeratinization disorders having an autoimmune component (Bos et al., 2005; Akiyama et al., 2017).

Interestingly, both duplication categories show a highly significant depletion for mitochondrial functions. Singletons show enrichment for these mitochondrial functions as well as basic cellular functions such as translation and transcription and for cellular components in general such as the ribosome and cytosol.

In terms of significantly depleted terms for each category, there are essentially no new terms relative to the enriched terms within the top 25, as each category seems to show depletion for terms that are enriched in the others. Singletons show depletion for functions such as immunity and reaction to external stimulus (enriched in SSDs) and functions associated with signal transduction (enriched in WGDs). Similarly, both duplication classes are depleted for functions enriched in the other or in singletons.

This sort of mirrored enrichment/depletion is not necessarily surprising given the largely opposite retention patterns for genes undergoing WGD vs SSD. However, this pattern does not fully hold on looking at the overlap in enrichment/depletion categories across all significantly enriched/depleted terms. Figure 3 shows the overlap in these sets of terms. What is immediately clear is that, for all three categories, the terms they show enrichment for are largely unique to that category, with the majority not seen as enriched or depleted in the others. Depleted terms show much greater overlap with the sets of terms enriched in other categories. Interestingly, the biggest, and most straightforward, overlaps in this respect seem to be between ohnologs and singletons (in both the enrichment and depletion sets for ohnologs, about a third show the opposite trend in singletons), rather than between ohnologs and SSDs.

This is actually the more intuitive result if we consider the biases affecting duplicate production and survival i.e. mutation, fixation and retention. Genes that have duplicated by WGD and genes that have remained in single copy are separated only the the bias in their retention as all genes will, by definition, have been duplicated and the duplicates fixed following a successful WGD event. Thus, any functions which are likely to be retained should be enriched in ohnologs and depleted in singletons. The same applies to functions which do not promote ohnolog retention; these should show depletion in ohnologs and enrichment in singletons, although the relationship may be complicated in this case as some of these functions are likely permissible in tandem duplicates.

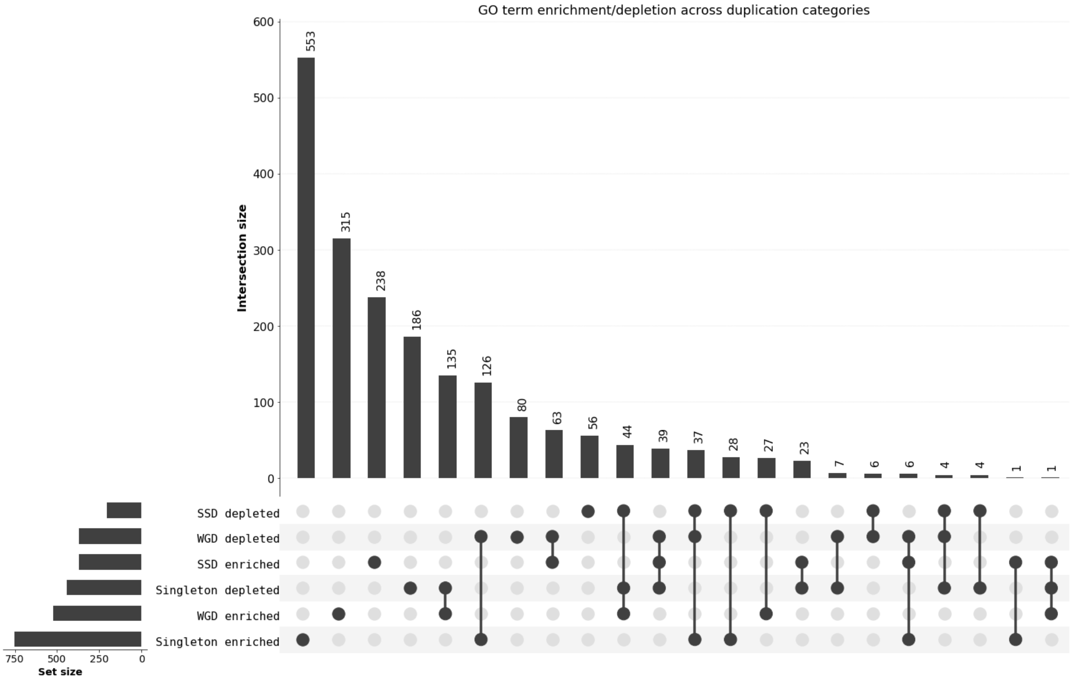

Figure 3: Overlap in functions enriched/depleted in different duplication categories. Overlap of GO terms significantly enriched or depleted in each duplication category. The categories overlapping in each set are indicated at the base of the plot.

Differences in duplicate age between SSDs and ohnologs are unlikely to be a major contributor to observed effects

One possible alternative explanation for the observed differences between duplicate is that they may reflect features which affect the longevity of duplicate retention independent of the mode of initial duplicate creation. This is a possibility as all ohnologs considered here originate from older duplication events of roughly the same age, however duplication events creating the tandem duplicates are spread across the vertebrate lineage up as far as human specific duplications, and so incorporate much younger duplications. Differences in the features of SSDs retained for different lengths of time have been reported previously in Woods et al. (2013), where it was demonstrated that essential genes were more likely to be retained in the long term despite an enrichment for non-essential genes in younger duplicates.

Duplicate age was tested as a possible confounder by considering genes which resulted from tandem duplications, but occurred at a similar time to the whole genome duplications. This should remove any effect due to mode of duplication and allow examination of the effect of duplicate age, assuming that the set of SSDs used is free from any potential ohnologs which have been mis-categorised. In order to be confident in the purity of the set used, SSDs which follow the duplication pattern of ohnologs (i.e. duplicated at the base of the vertebrate tree and not since) were removed (1952 genes).

Initially, the entire set of tandem duplicates (with potential ohnologs removed) was checked for any correlation between the features examined and duplicate age Figure 4. Although a number of features do show significant correlations, with 6 features showing $\rho > 0.1$ (CDS length, max expression, specificity of expression, PPIs, % GC3 content and number of introns), only specificity of expression shows any appreciable correlation (ρ = -0.27), indicating a negative relationship between tissue specificity and duplicate age. Further testing was carried out by comparing tandem duplicates of ohnolog age with younger tandem duplicates, separated by age. Results of these comparisons are presented in Table 5.

In almost all cases, the comparison between ohnolog-age SSDs and younger SSDs shows a much reduced or even opposite effect to the comparison between ohnologs and younger SSDs across age categories. Although there are still significant differences between the older and younger categories of SSDs, the reduction in effect size relative to the ohnolog comparison suggests that the effect observed can not be entirely explained by differences in duplicate age. Also, if duplicate age were the main factor behind these differences, we could expect to see increasing effect size as the difference in duplicate age increases. Such an effect is not particularly clear from these results (although it is potentially present in the cases of CDS length, expression specificity and max expression, somewhat confirming the correlation analysis of these features).

However, effect size is not consistent across the age categories either. A likely explanation for this is that this method is flawed in that it involved drawing somewhat arbitrary lines between

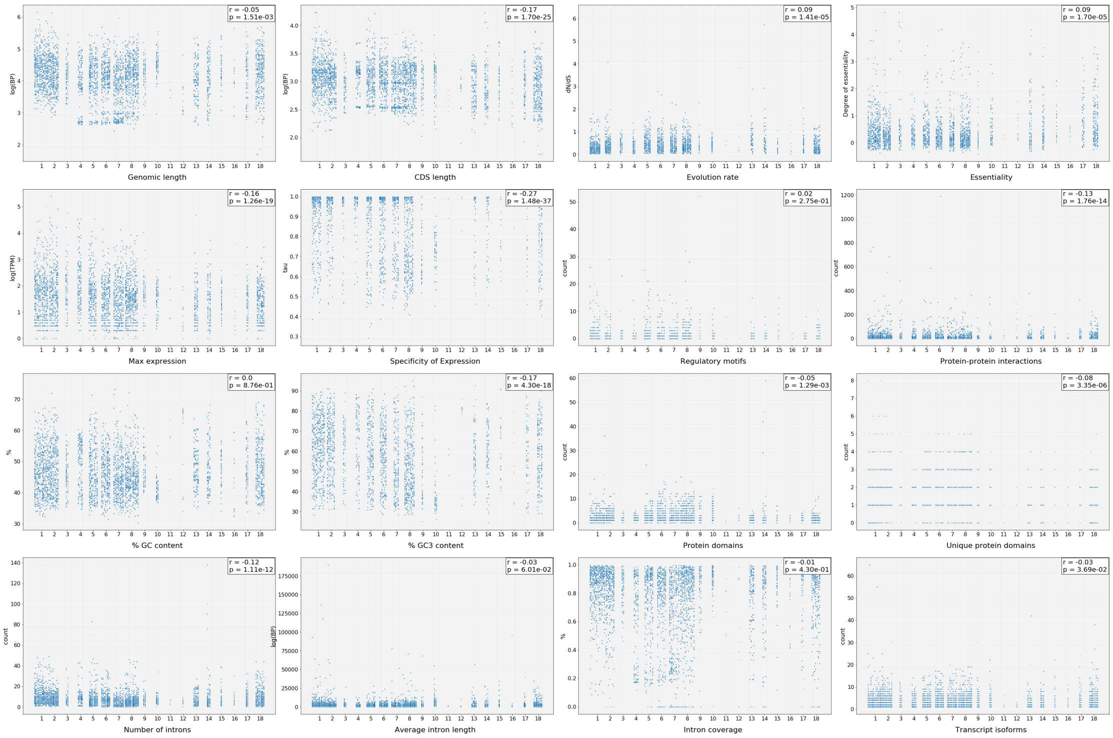

Figure 4: Relationship between genetic features and duplicate age. Values for ρ and p are given for each plot. The x axis represents duplicate age according to duplication node in Ensembl (1: Vertebrata, 2:Euteleostomi, 3:Sarcopterygii, 4:Tetrapoda, 5:Amniota, 6:Mammalia, 7:Theria, 8:Eutheria, 9:Boreoeutheria, 10:Euarchontiglires, 11:Primates, 12:Haplorrhini, 13:Simiiformes, 14:Catarrhini, 15:Hominoidea, 16:Hominidae, 17:Homininae, 18:*Homo sapiens*).

Table 5: Comparisons for duplicate groups split by age. Effect sizes are given for pairwise comparisons between the ohnolog-age duplicate groups and genes where the oldest duplication falls within the range of the duplication nodes given at the top. Non-significant differences are indicated with ns (cases where corrected p-value > 0.05 (n=16))

Feature	**Ohnolog-age duplicate group**	**Sarcopterygii-Amniota**	**Mammalia-Euarchontoglires**	**Primates**
Genomic length	Ohnologs	0.297	0.417	0.308
	SSDs of ohnolog age	0.199	0.283	0.19
	SSDs of ohnolog age (filtered)	0.132	0.163	0.096
	Potential ohnologs	0.253	0.335	0.242
CDS length	Ohnologs	0.208	0.336	0.345
	SSDs of ohnolog age	0.134	0.23	0.281
	SSDs of ohnolog age (filtered)	ns	0.117	0.227
	Potential ohnologs	0.179	0.278	0.332
Essentiality	Ohnologs	ns	0.109	-0.09
	SSDs of ohnolog age	ns	0.106	-0.116
	SSDs of ohnolog age (filtered)	ns	ns	-0.232
	Potential ohnologs	ns	0.134	-0.108
Max expression	Ohnologs	0.136	0.374	0.374
	SSDs of ohnolog age	ns	0.264	0.295
	SSDs of ohnolog age (filtered)	-0.128	ns	0.127
	Potential ohnologs	0.104	0.339	0.378
Evolution rate	Ohnologs	-0.339	-0.452	-0.3
	SSDs of ohnolog age	-0.223	-0.287	-0.164
	SSDs of ohnolog age (filtered)	-0.121	-0.122	ns
	Potential ohnologs	-0.278	-0.346	-0.213
% GC content	Ohnologs	ns	0.046	-0.093

Feature	Ohnolog-age duplicate group	Sarcopterygii-Amniota	Mammalia-Euarchontoglires	Primates
	SSDs of ohnolog age	ns	0.125	ns
	SSDs of ohnolog age (filtered)	ns	0.109	ns
	Potential ohnologs	ns	0.135	ns
% GC3 content	Ohnologs	ns	0.185	ns
	SSDs of ohnolog age	0.061	0.243	0.079
	SSDs of ohnolog age (filtered)	ns	0.25	ns
	Potential ohnologs	0.074	0.259	0.093
Specificity of Expression	Ohnologs	-0.286	-0.108	-0.08
	SSDs of ohnolog age	-0.225	ns	ns
	SSDs of ohnolog age (filtered)	ns	0.244	0.294
	Potential ohnologs	-0.285	ns	ns
Number of introns	Ohnologs	0.266	0.446	0.269
	SSDs of ohnolog age	0.155	0.321	0.137
	SSDs of ohnolog age (filtered)	0.122	0.252	0.078
	Potential ohnologs	0.192	0.356	0.171
Intron coverage	Ohnologs	0.211	0.306	0.15
	SSDs of ohnolog age	0.146	0.223	0.061
	SSDs of ohnolog age (filtered)	0.153	0.186	ns
	Potential ohnologs	0.17	0.244	0.075
Average intron length	Ohnologs	0.229	0.277	0.233
	SSDs of ohnolog age	0.167	0.19	0.164
	SSDs of ohnolog age (filtered)	0.128	0.108	0.108
	Potential ohnologs	0.208	0.225	0.201

Feature	**Ohnolog-age duplicate group**	**Sarcopterygii-Amniota**	**Mammalia-Euarchontoglires**	**Primates**
Regulatory motifs	Ohnologs	0.11	0.136	0.127
	SSDs of ohnolog age	0.078	0.085	0.098
	SSDs of ohnolog age (filtered)	ns	ns	ns
	Potential ohnologs	0.105	0.114	0.126
Protein domains	Ohnologs	0.086	0.045	0.225
	SSDs of ohnolog age	ns	-0.057	0.144
	SSDs of ohnolog age (filtered)	ns	-0.07	0.15
	Potential ohnologs	ns	-0.055	0.159
Unique protein domains	Ohnologs	0.11	0.095	0.248
	SSDs of ohnolog age	ns	ns	0.15
	SSDs of ohnolog age (filtered)	ns	ns	0.163
	Potential ohnologs	ns	ns	0.164
Transcript isoforms	Ohnologs	0.21	0.259	0.253
	SSDs of ohnolog age	0.122	0.136	0.16
	SSDs of ohnolog age (filtered)	ns	ns	ns
	Potential ohnologs	0.18	0.195	0.22
Protein -protein interactions	Ohnologs	0.259	0.415	0.351
	SSDs of ohnolog age	0.123	0.244	0.223
	SSDs of ohnolog age (filtered)	ns	ns	0.088
	Potential ohnologs	0.189	0.316	0.292

duplicate ages, possibly giving spurious results. For example, in the case of % GC content, there is no reason currently to think that tandem duplicates at the base of the vertebrate tree would be different specifically to duplicates in the mammalian lineage that pre-date the divergence of primates, but no other age category. Additionally, the method used to assign a 'duplicate age' to a given gene, taking the oldest duplicate node of the gene, does not account for gene families; a method such as the one used in Woods et al. (2013), where only duplicate pairs with no other close paralog were used, may be more suitable. However, as the method currently stands, there is no evidence that differences in duplicate age fully explain the differences observed between SSDs and ohnologs.

Feature importance rankings show ohnologs are defined by features related to constraint, SSDs are defined by features related to gene complexity

Following on from the results of the functional comparisons, it was decided that two random forest classifiers categorising each duplication category versus singletons would give the most easily interpreted measure of the relative importance for each feature towards duplicate creation and retention.

The ranking of features by importance for each model, along with plots showing how dependent features are on each other, are shown in Figure 5. The WGD based model shows high, relatively constant, rankings for evolution rate and essentiality, followed by specificity, with most other features largely grouping together. The SSD based model always shows specificity as the top ranking feature, with two groups of features showing lower importances. The higher ranking group mostly includes features relating to protein size and complexity of function. The observation that specificity is an important feature in distinguishing both types of duplicates from singletons is likely a reflection of the functional profile of these duplicates relative to singletons. Ohnologs are enriched for developmental and neural functions, while SSDs are enriched for immune functions, both of which are highly specific relative to singleton functions, which were mostly cellular components

As dependencies between features can affect measures of feature importance, the mean rankings should be interpreted in light of how predictive other features are of each individual feature. This is less of an issue in the WGD model, where only genomic length and number of unique domains are really predictive of each other, however this relationship may affect the rankings of these two features. Both are quite variable and their mean ranks are roughly the same, possibly indicating that feature importance is being 'shared' between the two features; as they are dependent on each other, adding one offers little improvement in prediction accuracy once the other is already included. Thus, each repeat of the model may assign importance to one or the other at random, causing the ranking of both to fluctuate. This is more of an issue in the SSD model, where there are multiple pairs of dependent features (genomic length and intron coverage, CDS length and intron count, domains and unique domains). The first two of these pairs show similar rankings and likely are another example of sharing of importance. However domains and unique domains show more separation with domains ranking higher, probably reflecting the fact that domains is more predictive of unique domains than vice versa. These dependencies not only potentially cause underestimations of feature importance, but also make it difficult to interpret which feature is actually important and which is just correlated. This issue may need to be examined further.

Additionally, the accuracy of the current models could be improved upon to potentially obtain more reliable estimates of feature impotance. Currently the WGD model has an accuracy of 68% and shows some bias towards classifying genes as ohnologs (singleton accuracy:65%, ohnolog accuracy:73%), while the SSD model also has an accuracy of 68% but shows less bias (singleton accuracy: 69%, SSD accuracy: 66%). This could be improved with further tuning of hyperparameters, but also we may need to look at whether or not we are giving the model homogeneous groupings, particularly in the case of the SSD model. For example, some singletons may be capable of initially duplicating by tandem duplication (e.g. the gene are short enough not to pose an issue for successful duplicate by NAHR) but will never be retained (e.g. the gene is dosage sensitive), or vice versa. Successful duplicates form a sort of hybrid of these two singleton groups and so it may be difficult to distinguish the two successfully. One possible way to examine if this is the case in the singleton group could be some form of clustering analysis, as used in Jiang et al.. The distribution of, for example, genomic and CDS lengths of singletons in Figure 2 may give support to this idea as singletons are generally longer than SSDs but do also show a tail of shorter genes, possibly genes which could (mechanistically) duplicate but are not retained.

Inadvertent inclusion of ohnologs in SSD datasets may have unintended effects: re-analysis of Lan & Pritchard

Given the differences between tandem duplicates and ohnologs noted here and in previous work, care should be taken in the definition of duplicate datasets to ensure that no unintended assumptions are made regarding the mode of duplicate creation. As SSDs and ohnologs differ in many respects, if ohnologs are unknowingly included in a dataset it could lead to differences between ohnologs and other duplicates being attributed to other factors.

One such example of these unintended consequences is examined here with re-analysis of work carried out in Lan and Pritchard (2016), in which the effects of genomic separation on shared expression constraint is examined in tandem duplicates. It is asserted that when tandem duplicates arise they are subject to co-regulation due to their proximity and that this constraint is relieved by genomic rearrangements at a later point which separate these duplicates. At this point, the duplicates may begin to diverge in expression and allow sub- or neo-functionalization to take place.

Results in the original work support the hypothesis that duplicate proximity may delay divergence; duplicates which have undergone sub-/neo-functionalization under the definition given

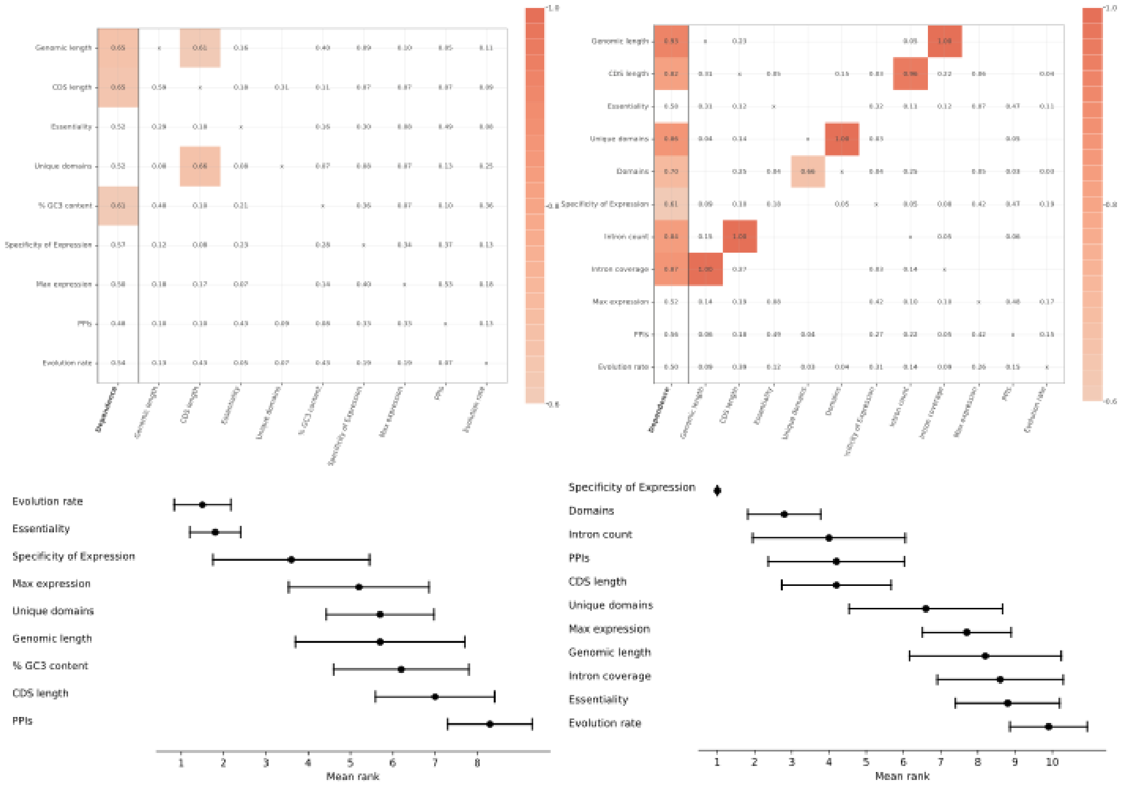

Figure 5: Feature dependencies and importance rankings. Plots for the WGD model are shown on the left, SSD model on the right. Dependence plots show how easily predicted each variable is in the left hand column, with a breakdown of how important each of the other predictors given across the rows. Importances are bounded at 0 and 1. The lower plots show the mean importance ranking of features across the 10 replicate runs; error bars represent standard deviation.

are enriched in genomically separated duplicates, with an increase in the proportion of these duplicates over time, and there is a significant difference in expression pattern correlation between separated and unseparated duplicates.

However, the definition of the duplicate dataset used here is problematic. Firstly, the gene pairs considered were limited to reciprocal best hit pairs, which is needlessly restrictive and likely biases the dataset towards highly conserved or very young duplicate pairs. Secondly, there is no explicit exclusion of ohnolog pairs, which is particularly concerning given the bias towards highly conserved pairs. The presence of WGD in the evolutionary history of humans is mentioned, however the possibility of the presence of ohnologs is dismissed as unlikely based on the synonymous divergence of the duplicate pairs considered i.e. it is assumed these duplicates are too young and likely post-date the WGD events. Synonymous divergence is likely a flawed method for dating duplication events in datasets which contain a mixture of ohnologs and tandem duplicates. Ohnologs are known to have higher sequence constraint than tandem duplicates, although most of the work on this topic has focused on non-synonymous divergence, which shouldn't affect this dating method, homogenisation between ohnologs may lead to underestimates in dating (as mentioned in Chapman et al. (2006)). Additionally, ohnologs possess functions which generally show high constraint in codon usage (Roux et al., 2017), a constraint which would affect synonymous divergence. For these reasons it may be inappropriate to measure divergence times of tandem duplicates and ohnologs on the same scale when looking at synonymous change, and to calibrate dating for dS estimates using singleton orthologs, as was carried out here.

Failure to remove ohnologs from the dataset seriously affects the initial assumptions of the model under test and also how results should be interpreted. Most of the pairs defined to be sub-/neo-functionalised were older and present on separate chromosomes, which was interpreted as evidence of delayed divergence with constraint relieved by duplicate separation. However, if we assume the presence of ohnologs, which originate on separate chromosomes by definition and are older than the tandem duplicates, it becomes less clear if these trends are supportive of the hypothesis or simply reflective of differences between ohnologs and tandem duplicates in terms of the likelihood of sub-/neo-functionalization.

Presence of ohnologs in the original dataset was checked at two levels of stringency, showing 394 pairs classified as ohnologs under the strict definition (27% of the dataset) and 605 pairs under the more relaxed defintion (42%). It is clear from Figure 6 that many of the pairs classified as undergoing sub-/neo-functionalization in the original paper are, in fact ohnologs and that the larger proportion of older pairs in this category is likely just the ohnolog pairs; from visual inspection there does not seem to be any noticeable increase in sub-/neo-functionalisation as age increases in the dataset with ohnologs removed. Similarly, Figure 7 does not seem to show any increase in the proportion of duplicates on different chromosomes with age once ohnologs have been removed.

Given these observations, it seems at least plausible that the reported link between duplication separation and sub-/neo-functionalization may have been a consequence of the inclusion of ohnolog pairs. Many of the most relevant pairs to the testing of this model (i.e. older, separated pairs defined as sub-/neo-functionalized) are removed on filtering for ohnologs. In fact, ohnologs in this dataset are significantly enriched for sub-/neo-functionalised pairs relative to the non-ohnologs ($p = 3.60 \text{ x}10^{-26}$, odds ratio = 3.82 for strict definition; $p = 9.14 \text{ x}10^{-41}$, odds ratio = 5.24 for relaxed definition; Fisher's exact test), which supports this idea. In spite of this, separated duplicates in the remaining dataset remain significantly enriched for sub-/neo-functionalized pairs relative to pairs on the same chromosome, although the effect is reduced ($p = 5 \text{ x}10^{-23}$ in Lan & Pritchard; $p = 1.99 \text{ x}10^{-15}$, odds ratio = 3.74 with strict definition ohnologs removed; p = 0.0007, odds ratio = 2.44 with relaxed definition ohnologs removed; Fisher's exact test). Similarly, on recreating the multiple regression model used to control for the effect of

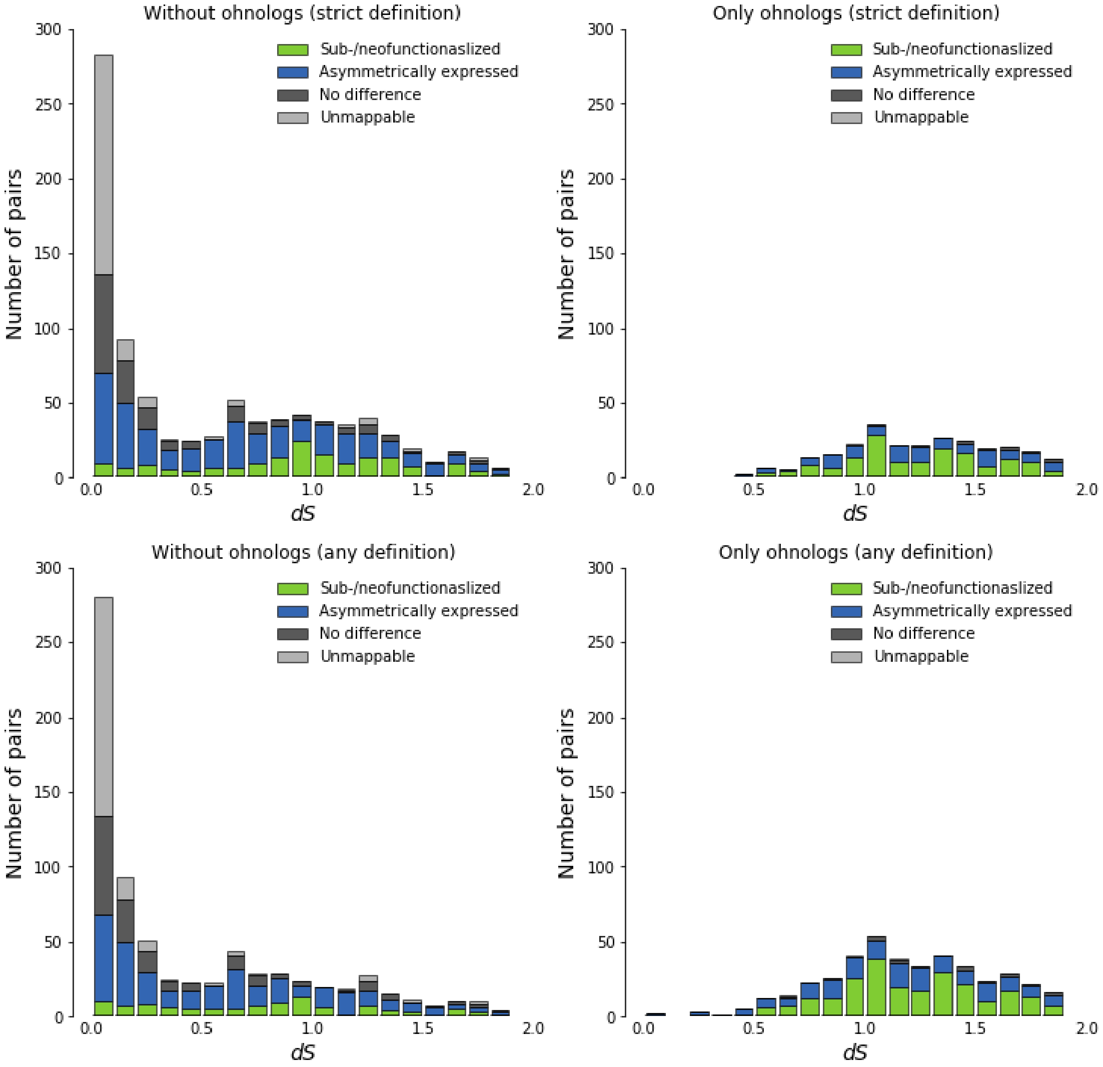

Figure 6: Expression pattern categories by age on removing ohnologs. Separation of the dataset into non-ohnologs and ohnologs shows that older sub-/neo-functionalized pairs are the main group affected by removal of ohnologs. This figure should be directly compared to Lan and Pritchard (2016) figure 2A.

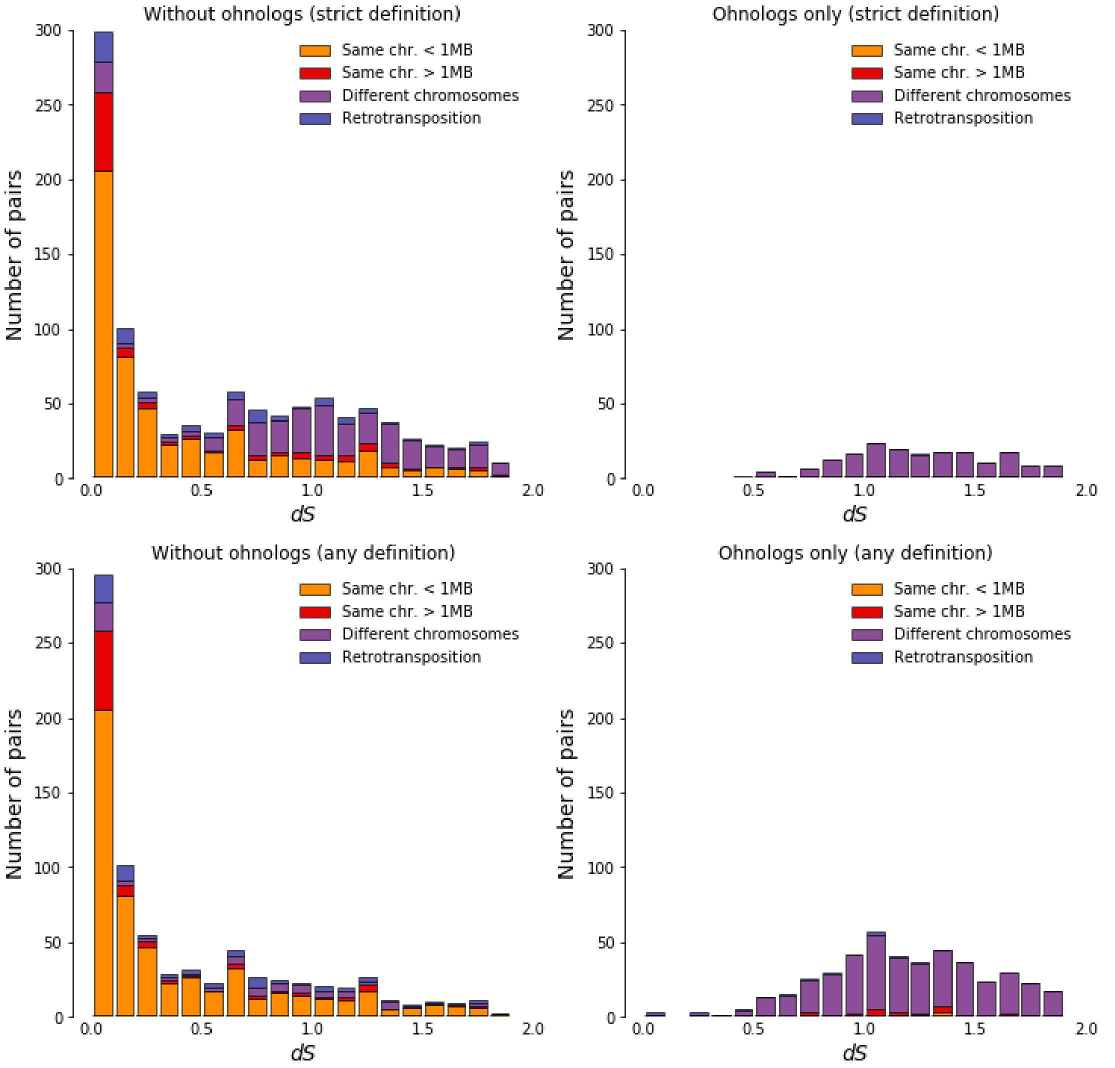

Figure 7: Duplicate proximity by age on removing ohnologs. Removal of ohnologs primarily affects older, separated duplicates. This figure should be directly compared to Lan and Pritchard (2016) figure 3A.

duplicate age, we find that separation still has an effect on expression correlation independent of duplicate age, but the significance and size of this effect is reduced on removal of ohnologs under both definitions. In Table 6, we can see that the coefficients for genomic separation (as well as the significance of the associated p-values testing H_0: coefficient $= 0$) are reduced on removal of ohnologs.

An interesting point which is not fully explored in the original paper is that (for the full dataset) the positive interaction term in this model implies *dS* (duplicate age) has no real effect on the expression correlation of separated duplicates. We can see this if we consider that the regression equation takes the form $correlation = \alpha - 0.207 * dS - 0.403 * distance + 1.98 * dS * distance$, where α is the intercept. As distance is a binary variable it is coded as same chromosome = 0 and different chromosome = 1. So, the equation simplifies to $correlation = \alpha - 0.207 * dS$ in the case where duplicates are on the same chromosome and $correlation = \alpha - 0.207 * dS - 0.403 + 0.198 * dS$ in the case where they are located on different chromosomes. Thus, for separated duplicates, the dS term and the interaction term essentially cancel out (in the original paper the co-efficients were -0.14 and 0.14 so this effect was even more pronounced). This is an interesting implication of this model as it would mean that, once duplicates have separated, further divergence has no effect on their expression correlation. This effect is lessened by the removal of ohnologs as this interaction term decreases while the dS term is largely unaffected. This may indicate that this implication is linked to the presence of ohnologs. For example, if dS is not a determinant of expression divergence in ohnolog pairs, then it would appear as if separated pairs were unaffected by further divergence when in fact the observation is due to most separated pairs being ohnolog pairs. This idea is supported by the fact that there is no significant correlation between expression correlation and dS for the ohnolog groups (r = 0.08, p = 0.1 for strict definition; r = 0.03, p = 0.5 for relaxed definition, Pearson correlation). However, there is a weak correlation between expression correlation and dS for separated duplicates in the datasets with ohnologs removed (r = -0.16, p = 0.008 for strict definition; r = -0.12, p = 0.27 for relaxed definition, Pearson correlation). The correlation is not statistically significant in the case of the dataset where relaxed defintion ohnologs are removed, however this may be due to lack of statistical power as this datset is lacking in pairs located on different chromosomes following the removal of ohnologs. Taking these results together, it seems likely that the interaction between dS and genomic separation in the original model was likely mostly an artifact of the inclusion of ohnologs rather than evidence for the radical effects of breaking synteny between duplicates, as claimed in the original paper.

Overall, it is not implausible that genomic separation of duplicate pairs allows greater expression divergence and therefore facilitates processes such as sub-/neo-functionalization. However, much of the evidence put forward to claim that sub-/neo-functionalization is delayed until separation occurs is affected by the inclusion of ohnologs in the dataset. Reassessing the results with ohnologs removed is also challenging in some ways as removal of ohnologs causes a severe drop in duplicate pairs which would be of importance to testing this hypothesis (older, separate pairs). The best approach would be to expand the definition of duplicate pairs beyond reciprocal best hits in order to have a more complete set to test on. This would also likely give a better time range as the reciprocal best hit requirement has given a large number of young duplicates (see Figure 6 for the proportion of duplicate pairs where RNA-Seq reads could not be mapped unambiguously at the lower dS ranges). While this requirement may give more reliable pairs, it also excludes older or faster evolving pairs. Many retained duplicates are faster evolving and

and here we are considering longer term fates of duplicates such as sub-/neo-functionalization, meaning that over-excluding either of these categories will not result in a good dataset for the tests required here.

Table 6: Linear regression models for each dataset used. P-values given are from Wald tests of the coefficients. Pairs which were classified as retroduplicated or which had their expression pattern quantified as 'unmappable' in the original analysis were excluded.

Variable	**Full dataset**		**Strict ohnologs removed**		**Relaxed ohnologs removed**	
	Coefficent	P-value	Coefficent	P-value	Coefficent	P-value
dS	-0.207	4.2x10^{-18}	-0.200	1.9x10^{-16}	-0.199	1.4x10^{-14}
Distance	-0.403	1.4x10^{-29}	-0.323	4.1x10^{-12}	-0.255	1.9x10^{-4}
dS*Distance	0.198	3.2x10^{-12}	0.137	5.3x10^{-5}	0.142	8.9x10^{-3}

4 Discussion and Conclusions

The results presented here broadly agree with existing work; ohnologs are more constrained and complex than genes which are capable of undergoing SSD and singletons are roughly intermediate between the two. This confirms the general trend which was indicated when taking previous work as a whole, but which was not abundantly clear due to the fragmented nature of existing findings. This issue with the existing body of work also precluded an assessment of relative importance of the observed differences in determining duplicate category, a topic which is examined here.

The importance rankings of features obtained also corroborate the trend of higher constraint in ohnologs/ lower constraint in SSDs, with features important in conservation such as evolution rate highly important in prediction of ohnologs and lower gene complexity a major factor in predicting SSDs. However, in both cases the accuracy of the model could be improved and therefore give more reliable results in this regard. One source of error in the case of the SSD model could be that the group of singletons under consideration are not a homogeneous group with regards to the factors which have prevented their duplication. Clustering analysis could be revealing in looking for sub groups.

Previously reported trends in enrichment/depletion of particular functional categories across duplication categories were also confirmed and expanded to examine the overlap in functional categories between duplicate groups and singletons. This allowed for some insight into how different biases affecting duplicate occurrence and retention might interact to form the patterns we observe here.

With these basic differences between duplicate groups confirmed, it should be possible to take forward the insights gained here to ask more complex questions about the nature of gene duplication and what factors exactly differentiate genes which duplicate by different means. For example, it may be possible to extract meaningful biological explanations for these basic patterns, i.e. to find out why we observe the trends we do. Possible examples of this might include associations between certain patterns of basic features and the functional categories associated with each type of duplication. Looking specifically at the results presented here, we find that SSD related genes are faster evolving and also that these genes are enriched for functions which are associated with faster evolving genes, such as immune functions, where an arms race situation exists (Marques and Carthew, 2007). This association lends itself to an explanation for both of these observations. As we know, functional redundancy exists at the point of creation for SSDs and a duplicate copy of a gene is very likely to be lost unless some pressure is created to retain it. This pressure is more likely to be created in situations where the gene is already fast evolving and there is pressure for continual changes. A similar example of combining observations to gain greater insight exists in the work presented in Roux et al. (2017). Here, an enrichment for nervous system expressed genes was observed in ohnologs retained in zebrafish following the teleost WGD. Other differences between genes with copies retained and not retained were also observed such as differences in divergence rate, codon usage bias and expression level, however these trends were found to only hold in the case of neural genes. This was interpreted as the neural expression patterns of these genes driving the other trends, likely due to increased pressure in neural genes to avoid translational error which could lead to protein aggregates and neural cell death. Both of these are examples of how observations regarding basic differences in duplicate categories can be combined to give greater insight into the processes shaping these patterns.

The same study discussed above (Roux et al., 2017) also addressed another open question in gene duplication which asks what the direction of the cause-effect relationship is in the case of differences between duplicate groups. In this specific case, the question of interest was whether neural expression patterns predisposed genes to being retained following WGD or retained genes had a tendency to gain neural expression patterns after the fact. This was answered by examining expression patterns in an outgroup which had not undergone this WGD (mouse) and it was found that mouse orthologs of the ohnologs examined also showed a strong bias for neural tissue expression, supporting the idea that this pattern is a result of biased retention rather than new expression patterns gained after the fact.

The interactions between gene function/expression pattern and the trends in basic features which have long been observed, and the question of whether these trends affect the initial duplication/retention process or result from it are both relatively unexplored questions in the topic of gene duplication. The results of this work provide valuable insights into basic differences between duplicates that will allow for more complex questions such as these to be addressed going forward.

Future directions

As discussed throughout, there are several ways that the current methods used here could be refined to give clearer and more reliable results. These refinements will be looked into going forward, however there are also several ways that the work completed so far could be built upon. Firstly, as discussed above, there are open questions regarding gene duplication which are possible to examine using the foundation laid here. Secondly, there is further confirmatory work that could be carried out.

Building on the current work could begin with examining the question of direction of causation. One of the more interesting feature differences to examine in this case would be the higher evolution rate observed in tandem duplicates, specifically because there are plausible mechanisms for higher evolution rate to either promote or result from SSD. The long accepted explanation has been that rate of evolution of a gene increases post-duplication due to the redundancy created by the duplicate copy releasing one copy from functional constraint. However, there has been little investigation into the alternative explanation that genes with an initially higher rate of evolution may be more likely to duplicate. Although attempts have been made in Davis and Petrov (2004) and O'Toole et al. (2018), the two studies give conflicting results. O'Toole et al. found that, in the case of primates, duplicable genes have higher rates of evolution prior to the duplication event, while Davis and Petrov found that, in the case of *S. cerevisiae* and *C. elegans*, singleton genes are faster evolving. These contradictory results may be explained by either differences in method or by real biological differences in the species considered. In both cases, an ancestral rate was approximated by looking at species other than the one where the duplication event occurred, however O'Toole et al. used outgroup species closely related to the species where duplicability was inferred (macaque and gibbon), while Davis and Petrov used species for rate estimation that were highly diverged (*D. melanogaster* and *A. gambiae*). By restricting the test group to genes which have orthologues across such a large timespan, Davis and Petrov may not have used a dataset which is entirely representative of duplicate genes as a group. Equally, the differences in results between the two studies may be due to real biologial differences in how selection forces function in primates versus yeasts/nematodes, as effective population sizes differ greatly between these groups. In order to clarify the situation, we intend to carry out similar work to that in O'Toole et al. (2018) in the *Drosophila* lineage, using closely related species to estimate an ancestral rate of evolution. Results of this work should determine if there are real differences between species due to population size.

The same logic of looking at outgroups where the specific duplication event has not taken place can be expanded to address the timeline for other features, including in the case of WGD (as carried out by Roux et al. (2017)). However, ohnologs resulting from the vertebrate 2R duplications are likely not the best candidates for this type of analysis as there is a shortage of closely related outgroups in this case, making it difficult to assess features such as evolution rate as the divergence between species is high. Other WGD events such as those in yeast or teleosts would be better suited for this work.

In addition to the extension of the results shown here, there is further work which could be carried out to confirm the nature of the differences between ohnologs and SSDs. Although here we have confirmed that, data availability allowing, other vertebrates show the same patterns as humans, all of these ohnologs have arisen from the same duplication event. An examination of patterns for the ohnologs arising from the teleost specific duplication (3R duplication) could confirm that WGDs in vertebrates follow the same general pattern any time they occur, which is not necessarily a foregone conclusion as work on genes retained following the salmon specific WGD has not shown 3R ohnologs to be any more likely to be retained (Lien et al., 2016).

Outside of vertebrates, the same work could also be repeated in other groups such as yeast or plant species. The results of this analysis could be interesting to compare to human results to determine if there are any major differences caused by the differences between species. We would expect to see much the same patterns, excluding enrichment for functions which do not exist in these other species, such as neural development, as most of the work confirmed here was originally carried out in yeast and plants have shown remarkably similar results, down to enrichment for immune function in SSDs (Qiao et al., 2018) despite large differences in immune systems between humans and plants. Given this broad similarity across such diverse species, any differences that are observed would warrant explanation and could offer further insight into the mechanisms of duplication.

References

Akiyama, M. et al. Autoinflammatory keratinization diseases. *Journal of Allergy and Clinical Immunology*, 140(6):1545–1547, 2017.

Amoutzias, G.D. et al. Posttranslational regulation impacts the fate of duplicated genes. *Proceedings of the National Academy of Sciences*, 107(7):2967–2971, 2010.

Aury, J.M. et al. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature*, 444(7116):171–178, 2006.

Babenko, V.N. Prevalence of intron gain over intron loss in the evolution of paralogous gene families. *Nucleic Acids Research*, 32(12):3724–3733, 2004.

Bailey, J.A. et al. Segmental duplications: organization and impact within the current human genome project assembly. *Genome research*, 11(6):1005–17, jun 2001.

Banerjee, S., Feyertag, F. and Alvarez-Ponce, D. Intrinsic protein disorder reduces small-scale gene duplicability. *DNA Research*, 24(4):435–444, 2017.

Birchler, J.A. and Veitia, R.A. Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences*, 109(37):14746–14753, 2012.

Blomme, T. et al. The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biology*, 7:R43, 2006.

Bos, J. et al. Psoriasis: dysregulation of innate immunity. *British Journal of Dermatology*, 152 (6):1098–1107, 2005.

Brunet, F.G. et al. Gene Loss and Evolutionary Rates Following Whole-Genome Duplication in Teleost Fishes. *Molecular Biology and Evolution*, 23(9):1808–1816, 2006.

Chapman, B.A. et al. Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. *Proceedings of the National Academy of Sciences*, 103(8):2730–2735, 2006.

Chen, L., DeVries, A.L. and Cheng, C.H.C. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proceedings of the National Academy of Sciences*, 94(8):3811–3816, 1997.

Conant, G.C. and Wolfe, K.H. Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Molecular Systems Biology*, 3, 2007.

Davis, J.C. and Petrov, D.A. Preferential Duplication of Conserved Proteins in Eukaryotic Genomes. *PLoS Biology*, 2(3):e55, 2004.

Dehal, P. and Boore, J.L. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLoS Biology*, 3(10):e314, 2005.

Des Marais, D.L. and Rausher, M.D. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature*, 454(7205):762–765, 2008.

Force, A. et al. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545, 1999.

Fritz, C.O., Morris, P.E. and Richler, J.J. Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, 141(1):2–18, 2012.

Gerstein, A.C. et al. Genomic Convergence toward Diploidy in Saccharomyces cerevisiae. *PLoS Genetics*, 2(9):e145, 2006.

GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550(7675): 204–213, 2017.

Gu, X. Evolution of duplicate genes versus genetic robustness against null mutations. *Trends in Genetics*, 19(7):354–356, 2003.

Guan, Y., Dunham, M.J. and Troyanskaya, O.G. Functional Analysis of Gene Duplications in Saccharomyces cerevisiae. *Genetics*, 175(2):933–943, 2007.

Hakes, L. et al. All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biology*, 8(10):R209, 2007.

He, X. and Zhang, J. Gene Complexity and Gene Duplicability. *Current Biology*, 15(11):1016–1021, 2005a.

He, X. and Zhang, J. Rapid Subfunctionalization Accompanied by Prolonged and Substantial Neofunctionalization in Duplicate Gene Evolution. *Genetics*, 169(2):1157–1164, 2005b.

He, X. and Zhang, J. Higher Duplicability of Less Important Genes in Yeast Genomes. *Molecular Biology and Evolution*, 23(1):144–151, 2006.

Hittinger, C.T. and Carroll, S.B. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature*, 449(7163):677–681, 2007.

Jaillon, O. et al. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, 431(7011):946–957, 2004.

Jiang, W.k. et al.

Jiao, Y. et al. Ancestral polyploidy in seed plants and angiosperms. *Nature*, 473(7345):97–100, 2011.

Jun, J. et al. Duplication Mechanism and Disruptions in Flanking Regions Determine the Fate of Mammalian Gene Duplicates. *Journal of Computational Biology*, 16(9):1253–1266, 2009.

Kassahn, K.S. et al. Evolution of gene function and regulatory control after whole-genome duplication: Comparative analyses in vertebrates. *Genome Research*, 19(8):1404–1418, 2009.

Keller, T.E. and Yi, S.V. DNA methylation and evolution of duplicate genes. *Proceedings of the National Academy of Sciences*, 111(16):5932–5937, 2014.

Klüver, N. et al. Divergent expression patterns of Sox9 duplicates in teleosts indicate a lineage specific subfunctionalization. *Development Genes and Evolution*, 215(6):297–305, 2005.

Kryuchkova-Mostacci, N. and Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Briefings in Bioinformatics*, 18(2):bbw008, 2016.

Lan, X. and Pritchard, J.K. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science*, 352(6288):1009–1013, 2016.

Lien, S. et al. The Atlantic salmon genome provides insights into rediploidization. *Nature*, 533 (7602):200–205, 2016.

Liu, Y. et al. Structure and Evolutionary Origin of Ca2+-Dependent Herring Type II Antifreeze Protein. *PLoS ONE*, 2(6):e548, 2007.

Lynch, M. and Conery, J.S. The evolutionary demography of duplicate genes. *Journal of Structural and Functional Genomics*, 3:35–44, 2003.

Macqueen, D.J. and Johnston, I.A. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proceedings of the Royal Society B: Biological Sciences*, 281(1778), 2014.

Makino, T. and McLysaght, A. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proceedings of the National Academy of Sciences*, 107(20): 9270–9274, 2010.

Marques, J.T. and Carthew, R.W. A call to arms: coevolution of animal viruses and host innate immune responses. *Trends in Genetics*, 23(7):359–364, 2007.

Nakatani, Y. et al. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Research*, 17(9):1254–1265, 2007.

Ohno, S. *Evolution by gene duplication.* Springer, New York, 1970.

O'Toole, Á.N., Hurst, L.D. and McLysaght, A. Faster Evolving Primate Genes Are More Likely to Duplicate. *Molecular Biology and Evolution*, 35(1):107–118, 2018.

Oughtred, R. et al. The BioGRID interaction database: 2019 update. *Nucleic Acids Research*, 47(D1):D529–D541, 2019.

Putnam, N.H. et al. The amphioxus genome and the evolution of the chordate karyotype. *Nature*, 453(7198):1064–1071, 2008.

Qian, W. and Zhang, J. Genomic evidence for adaptation by gene duplication. *Genome Research*, 24(8):1356–1362, 2014.

Qiao, X. et al. Different Modes of Gene Duplication Show Divergent Evolutionary Patterns and Contribute Differently to the Expansion of Gene Families Involved in Important Fruit Traits in Pear (Pyrus bretschneideri). *Frontiers in Plant Science*, 9, 2018.

Rice, A.M. and McLysaght, A. Dosage-sensitive genes in evolution and disease. *BMC Biology*, 15(1):78, 2017.

Roux, J., Liu, J. and Robinson-Rechavi, M. Selective Constraints on Coding Sequences of Nervous System Genes Are a Major Determinant of Duplicate Gene Retention in Vertebrates. *Molecular Biology and Evolution*, 34(11):2773–2791, 2017.

Roy, S.W. and Penny, D. On the Incidence of Intron Loss and Gain in Paralogous Gene Families. *Molecular Biology and Evolution*, 24(8):1579–1581, 2007.

Scannell, D.R. et al. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*, 440(7082):341–345, 2006.

Session, A.M. et al. Genome evolution in the allotetraploid frog Xenopus laevis. *Nature*, 538 (7625):336–343, 2016.

Singh, P.P., Arora, J. and Isambert, H. Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes. *PLOS Computational Biology*, 11(7):e1004394, 2015.

Strobl, C. et al. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, 2007.

van Hoek, M.J.A. and Hogeweg, P. Metabolic Adaptation after Whole Genome Duplication. *Molecular Biology and Evolution*, 26(11):2441–2453, 2009.

Wang, T. et al. Identification and characterization of essential genes in the human genome. *Science*, 350(6264):1096–1101, 2015.

Wolfe, K. and Shields, D. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387(June):708–13, 1997.

Woods, S. et al. Duplication and Retention Biases of Essential and Non-Essential Genes Revealed by Systematic Knockdown Analyses. *PLoS Genetics*, 9(5):e1003330, 2013.

Zerbino, D.R. et al. Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761, 2018.

Zhu, Y., Lin, Z. and Nakhleh, L. Evolution After Whole-Genome Duplication: A Network Perspective. *G3: Genes—Genomes—Genetics*, 3(11):2049–2057, 2013.