Dessimoz Lab,
Center for Integrative Genomic &
Department of Ecology and Evolution,
University of Lausanne,
Switzerland

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Detecting whole genome duplication using hierarchical orthologous groups

First Step Project

MSc Molecular Life Science (MLS) - Bioinformatics

By Robin Hofmeister

Directed by Prof. Christophe Dessimoz

Supervised by Clément Train

ABSTRACT

Whole genome duplication events are established as largely responsible for the large genome size, the complexity and the evolution of most eukaryotic organisms. Since a long time discussed for their involvement in the evolution of vertebrates (2R) and fishes (3R), such events are difficult to detect and locate because of the fast dynamics of the genome evolution after it duplicated. Analysis reveal that ancestral genomes contain a lower rate of duplicated genes than extant genomes, suggesting that their longer evolutionary time has resulted in the loss of many genes. This led us to develop different methods to detect whole genome duplication. The first method based on the general duplication rate allows to identify seven genomes that are referenced in literature to have undergone a whole genome duplication: *G. max, P. trichocarpa, G. raimondii, T. aestivum, P. patens, E. tef, P. tetraurelia.* The second method is based on the genome duplication rate according to the duplication rate among its lineage. It allows to identify three referenced ancestral genomes: *Solanum, Musa acuminata* and *Brassica.* Whole genome duplication in ancestors of vertebrates and fishes could not been detected with previous methods. However, a pattern suggesting a whole genome duplication after the divergence of lampreys has been identified. It seems correspond to the second round of whole genome duplication.

1.INTRODUCTION **4**

2. METHODS **6**

2.1 Dataset 6

2.2 Orthology Inference and HOG reconstruction 7

2.3 HOGs exploration: Pyham 7

2.4 Benchmarking 7

2.5 Analysis: ratios 8

2.6 Analysis: WGD detection in extant genomes 9

2.7 Analysis: WGD detection in ancestral genomes 9

3. RESULTS **10**

3.1 Preview 10

3.2 Extant genome 12

3.3 Ancestral genome 12

3.4 Vertebrates and fish-specific WGD hypothesis 12

4. DISCUSSION **13**

4.1 Potential WGD in extant genomes 13

4.2 Potential WGD in ancestral genomes 15

4.3 False negatives 15

4.4 Limitations 16

5. PERSPECTIVES **17**

6. SUPPLEMENTAL DATA **18**

7. ANNEXE **22**

8. REFERENCES **23**

1. INTRODUCTION

During evolution, many eukaryotic organisms have acquired their complexity and large genome size through event called whole genome duplication (WGD) - an event in which the resulting genome is composed of an additional copy of its entire genome. Bases on the relative genome size, Susmo et al. proposed the first hypothesis of whole genome duplication. It assumes that vertebrates evolution has been shaped by two round of whole genome duplication. Occurring early in the ancestor of vertebrates, these WGD lead to an actual vertebrates genome representing paleopolyploidy. The importance of gene redundancy is underlined, suggesting duplication as a more important mechanism in evolution than natural selection (S. Ohno, 1970-1999). The main arguments was that after gene duplication, one copy remains the same while the additional copy evolve to take a new function - called later neo-functionalization (Dehal et al., 2005). More recently, evidences demonstrated that most fish might have more genes than human. This leads to a second hypothesis, suggesting a third round (3R) of whole genome duplication occurring only in the Teleostei clade (Meyer et al., 2005). These hypotheses are supported and strongly discussed during the next decades, but still remains unsolved. Such events are difficult to defined because several mechanisms can lead to an increase of number of genes in the genome, such as tandem duplication, segmental duplication or transposable elements. Recent improvement of sequencing techniques might help to discriminate over these different process.

Whole genome duplication events are established as largely responsible for evolution and speciation of species. It has been demonstrated - with a reconstruction of the fish ancestral genome - that interchromosomal rearrangements occurs only rapidly after the fish duplication, suggesting an increase of genomic changes just after event of WGD (Kasahara et al., 2007). Moreover, as already proposed by Ohno et al., duplicated genes can have different fate (S. Ohno, 1970-1999). After a whole genome duplication, pair of genes has a short lifetime before the additional copy diverge functionally. It can result in (i) a sub-functionalization, where genes shared the ancestral function, (ii) a neo-functionalization, where the additional copy evolve to a

new function, (iii) a non-functionalization, where the additional copy is silenced to become a pseudogene, (iv) in a super-functionalization, where both genes keep original function, or (v) in the loss of the additional copy of the duplicated gene. Except for genes lost, relationship between these pairs of duplicated genes can be traced back by a concept called homology - a shared ancestry. Homology can be subdivided in two groups. Paralogous genes are defined as genes related through duplication from the same gene in the common ancestor, while orthologous genes are related through speciation (Fitch, 1970; Gabaldón et al., 2013). Orthologous genes from two species allow to trace back to an ancestral gene present in a common ancestor of these two species. Thus, orthology provides the most effective way to quantify similarities and differences between the genome composition of different species. The concept of orthology genes become more complicated if a gene has more than one ortholog in a given genome - so called co-orthologs.

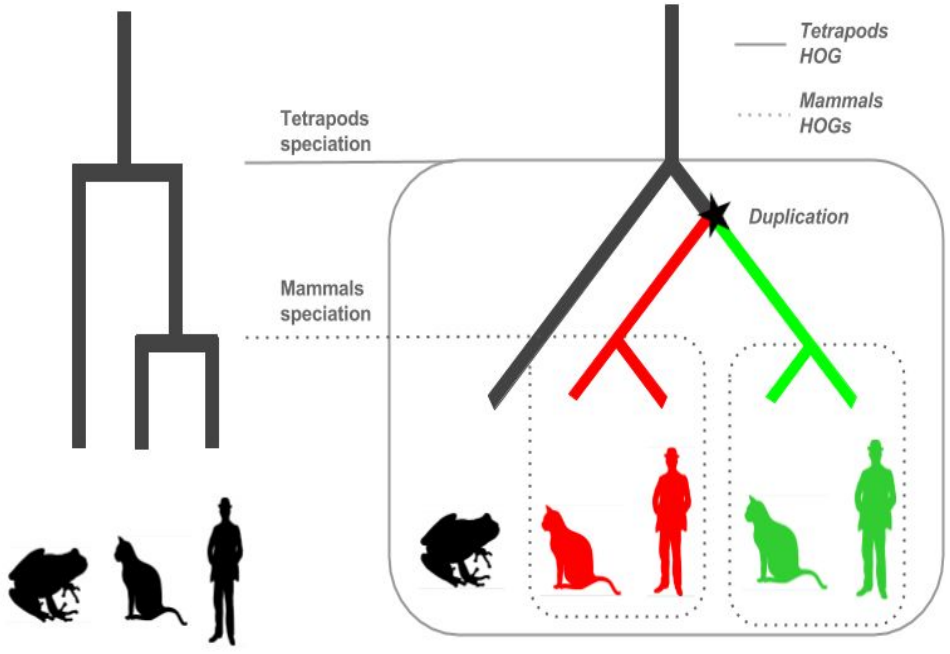


Figure 1. *Hierarchical Orthologous Groups (HOGs).* Species tree (**left**) and its associated gene tree (**right**) representing the evolutionary relationship of five genes resulting from the same common ancestor. Three HOGs can be identify here: one HOG at the tetrapods level (**dotted line**) and two smaller HOGs at the mammalian level (**full line**), due to a duplication event (**star**). Orthologs genes are indicated with the same color (red or blue) while paralogs genes are indicated with different colors (red versus blue). The duplication event after the first speciation event leads to in-paralogs genes into the Tetrapods HOG.

Therefore, Orthologous group is defined as a set of homologous genes that come from the same ancestral gene after speciation event. It includes all genes that shared orthologous relationship, such as orthologs and co-orthologs, but also paralogs genes that have evolved through duplication after the speciation event, then called in-paralogs (Gabaldón et al., 2013; Boeckmann et al., 2011).

A decade of sequencing technologies improvement has allowed to increase genomic data available. It is then possible to compare more than two gene at a time, at more than one taxonomic range. Orthologous groups are not scaling for these analysis because they don't provide information about evolutionary events - such as genes duplicated or lost. To solve this, the concept of Hierarchical Orthologous Groups (HOGs) is defined as "sets of genes that have descended from a single common ancestor within a taxonomic range of interest" (Altenhoff et al., 2013 ; Boeckmann et al., 2011; Sonnhammer et al., 2014). It allows to reconstruct either the whole gene family history by taking one HOG for all levels, or an ancestral genome by taking all HOGs at a given level. It can also be used to compare content between genomes.

In the following analysis, we will use the evolutionary relationship described by Hierarchical Orthologous Groups to detect whole genome duplication among the Eukaryotic clade.

2. METHODS

2.1 Dataset

The Orthologous MAtrix (OMA) project is a algorithm used to infer large scale orthology among complete genomes, based on the protein sequences. The OMA browser is the associated web database for orthology predictions among more than 2000 complete genomes. We retrieved from the OMA browser all the 327 full genome sequences from the eukaryotic clade with their related all-against-all alignments (Schneider et al., 2007; www.omabrowser.org).

2.2 Orthology Inference and HOG reconstruction

The OMA project provides a standalone version of the OMA algorithm that can be run on custom datasets (www.omabrowser.org/standalone). The OMA algorithm infer pairwise orthology relations and HOGs in 3 steps: (i) All-Against-All alignments of all protein sequence pairs from all the genomes and upgrade as homologs all aligned pair of sequences with sufficient alignment score . (ii) infer orthologs from closest pairs of sequences between genomes pairs, (iii) reconstruct HOGs from pairwise orthologous relations using GETHOGs algorithm ("Graph-based Efficient Technique for Hierarchical Orthologous Groups") The algorithm outputs an orthoXML (standard file format for HOGs) file containing all HOGs (Altenhoff et al., 2013; Train et al., 2017).

2.3 HOGs exploration: Pyham

To manipulate and analyse evolutionary informations contained in HOGs, we use Pyham (Python HOG Analysis Method), a python library to facilitate exploration and visualisation of orthoXML. We use the built-in functionalities of Pyham to compare ancestral genomes based on their ancestral genes evolution (induced by HOGs). This allows to identify how ancestral genes have evolved in between two levels by a classification of genes into 4 categories: (i) 'duplicated' if the ancestral gene duplicated one or more time, (ii) 'lost' if the ancestral gene have been lost in between the two levels, (iii) 'gained' if the gene can only be found in the younger genome compared and (iv) 'identical' if the ancestral gene was not affected by one of the previously mentioned evolutionary events.

2.4 Benchmarking

In order to assess the accuracy of our algorithm, we use case of whole genome duplications found in literature as reference. The following whole genome duplication are referenced:

(Panchy et al., 2016; Cheng et al., 2016; Cannarozzi et al., 2014; Aury et al., 2006; Ma et al., 2009; Scannell et al., 2007; Plant Genome Duplication Database, http://chibba.agtec.uga.edu/duplication/)

- Used as reference: *Glycine max, Populus trichocarpa, Gossypium raimondii, Triticum aestivum, Zea mays, Physcomitrella patens, Eragrostis tef, Paramecium tetraurelia, Rhizopus oryzae, Solanum, Brassica, Brassicaceae, Musa acuminata, Poaceae* (a detailed table is available in *Annexe 1*).

- *2R hypothesis:* comparison of sequences of sea squirt, lamprey, elephant shark, tunicate, lancelet and bony fish allows to precise the location of the potential two rounds of whole genome duplication (Putnam et al., 2008; Holland et al., 1994). The first round is located after the divergence of lancelets (*B. floridae)* and tunicates, and before the divergence of lampreys (*P. marinus)*. The second round is located after the divergence of Lampreys and before the divergence of Chondrichthyes. In Oma browser, corresponding levels are *Vertebrata* and *Euteleostomi*.
- *3R hypothesis:* it has been established that the fish-specific whole genome duplication has shaped the evolution of Teleostei (Glasauer et al., 2014; Sato et al., 2010). Thus its associated location in OMA browser is after the divergence of *L. oculatus* and the *Clupeocephala* clade -belongs to Teleostei.

According to the early divergence and the large evolutionary time of these clades, we will try to detect a pattern resulting from these rounds of whole genome duplication through a graphical representation of the rate of duplicated genes among all lineage.

2.5 Analysis: ratios

In order to detect whole genome duplication, we try to detect abnormal pattern in number of ancestral genes. To proceed, we design several ratio based on behavior genes in between taxonomic ranges. The first ratio called *R.naive* is calculated based on the total number of genes in the child genome compared to the total number of genes in its parent genome. It is a rough estimate of the genome wide gene behavior. The second ratio called *R.shared* refine the first ratio by removing genes that have no shared relation between the child and its parent genome (i.e. genes lost and gained). It is a rough estimate of the relative expansion of genes between parent and child. The third ratio called *R.duplicated* is calculated based on the number of genes descending

from the parents genome through duplication event. It is an estimator of the duplication rate between the parent genome and its child genome.

2.6 Analysis: WGD detection in extant genomes

Ratios previously described are calculated for all genomes retrieved from the Oma browser. We try to detect ratios significantly different between genomes, that could explain a whole genome duplication. A threshold is set for each ratio in order to get the more true positive and the less false positive results, using referenced cases of WGD. It has been found an optimal combination for these ratio set to (*R.naive, R.shared, R.duplicated) = (1, 1.1, 0.5)*. All genomes are sorted according to this combinaison. Those for which the three ratios are above thresholds are selected as good candidates for a whole genome duplication. The *R.duplicated* is the ratio the more useful to sort between genomes. It has been found that if *R.duplicated* is decreased to include all WGD known (the one with the weaker ratio being *Z. mays* : 0.367), it get eleven more false positives. If the ratio is increased to exclude all false positives (the one with the highest ratio being *S. moellendorffii :* 0.799*),* it get only two results - which are true positives - but six genomes known to have undergone a WGD are excluded. Moreover, species with only one or two ancestors are also excluded from the analysis because of the following. Ancestral genomes close to root are suggested to have undergone a long evolution time and generally more loss than ‘recent’ ancestral genomes. As the tree is built from the bottom with the extant genomes sequenced, ancestral genomes lose in resolution and in fidelity when we go up in the tree. If an extant genome is compared directly to the ancestral *Eukaryota* genome, it might introduce a bias in the analysis.

2.7 Analysis: WGD detection in ancestral genomes

To detect WGD in internales nodes, a research is made in order to find a pattern in taxonomic lineage where a WGD in known. It is found that when a WGD happened in an ancestral genome in position *i* in the lineage, his proportion of duplicated genes is largely higher than in the ancestral genome in position *(i-1)* or *(i+1)* (*supplemental data 1*). A second algorithm is created with these differences implemented. The sorting is

made in order to have a least twice more duplication in ancestral genome *i.* The second ratio -proportion of duplicated gene- is also add to the algorithm. According to the general decrease of proportion of duplicated genes in internal nodes (*Supplemental data 2.b*), the sorting for the second ratio can be made with a value weaker than for the leaves. Parameters are calibrated in order to obtain the more true positives and the least false positives.

3. RESULTS

3.1 Preview

A first visualization of informations contain in HOGs and extract by Pyham indicates that leaves and internal nodes (respectively extant genomes and ancestral genomes) have generally the same amount of genes (*Supplemental data 2.a*). This leads us firstly to think that the ancestral genomes have been reconstructed with a certain fidelity and resolution. On the other hand, the percentage of genes resulting from a duplication is clearly higher in extant genomes that in ancestral genomes (*Supplemental data 2.b*). Knowing that it does exists WGD in nodes, this information should be considered carefully because this difference in duplicated genes could introduce a bias for the next investigations. Moreover, some species have a very small number of genes compared to their close related species. For example *Eimeria tenella* contains only 214 genes, while its close related species, *E. maxima* and *E. acervulina,* contain 6057 and 6862 genes respectively. This might be due to sequencing errors for *E. tenella*. Such species might lead to an incomplete reconstruction of the ancestral genome, in which the number of genes might be under-estimated because of this.

In order to detect WGD events, three ratios were used (see methods). Distributions of these ratios over species are shown on *figure 2.a*. It can be noticed that some species with an initial amount of genes in the child genome largely higher than in the parent genome (*R.naive*>1), are excluded by the second ratio (*R.shared*<1). It means that such species have gained a lot of genes, which has biased the first ratio (*Reticulomyxa filosa* : 15% of loss, 87% of gain). On the other hand, some species

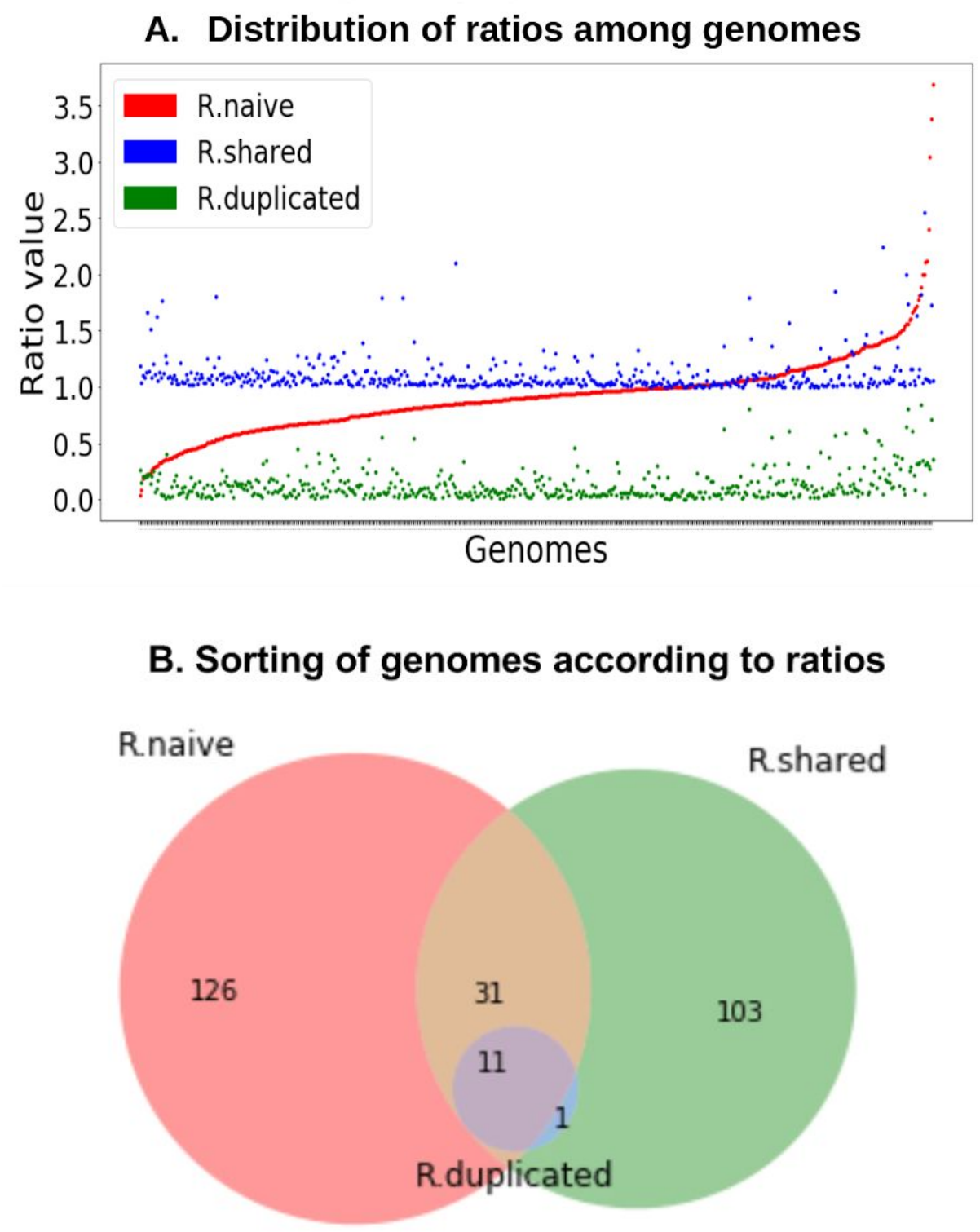


Figure 2: ***Genomes distribution according to ratios. A.*** 537 genomes are aligned on the x axis with their three corresponding ratio values (see methods) on the y axis. Genomes are sorted according to *R.naive.* **B.** 537 genomes are classified in five groups according to the three ratios set in the algorithm. **red**: R.naive>1, R.shared<1, R.duplicated<0.5 ; **green**: R.naive<1, R.shared>1, R.duplicated<0.5 ; **brown**: R.naive>1, R.shared>1, R.duplicated<0.5 ; **blue**: R.naive<1, R.shared<1, R.duplicated>0.5 ; **purple**: R.naive>1, R.shared>1, R.duplicated>0.5. Purple and blue area are the selected genomes.

with an important amount of genes lost are excluded by the first ratio (*R.naive*<1), while the second ratio includes them (*Schistosoma mansoni* : 509% of loss, 58% of gain). Sames results can be visualized with a Venn diagram, which represents three groups corresponding to the three differents ratio. Genomes with massive gene gained are excluded from a WGD hypothesis by the second ratio, while genome with massive gene loss are included (*figure 2.b*). We conclude by this first overview that the first ratio - *R.naive* - is not very helpful because of the large quantities of loss and gain. However, some extant genomes or ancestral genomes correspond to the characteristics of a whole genome duplication and still need to be investigated.

3.2 Extant genome

These three ratios are implemented and calibrated in an algorithm (see Methods), then tested on eukaryotic genomes. To assess the power of those ratios , known whole genome duplications are compared to the results. From these 14 referenced WGD, 7 extant genomes are detected: *G. max, P. trichocarpa, G. raimondii, T. aestivum, P. patens, E. tef, P. tetraurelia.* Two extant genomes and four ancestral genomes are not detected - false negatives. Four extant genomes are detected, but not referenced as known WGD, and so require further investigation - *B. napus, S. moellendorffii, C. cinerea*, *M. esculenta* (*supplemental data 3*).

3.3 Ancestral genome

The second method developed allows us to detect three more whole genome duplications, that occur in the following ancestral genomes: *Musa acuminata, Solanum, Brassica* and *Entamoeba* (*supplemental data 4*). The first three are referenced to have undergone a WGD and are therefore true positives results. However, investigations on *Entamoeba* genome are required.

3.4 Vertebrates and fish-specific WGD hypothesis

Hypothesis a whole genome duplication in vertebrates and fish could not be detected by the previous algorithms, because of the small proportion of duplicated genes.

However, a pattern can be noticed on the *figure 3*, indicating a relatively large increase of duplicated genes before the divergence of Tetrapods and Fishes. This result excludes Lampreys (*P. marinus*) from the duplication event and thus would be the second round of duplication rather than the first (Holland et al., 1994; Ward et al., 1981; Putnam et al., 2008).

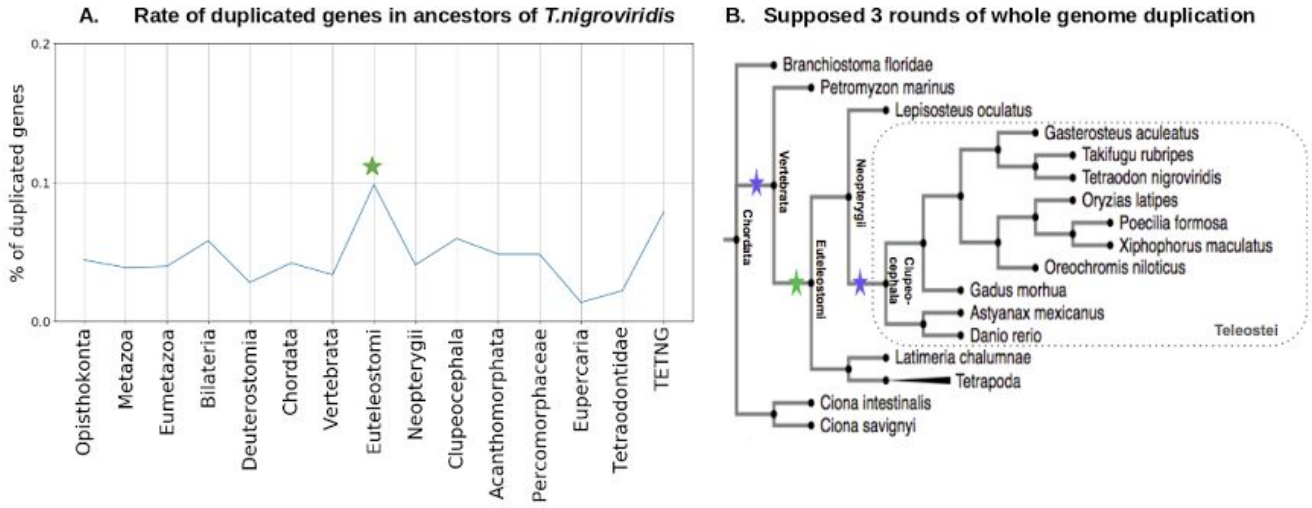


Figure 3: ***Supposed whole genome duplication in early vertebrates.*** **A.** Distribution of the rate of duplicated genes among ancestors of *T. nigroviridis* and **B.** its associated phylogeny in OMA browser. **Blue stars** indicate the supposed first and third round of whole genome duplication. **Green stars** indicate the supposed and detected second round of whole genome duplication. The lacks of outgroup for Teleostei (**dotted line**) makes the fish-specific whole genome duplication difficult to detect.

4. DISCUSSION

4.1 Potential WGD in extant genomes

Investigations on the false positives reveal some characteristics of their genomes that could explain their detection by the algorithms. It is found that the nuclear genome of *Brassica napus* contains a lot of sequences coming from *B. oleracea.* Its chloroplast phylogeny indicates also several sequences coming from *B. oleracea.* It is first surprising because OMA browser defines them as sister species diverging from the same ancestor - *Brassica*. Actually, *B. napus* is formed by multiples fusion events between ancestor of *B. rapa* and *B. oleracea*. Moreover, *B. rapa* genome is known to have undergone a WGD (Palmer et al., 1983; Parkin et al., 2005; Panchy et al., 2005). As the ancestral genome *Brassica* has been reconstructed from only *B. napus* and *B.*

oleracea genomes, it is clear that the proportion of duplicated and identical genes are overestimated when *B. napus* genome is compared to *Brassica* genome. We cannot conclude that a whole genome duplication occurs in *B. napus.* We suggest that the algorithm detect this genome because of the bias created by the lacks of species for the *Brassica* ancestral genome reconstruction.

Investigation on *Selaginella moellendorffii* reveals no evidences for a WGD event. Taking into account its time of divergence from the *Tracheophyta* lineage (~400 million years ago), the most plausible hypothesis is that a lot of small duplication events occurs independently in its genome, increasing the proportion of duplicated genes. Moreover, its genome is composed of a lot of duplicated segments, including 37.5% of transposable elements (Banks et al., 2011; Weng et al., 2005) . We can therefore conclude that the early divergence from the *Tracheophyta* lineage and the independant accumulation of duplicated segments lead to a large proportion of duplicated genes in its genome, without a WGD event. To confirm this hypothesis, it can be noticed some ancestral genes that have duplicated up to 105 times in *S. moellendorffii,* suggesting several independent duplications.

The *Coprinopsis cinerea* genome reveals several evidence of small independant duplication, that can be responsible for the increase of duplicated genes. The presence of seventeen non-allelic laccase gene is identified in its genome, probably reflecting a recent event of gene duplication. Moreover, three serine proteinases are identified as homologous to two serine proteinases in *Agaricus bisporus* - a close related specie (Kilaru et al., 2005; Heneghan et al., 2008). This suggest either a local duplication or a common ancestor for these species. Finally, a duplication of pheromone receptors is known to occur in the ancestor of *Schizophyllum commune* and *Coprinopsis cinerea* (Devier et al., 2008). Phylogenetic relationship between *A. bisporus, C. cinerea* and *S. commune* are solve as a polytomy on OMAbrowser (www.omabrowser.org). It does not allow to replace the pheromone receptor duplication nor the proteinases duplication. However, it confirm a common ancestor for *A. bisporus* and *C. cinerea.* In this case, we can conclude that we lack information and no WGD can be defined.

No clear evidences of whole genome duplication are found concerning *Manihot esculenta.* However, its genome indicates large duplicated region regarding to castor

bean genome, and several duplicated loci (Prochnik et al., 2011; Fregen et al., 1996). It can be a good candidate for future investigation of whole genome duplication that are made possible thanks to new sequencing techniques.

4.2 Potential WGD in ancestral genomes

It has been found that *Entamoeba* species have a heterogenous amount of DNA. Actually, they have the possibility to contain multiples nuclei, and each of them can contains a different amount of DNA (Mukherjee et al., 2008). It has been established that *E. histolytica* can achieve several endo-reduplicative cycles without mitosis, and that the number of endo-reduplicative cycles undergoes many variations (Lohia et al., 2007). Moreover, *E. histolytica* and *E. dispar* genomes contain several segmental duplication -particularly in rRNA genes- and tandem duplication - particularly in tRNA genes (Lorenzi et al., 2010). Finally, *Entamoeba* phylogeny is solved in OMA browser as a polytomy (www.omabrowser.org), which might introduce a bias in the reconstruction of the *Entamoeba* ancestral genome. Knowing all these characteristics of *Entamoeba* species, it is difficult to conclude to a whole genome duplication. This ancestral genome contains a large proportion of duplicated segments and has the possibility to have multiples nuclei, which favor its detection by ratios.

4.3 False negatives

Some genomes known to have undergone a whole genome duplication are still not detected by any created algorithms. *Z. mays* and *R. oryzae* have a relatively high percentage of duplication in the genes coming from the their ancestral genomes (36% and 39%, respectively), but if we decrease the second ratio to include them in the results, we get a too large proportion of false positive. Both species have the largest proportions of genes classified as ‘identical’ (40% and 34%, respectively), and also a large proportion of genes classified as ‘lost’ (*supplemental data 5*). Regarding to Pyham classification of genes (see Methods) and to the fate of duplicated genes, a pairs of duplicated genes from which one gene is lost is classified as ‘identical’. If both genes are lost, the corresponding ancestral HOG is classified as ‘lost’. Therefore, *Z. mays* and *R. oryzae* undergo an underestimation of the number of duplicated genes,

which is responsible of their exclusion by ratios. Genes classification induce a bias in the calculated number of duplicated genes. On the other hand, ancestral genomes *Brassicaceae* and *Poaceae* have a small proportion of duplicated genes. However, they have the largest proportion of genes classified as 'identical' among the referenced ancestral whole genome duplication (*supplemental data 6*). As previous for the species, the number of duplicated genes is biased. It can also be noticed a lack of plant species, which does not allow a complete reconstruction of ancestral genomes.

4.4 Limitations

The main cause that makes whole genome duplication difficult to detect is the dynamic of the genome after its duplication. Different fate of duplicated genes lead to a large amount of genes lost, that result in an underestimation of the proportion of duplicated genes. It can be considered that loss are bigger when the evolution time between two nodes is long. Thus, there is generally more loss in ancestral genomes close to *Eukaryota* genome -the root- and a whole genome duplication is more difficult to detect. Because of this fact, species with only one or two ancestors are excluded from a WGD hypothesis since the beginning. It is suggested that if they are detected, it is because of their rapid divergence and their long evolutionary time, which allowed to accumulate more independent duplication. Four extant genomes are excluded because of this : *Emiliania huxleyi, Naegleria gruberi, Guillardia theta and Perkinsus marinus*. Confirming the previous hypothesis, all these species contains genes that have duplicated several times: *P. marinus* up to 46 times, *N. gruberi* up to 657 times, *E. huxleyi* up to 1122 times and *G. theta* up to 1164 times.

Ancestral genomes are reassembled according to their descending genomes availables. The resolution of an ancestral genomes is decrease if some of its descending genomes are missing (potential HOGs conserved only in the missing specie will not be part of the ancestral genome). More extant genomes would be required to precise the location of the two rounds of WGD in vertebrates and fishes. Add Sharks to the phylogenetic tree would create a new nodes, and thus a new ancestral genome between *Vertebrata* and *Euteleostomi*.

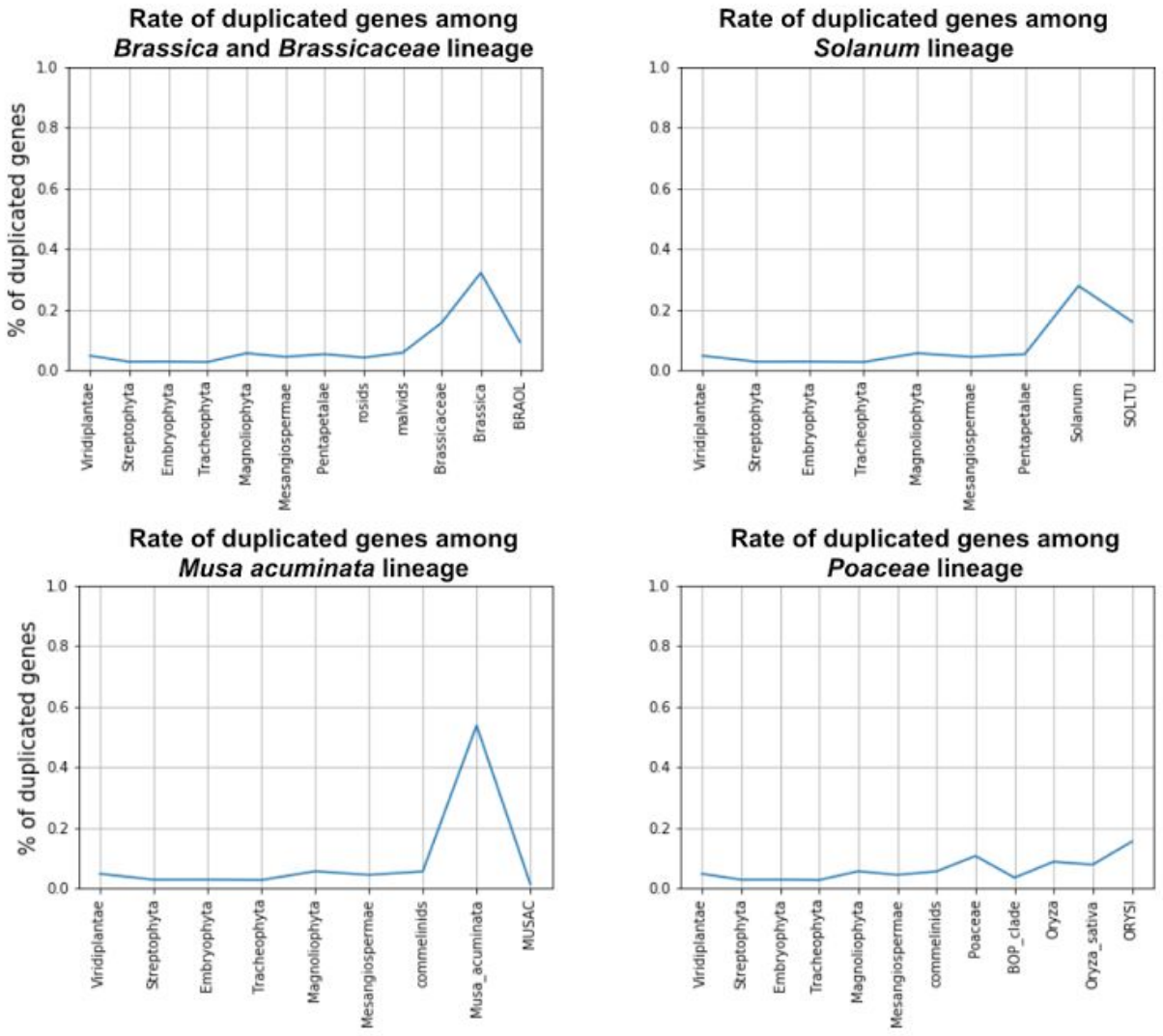
5. PERSPECTIVES

Methods used in these investigations to detect whole genome duplication have a certain power and several limitations. Extant genomes that have diverged late provides the best support to test our methods, while extant genomes that have diverged early have had time for accumulate a lot of independant duplication and are generally detected as false positives. Due to the large evolutionary time and to the dynamics of the genome after its duplication, ancestral whole genome duplication are more difficult to detect with these methods. To improve this analysis, it is possible to build the phylogenetic genes tree from each HOG and look for massive duplications in taxa suspected of having undergone a whole genome duplication. Moreover it is possible to use the principle of synteny to support predictions. Ancestral synteny can be reconstruct from the extant genomes and thus help us to confirm predictions.

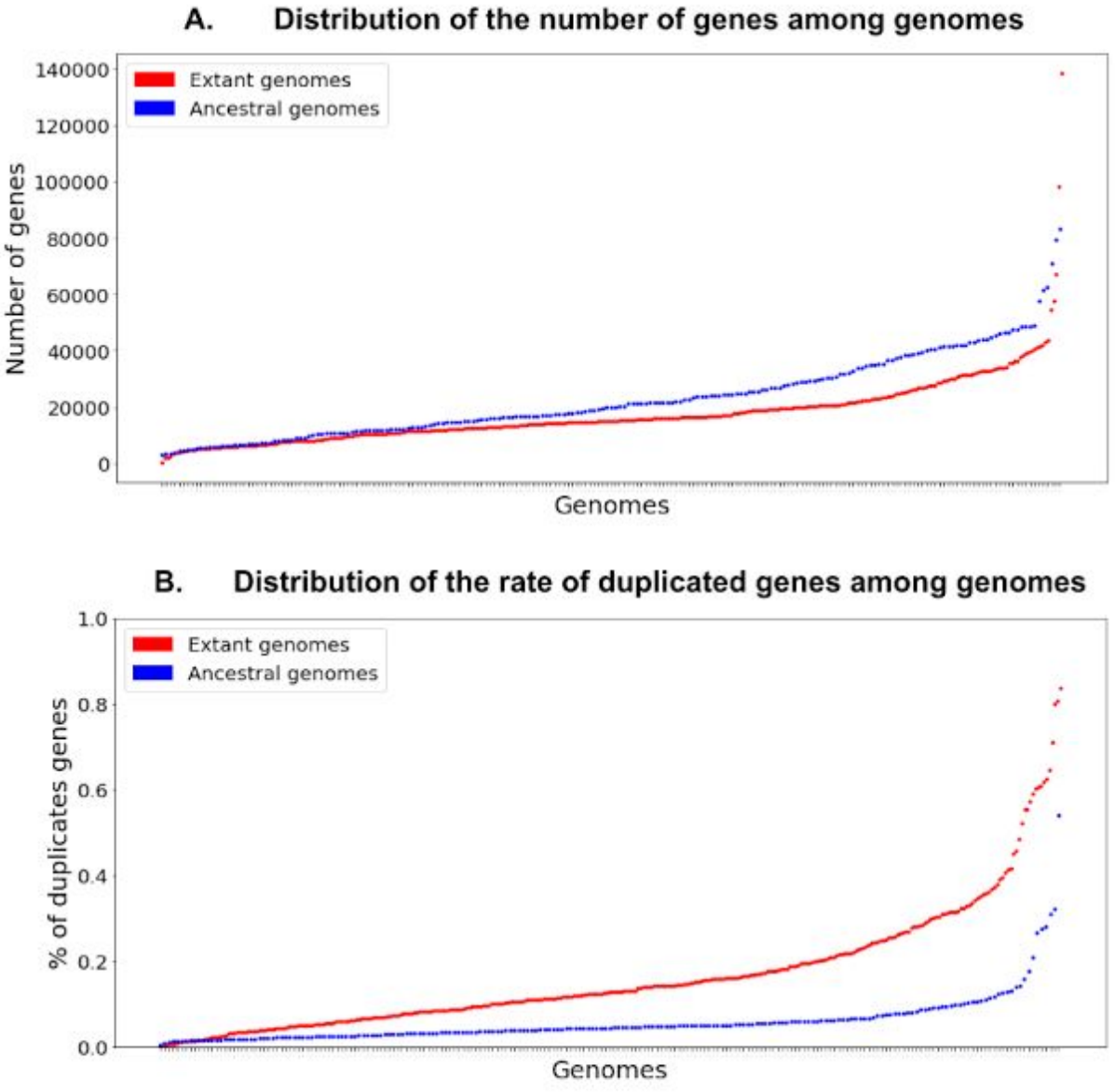
6. SUPPLEMENTAL DATA

Code hosted at : www.github.com/RobinHofmeister/FirstStepProject_MSc-MLS-A2017

Supplemental data 1. *Distribution of the rate of duplicated genes among lineage of ancestral genomes referenced to have undergone a whole genome duplication*.



Supplemental data 2. *Genomes characteristics*. 327 extant genomes (**red points**) and 210 ancestral genomes (**blue points**) are aligned on the x axis with their associated **a)** number of genes and **b)** proportion of duplicated genes on the y axis.



Supplemental data 3. *Table of the extant genomes detected with the first algorithm and their associated ratios*.

		R.naive	*R.shared*	*R.duplicated*
referenced	*G. max*	1.56	1.73	0.8
	P. trichocarpa	1.37	1.38	0.62
	G. raimondii	1.89	1.82	0.83
	T. aestivum	1.66	1.15	0.6
	P. patens	1.14	1.57	0.6
	E. tef	1.03	1.37	0.62
	P. tetraurelia	3.38	1.73	0.71
	C. cinerea	1.37	1.47	0.6
	M. esculenta	1.08	1.36	0.55
	S. moellendorfii	1.06	1.8	0.8
	B. napus	1.38	1.17	0.52

Supplemental data 4. *Table of the ancestral genomes detected with the second algorithm and their associated ratios*.

		R.naive	*R.shared*	*R.duplicated*
referenced	*Brassica*	1.72	1.63	0.32
	Musa acuminata	0.8	1.4	0.53
	Solanum	0.75	1.39	0.28
	Entamoeba	1.07	1.11	0.31

Supplemental data 5. *Characteristics of the extant genomes referenced in literature to have undergone a whole genome duplication*.

	% total genes duplicated	% total genes identical	% ancestral genes lost
G. max	0.68	0.16	0.25
P. trichocarpa	0.5	0.3	0.2
G. raimondii	0.73	0.14	0.09
T. aestivum	0.4	0.25	0.06
P. patens	0.27	0.17	0.68
E. tef	0.48	0.29	0.42
P. tetraurelia	0.27	0.11	0.26
Z. mays	0.4	0.25	0.06
R. oryzae	0.22	0.34	0.29

Supplemental data 6. *Characteristics of the ancestral genomes referenced in literature to have undergone a whole genome duplication*.

	% total genes duplicated	% total genes identical	% ancestral genes lost
Brassica	0.28	0.49	0.07
Musa acuminat	0.4	0.34	0.57
Solanum	0.25	0.58	0.51
Poaceae	0.09	0.7	0.14
Brassicaceae	0.15	0.63	0.3

7. ANNEXE

Annexe 1: *table with referenced whole genome duplication and their associated match in OMA browser phylogeny.*

Referenced WGD	Match in OMA browser	Note
ancestor of *S. lycopersicum* and *S. tuberosum*	*Solanum*	
M. domestica	no	missing
G. max	yes	
ancestor of *G. max, M. truncatella* and *P. vulgaris*	no	solve as polytomy, *P.vulgaris* missing
ancestor of *P. trichocarpa* and *S. purpurea*	*P. trichocarpa*	*S. purpurea* missing
G. raimondii	yes	
T. aestivum	yes	
Brassica	yes	
B. rapa	no	missing
ancestor of *B. rapa* and *A. thaliana*	*Brassicaceae*	
E. grandis	no	missing
A.coerulea	no	missing
Z. mays	yes	
P. virgatum	no	missing
S. polyrhiza	no	missing
ancestor of *Z. mays* and *O. sativa*	*Poaceae*	
P. patens	yes	
Musa acuminata	yes	
E. tef	yes	
P.. tetraurelia	yes	
R. oryzae	yes	
Yeast		phylogeny solve as polytomy with *K. lactis*

8. REFERENCES

Altenhoff, A. M. et al. (2013) ‘Inferring Hierarchical Orthologous Groups from Orthologous Gene Pairs’, PLoS ONE, 8(1). doi: 10.1371/journal.pone.0053786.

Aury, J. M. et al. (2006) ‘Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia’, Nature, 444(7116), pp. 171–178. doi: 10.1038/nature05230.

Bank, J, Albert, V. a, Aono, N. and Aoyama, T. (2011) ‘Content Associated With the Evolution of Vascular Plants’, Science, 332(6032), pp. 960–963. doi: 10.1126/science.1203810.The.

Boeckmann, B. et al. (2011) ‘Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees’, Briefings in Bioinformatics, 12(5), pp. 423–435. doi: 10.1093/bib/bbr034.

Cannarozzi, G. et al. (2014) ‘Genome and transcriptome sequencing identifies breeding targets in the orphan crop tef (Eragrostis tef)’, BMC Genomics, 15(1), p. 581. doi: 10.1186/1471-2164-15-581.

Cheng, F. et al. (2015) ‘Genome triplication drove the diversification of Brassica plants’, The Brassica rapa Genome, (February), pp. 115–120. doi: 10.1007/978-3-662-47901-8_10.

Dehal, P. and Boore, J. L. (2005) ‘Two rounds of whole genome duplication in the ancestral vertebrate’, PLoS Biology, 3(10). doi: 10.1371/journal.pbio.0030314.

Devier, B. et al. (2009) ‘Ancient trans-specific polymorphism at pheromone receptor genes in basidiomycetes’, Genetics, 181(1), pp. 209–223. doi: 10.1534/genetics.108.093708.

Fregene, M. (1997) ‘A molecular genetic map of cassava’, pp. 431–441.

Gabaldón, T. and Koonin, E. V. (2013) ‘Functional and evolutionary implications of gene orthology’, Nature Reviews Genetics. Nature Publishing Group, 14(5), pp. 360–366. doi: 10.1038/nrg3456.

Glasauer, S. M. K. and Neuhauss, S. C. F. (2014) ‘Whole-genome duplication in teleost fishes and its evolutionary consequences’, Molecular Genetics and Genomics, 289(6), pp. 1045–1060. doi: 10.1007/s00438-014-0889-2.

Heneghan, M. N. et al. (2009) ‘Characterization of serine proteinase expression in Agaricus bisporus and Coprinopsis cinerea by using green fluorescent protein and the A. bisporus SPR1 promoter’, Applied and Environmental Microbiology, 75(3), pp. 792–801. doi: 10.1128/AEM.01897-08.

Kasahara, M. et al. (2007) ‘The medaka draft genome and insights into vertebrate genome evolution’, Nature, 447(7145), pp. 714–719. doi: 10.1038/nature05846.

Kilaru, S., Hoegger, P. J. and Kües, U. (2006) ‘The laccase multi-gene family in Coprinopsis cinerea has seventeen different members that divide into two distinct subfamilies’, Current Genetics, 50(1), pp. 45–60. doi: 10.1007/s00294-006-0074-1.

Lohia, A. et al. (2007) ‘Genome re-duplication and irregular segregation occur during the cell cycle of Entamoeba histolytica’, Bioscience Reports, 27(6), pp. 373–384. doi: 10.1007/s10540-007-9058-8.

Lorenzi, H. A. et al. (2010) ‘New assembly, reannotation and analysis of the Entamoeba histolytica genome reveal new genomic features and protein content information’, PLoS Neglected Tropical Diseases, 4(6). doi: 10.1371/journal.pntd.0000716.

Lynch, V. J. and Wagner, G. P. (2009) ‘Multiple chromosomal rearrangements structured the ancestral vertebrate Hox-bearing protochromosomes’, PLoS Genetics, 5(1). doi: 10.1371/journal.pgen.1000349.

Ma, L. J. et al. (2009) ‘Genomic analysis of the basal lineage fungus Rhizopus oryzae reveals a whole-genome duplication’, PLoS Genetics, 5(7). doi: 10.1371/journal.pgen.1000549.

Meyer, A. and Van De Peer, Y. (2005) ‘From 2R to 3R: Evidence for a fish-specific genome duplication (FSGD)’, BioEssays, 27(9), pp. 937–945. doi: 10.1002/bies.20293.

Mukherjee, C., Clark, C. G. and Lohia, A. (2008) ‘Entamoeba shows reversible variation in ploidy under different growth conditions and between life cycle phases’, PLoS Neglected Tropical Diseases, 2(8), pp. 1–9. doi: 10.1371/journal.pntd.0000281.

Ohno, S. (1999) ‘Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999.’, Seminars in cell & developmental biology, 10(5), pp. 517–522. doi: 10.1006/scdb.1999.0332.

Palmer, J. D. et al. (1983) ‘Chloroplast DNA evolution and the origin of amphidiploid Brassica species.’, TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik, 65(3), pp. 181–9. doi: 10.1007/BF00308062.

Panchy, N., Lehti-Shiu, M. D. and Shiu, S.-H. (2016) ‘Evolution of gene duplication in plants’, Plant Physiology, 171(August), p. pp.00523.2016. doi: 10.1104/pp.16.00523.

Parkin, I. A. P. (2005) ‘Segmental Structure of the Brassica napus Genome Based on Comparative Analysis With Arabidopsis thaliana’, Genetics, 171(2), pp. 765–781. doi: 10.1534/genetics.105.042093.

Prochnik, S. et al. (2012) ‘The Cassava Genome: Current Progress, Future Directions’, Tropical Plant Biology, 5(1), pp. 88–94. doi: 10.1007/s12042-011-9088-z.

Putnam, N. H. et al. (2008) ‘The amphioxus genome and the evolution of the chordate karyotype’, Nature, 453(7198), pp. 1064–1071. doi: 10.1038/nature06967.

Sato, Y. and Nishida, M. (2010) ‘Teleost fish with specific genome duplication as unique models of vertebrate evolution’, Environmental Biology of Fishes, 88(2), pp. 169–188. doi: 10.1007/s10641-010-9628-7.

Scannell, D. R. et al. (2007) ‘Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication.’, Proceedings of the National Academy of Sciences of the United States of America, 104(20), pp. 8397–8402. doi: 10.1073/pnas.0608218104.

Schneider, A., Dessimoz, C. and Gonnet, G. H. (2007) ‘OMA Browser - Exploring orthologous relations across 352 complete genomes’, Bioinformatics, 23(16), pp. 2180–2182. doi: 10.1093/bioinformatics/btm295.

Sonnhammer, E. L. L. et al. (2014) ‘Big data and other challenges in the quest for orthologs’, Bioinformatics, 30(21), pp. 2993–2998. doi: 10.1093/bioinformatics/btu492.

Train, C. M. et al. (2017) ‘Orthologous Matrix (OMA) algorithm 2.0: More robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference’, Bioinformatics, 33(14), pp. i75–i82. doi: 10.1093/bioinformatics/btx229.

Weng, J.-K., Tanurdzic, M. and Chapple, C. (2005) ‘Functional analysis and comparative genomics of expressed sequence tags from the lycophyte Selaginella moellendorffii.’, BMC genomics, 6, p. 85. doi: 10.1186/1471-2164-6-85.