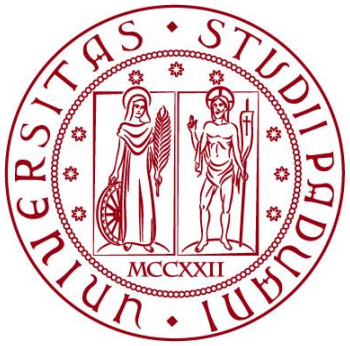
UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI BIOLOGIA

Corso di Laurea magistrale in Biologia Evoluzionistica



TESI DI LAUREA

Computational estimation of genome size and its evolution in relation to body size in the fast adaptive radiation of African cichlids

Relatore: Prof.ssa Chiara Papetti
Dipartimento di Biologia

Correlatore: Prof. Hugo Gante
Department of biology, KU Leuven

Laureando: Pietro Antolini

ANNO ACCADEMICO 2022/2023

INDICE

1. INTRODUCTION

Adaptive radiations are key phenomena that scientists investigate to understand how speciation occurs and to answer many of our evolutionary questions. They consist in the rapid diversification of an ancestral population into several ecologically different species, associated with adaptive morphological or physiological divergence (Schluter et al., 2000).

1.1. East African cichlids radiation

Among all the examples of this process, the cichlid radiations that occurred in the African Great Lakes are some of the most spectacular and the biggest among vertebrates. Thousands of species of these fishes, several not formally described yet, emerged in a variety of ecological and morphological forms, in a span of time that, geologically speaking, could be considered a blink of an eye. Although phylogenetically linked, every big lake, and even some of the smaller ones, hosts its own radiation that gave life to a plethora of endemic species, whose complex evolutionary history and rapid explosion have been thrilling scientists for decades. Lake Tanganyika cichlid radiation (250 species) is the oldest of the area and the most diverse from an ecological, behavioural, morphological and genetical point of view (Salzburger et al., 2014; Svardal, H., Salzburger, W. & Malinsky, 2021). Taxonomically endemic species are grouped in 16 tribes (14 for some authors, 12 for others), with quite different for number of species. Apart of representatives of the three tribes Coptodonini, Oreochromini and Tylochromini that come from a secondary colonization, all the other tribes are endemic, having evolved and diversified in situ, with a most recent common ancestor that lived 9.6 My ago (Ronco et al., 2021). One of these tribes, Haplochromini, have an incredible history of diversification since they colonized the rivers near Tanganyika and then invaded the lacustrine habitats of Malawi and Victoria, giving life to the entire radiation in these two lakes, in addition to a secondary radiation of this group in Tanganyika (Danley et al., 2012). Lake Malawi radiation (800-1000 species), according to Salzburger, 2018, started 800.000 years ago when the lake transitioned to a more or

less closed system and the current deep-water conditions arose. The delimitation of this impressive number of species is more complex since they diverged recently, and this is reflected by different subgroup proposals from different authors (Malinsky et al., 2018; Salzburger, 2018). Lake Victoria Species Flock (LVSF) counts approximatively 700 species and inhabits not only Lake Victoria but also nearby water bodies and even North African ones (only a couple of species). It is the product of the youngest of the African cichlid radiations with an estimated onset dated between 100.000 and 200.000 years ago, but evidence show that most of species diversified only after the recolonization of the Lake Victoria after a period of complete desiccation, 15.000 years ago.

Geological and paleoclimatic studies showed that the unique event of African cichlid diversification had been possible only through the establishment of particularly favourable conditions. First, East Africa was subverted by the rise of the East African Rift System (EARS) at the border between Tanzanian and African plates, whose two branches started soaring respectively around 30-35 My ago for the eastern part and around 25-12 My ago for the western one. This uplift involved the formation of half graben structures and the consequent creation of basins where the water could be collected. Both Lake Tanganyika (9-12 My ago) and Lake Malawi (>8.6 My ago) formed in this way. Moreover, this new geomorphological conformation brought by the rise of the rift, entailed a change of river flow towards a topographic low between the two branches of the EARS that resulted in the formation of Lake Victoria (>0.4-1.6 My ago) (Danley et al., 2012). The consequence of this process was the opening of new-born available ecological niches to exploit without competitors. (Figure 1).

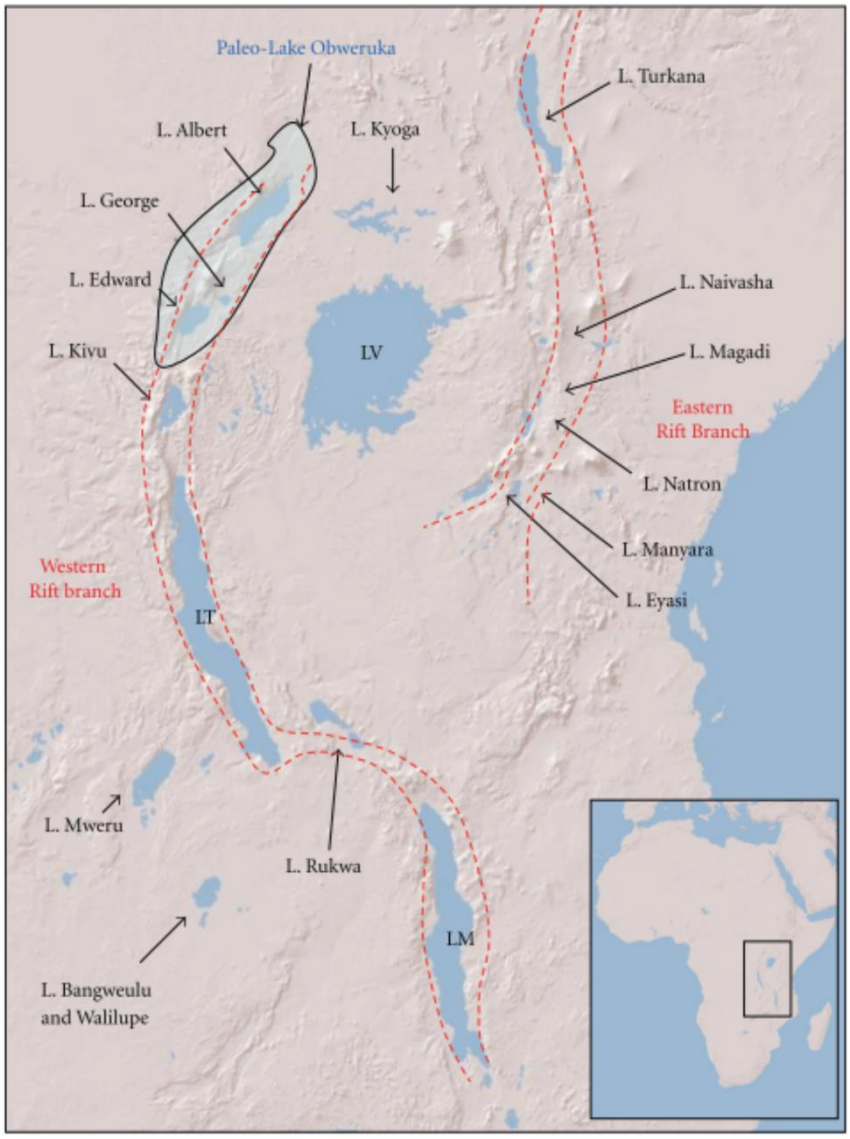


Figure 1. Geographic position of the study region and location of the East African rift. The approximate locations of the two main branches of the East African rift system are displayed in red-dashed line (LV: Lake Victoria, LT: Lake Tanganyika, and LM: Lake Malawi). Credits: Danley et al., 2012.

In this scenario the success of founder cichlids is understandable and corroborated by the fact that other taxa of the area, for example ostracods and gastropods but also non-cichlid fish families, faced a similar rapid diversification within these lakes, which gave birth to several endemic species (Salzburger, 2018). But even though the process of adaptive radiation is common across all these taxa, the number of endemic species of cichlids that were generated is at least one order of magnitude higher in comparison. So, the ecological opportunity provided by the African Great Lakes explains the diversification but not its width and its maintenance. Other

factors peculiar to this fish taxon should have played a role in this outstanding display of life.

1.1.1. Cichlidae

Cichlids include the highest number of genera of teleost with around 1700 species recognized to date, but more than 3000 estimated (Salzburger, 2018). They show a Gondwanan distribution with species that inhabits fresh and brackish waters of India, Sri Lanka, Middle East, Africa, South and Central America, Mexico and even Texas.

The diversity within this group is not only taxonomic, but it permeates different levels of biological organization. Morphologically speaking they show a wide range of body shapes (from almost roundish to elongated) and sizes (from 3 cm to almost 1 m), but are the structure that fulfil an ecological purpose, as for example the upper jaw and the peculiar pharyngeal jaw apparatus, that shows the highest variety in this group (Ronco et al., 2021). This is not a surprise because, in the case of adaptive radiation, diversification occurs via niche specialization so, a strong association is expected in the extant fauna between the environment occupied by a species and the specific morphological features used to exploit it (Schluter et al., 2000). For this reason, cichlids show also a great ecological variability. They occupy multiple rings in the lacustrine and riverine food chain, ranging from pelagic fish predators to benthic algae grazers, from scale eaters to planktivorous and detritivores. They inhabit rocky, weedy, muddy and sandy substrates and some species are shell dwellers, so they use barks of other animals as a refugia. Although few species inhabit primarily brackish water or even salt water, cichlids occur mostly in freshwater, either in streams, rivers and lakes, where they range from the shallow water habitats to the deeper ones (Salzburger, 2018).

This group also shows a broad behavioural repertoire, beyond the one proper of a particular feeding niche. Many species have dominant males who use aggression displays to defend their territories and courting to attract females, while others form large schools (Salzburger, 2018). The mating behaviour is either substrate brooding or mouth brooding with both groups that show a wide range of intraspecific differences. The former system usually results in a lower investment in parental

care, while the latter can show maternal, paternal or biparental care (Keenlyside, 1991).

Pigmentation has a fundamental role in both intersexual and intrasexual rituals, so it is not surprising that cichlid species display an incredible array of colours and patterning. One of the most amazing examples is the evolutionary innovation of the anal fin egg-spots in haplochromines, coloured circular markings that vary substantially in colour, shape, number and arrangement between species and even within species (Figure 2). In the mouth brooding mating system of this tribe, egg-spots are presented by the male to the female who responds by snatching and bringing her mouth close to the male's genital opening. Here sperm are discharged and can fertilize the eggs inside the female's mouth (Santos et al., 2014).

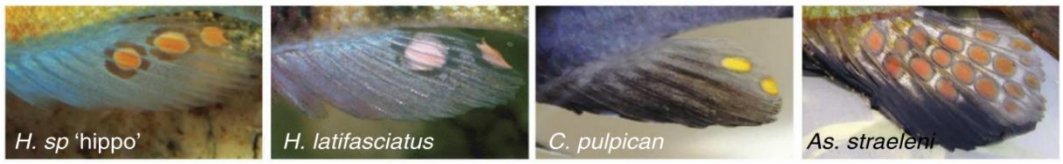


Figure 2. Anal fin Haplochromines egg-spots. Credits: Santos et al., 2014.

Interestingly, wherever they occur, cichlids show a strong tendency to form species flocks, an amazing display of diversification (Salzburger, 2018) that according to Chenuil et al. (2018) occurs when a monophyletic taxon displays high speciosity in an area in which it is endemic (the exclusive occurrence of a species or higher-level taxon in a confined geographic area), high ecological diversity among species, and if it dominates the habitat in terms of biomass. Species flock are most of the times outcome of adaptive radiations, so this taxon appears particularly prone to give life to these spectacular evolutionary explosions. But why? The answer is probably hidden in their genomic features, so an increasing number of studies started investigating the DNA of these fishes, supported by the great innovations in sequencing techniques and data analysis.

1.1.2. Cichlid genomes

Understanding the genomic features of a species or a group requires valuable assemblies and the possibility to confront sequences at an interspecific but also intraspecific level with more or less phylogenetically close species. For this reason,

a reliable phylogenetic reconstruction is needed. However, in a scenario like the one of the African Great Lakes cichlid radiation this task can be quite challenging because the speciation is young (and so genomes are more similar) and because of the high amount of incomplete lineage sorting (ILS) and introgression present in their genomes. ILS is typical in rapidly speciating lineages, and it consists of the maintenance of ancient polymorphism across different speciation events. It generates incongruences between gene tree and species tree topologies (Maddison et al., 2006). On the other hand, introgression is the transfer of genetic material from a species to another due to backcrossing of hybrids with one of the parental species, that can easily result into misleading phylogenies if not previously identified (Aguillon et al., 2022). Salzburger (2018) defines genomes of rapid diversifying cichlids as mosaic genomes, where different small fragments of the sequence tell different evolutionary histories.

From the sequencing of the first five phylogenetically representative species by Brawand et al. (2014) the amount of genomic data on African cichlids has grown and several studies have proposed phylogenies based on genome wide markers and started comparative genomic investigations.

Accelerated gene evolution was found in Brawand et al.'s (2014) five species with high rates of non-synonymous to synonymous substitutions in the coding sequence compared with Nile Tilapia. Even the gene regulatory portion of the genome shows a greater dynamic compared to other fish species, with higher rate of nucleotide substitutions and of insertions and deletions, evidence of relaxed purifying selection in cichlid genomes.

The common idea is that natural selection is the main force that drives cichlid diversification, supported by the several examples of convergent evolution of similar ecotypes across different radiations, but a key role is played also by strong sexual selection as suggested by the high variability in male nuptial colorations (Brawand et al., 2014; Salzburger, 2018).

The raw material on which evolutionary forces can act to generate adaptation to new environments is the genetic variation at functional loci, linked with fitness-related phenotypes that can be inherited by the next generation. In spite of their great phenotypic diversity, Great African Lakes cichlid genomes show relatively low mean genetic variation compared with other vertebrates. Moreover, most of

their allele variation is shared by species of the same radiation, but surprisingly also by those of different ones (Svardal et al., 2021). The youth of the diversification and the inevitable gene flow of a sympatric speciation, clearly explains part of this DNA sharing and similarity, but a key role could have been played by the climate changes that interested Eastern Africa from the onset of cichlid explosion. Alternation of arid and wet climates caused cyclic low lake levels and even desiccations (Danley et al., 2012), mirrored by fragmentation and contraction of fish populations. During those bottlenecks the action of genetic drift probably strongly reduced the existing genetic diversity, and the gathering of survivors in refugia supported hybridization between them.

Overall, these results suggest that low genome-wide nucleotide diversity levels may not limit rapid adaptation and speciation. Though, Svardal et al. (2020) noticed also that even if low, genetic variation is quite variable across the genome, with extended outlier regions of high diversity, introducing the possibility that variation is higher at functional loci in cichlids.

Regarding how this genetic variation arose, the idea is that it was accumulated before the beginning of the radiation during a period of relaxed selective constraints through different evolutionary mechanisms (Brawand et al., 2014; Salzburger, 2018), and apparently not due to an increase in the nucleotide mutational rate whose estimate is too slow to explain the actual divergence (Malinski et al., 2018).

First, high rates of lineage specific gene duplication were found in this group. In particular, in the common ancestor of the East African cichlids it was estimated 4.5-6-fold higher than in other clades and it is even higher in the common ancestor of haplochromines (Brawand et al., 2014). Duplication offers divergence possibilities allowing neofunctionalization of the new produced copy and/or modification of the duplicated gene protein expression pattern (Zhang, 2003).

Transposable elements (TE) are another source of variation. They are repetitive elements able to replicate and insert themselves in the genome, including upstream of genes modifying their expressions. TE, with their potential of producing large changes in genome, can also create genetic interspecific incompatibilities favouring speciation (Serrato-Capuchina et al., 2018). An example of their evolutionary potential is the insertion responsible for the development of the egg-spots in haplochromines (Santos et al., 2014). Three or four waves of TE expansion were

detected in cichlid genomes including a group-specific burst of one of the TE families (Brawand et al., 2014).

Another dynamic that can boost genetic variation is recombination, since it can assemble new genotypes producing combination of alleles previously not explored by evolution. Its contribution is stronger the more distinct are the parents involved, as for example during hybridization between two species (Mallet, 2007). Cichlid hybridization is largely reported even between different lake radiations, and introgressed material can be found throughout all these fish genomes (Brawand et al., 2014; Stelkens et al., 2015; Malinski et al.; 2018; Keller et al., 2013). Those results could be only the by-product of a limited geographical and reproductive isolation, but experimental work (Stelkens et al., 2009) and observation in the wild (Nichols et al., 2015) testify that hybridization in cichlids can produce new and extreme phenotypes whose success seems to depend on available ecological resources outside of the parental niches (Selz & Seehausen, 2019). It is a general consensus now that hybridization in East African cichlids has been an important factor for establishing and maintaining genetic variation (Gante et al., 2016; Salzburger, 2018; Svardal et al., 2021) with evidence of this process before or early in the adaptive radiation of LVSF (Meier et al., 2017) and Malawi cichlids (Svardal et al., 2020) and between the early lineages of Tanganyika (Irisarri et al., 2018).

All these dynamics other than generating new genetic diversity, can have an impact on genome size (GS) contained in the cellular nucleus. Duplication contribution is probably the most straightforward: having two, or more, copies of the same gene retained, the genome will increase its size. Transposable elements, when not constrained, can replicate in an uncontrolled way and insert themselves multiple times in the host genome, causing genomic expansions. Finally, hybrids show a genomic content quite variable from the mean of the two parental species. This seems to depend on various factors including genomic features of parental species, transmission bias, genomic rearrangements but the most evident differences are linked to TE expansions (Romero-Soriano et al., 2016). Hybridization appears able to break down mechanisms of TEs repression present in the parental lineages releasing their evolutionary and replication potential in hybrids (Dion-Cotè et al., 2014; Wright, 2017).

Brawand et al. (2014) found relatively similar GS for his five species, but the estimation came from assemblies, which crafting methods usually struggle to recreate repetitive regions of the genome (Sun et al., 2018), regions that are precisely the main contributors for GS changes (Tenaillon et al., 2010; Chalopin et al., 2015).

From these premises, it seems plausible to expect intra- and interspecific GS variation in African Great Lakes cichlids, but no studies have demonstrated it yet.

1.2. Genome size

Genome size (GS) is commonly described as the amount of haploid nuclear DNA of an organism, typically measured in Megabases or picograms (1 pg = 978 Mb) (Doležel et al., 2010).

It is a highly variable trait across eukaryotes with differences that range at least four orders of magnitude, as for example from the 19 Megabases of *Pratylenchus coffeae*, a parasitic nematode, to the 130 Gigabases of the marbled lungfish, *Protopterus aethiopicus* (Gregory, 2005) the animal with the biggest GS known to date. Although considerably smaller, differences within species, between sexes but also in general between individuals, are also present (Marescalchi et al., 1998; Jeffery et al., 2016; Neiman et al., 2011; Romero-Soriano et al., 2016).

This striking variation does not correlate neither with organismal complexity, nor with number of genes (at least for eukaryotes) (Thomas, 1971), a puzzling scenario that scientists were not able to explain for a long time, to the point that they referred to it as the C-value paradox.

Today we know that the nature of most of this extra DNA is non-coding and repetitive, with the coding part of the genome that correlates strongly and negatively with GS (Elliot and Gregory, 2015), but the reasons of its evolution and the mechanisms of its accumulation are still matter of debate.

1.2.1. Mutational mechanisms

Whatever the evolutionary scenarios of GS change might be, they must involve mutational mechanisms of addition and loss of DNA. Those differ for the entity of

the modification and so for the evolutionary time scale over which they can be effective (Petrov, 2001). The genome-size variants that arise from them sometimes affect phenotype and thus have to go through natural selection before becoming fixed. It is also likely that, within a certain range, genome-size variants could be of such similar selective values that their ultimate fates are determined primarily by neutral drift (Petrov, 2001). Moreover, singularly these mechanisms can have fitness-related effects that can go under selection independently from their impact on GS.

Polyploidy or whole genome duplication is probably the most obvious of these mechanisms, with a huge impact on genome size and its evolution from a generation to the next. Usually, after this event the genome faces a series of chromosomal rearrangement and a process of deduplication (loss of most of the duplicated genes), both events that can mask the signs of an ancient polyploidization. This process is expected to cause a rapid and huge increase of GS and to carry all the phenotypic effects associated with those kinds of events (e.g., bigger cells, possible increase in developmental time) (Wright, 2017).

Another chromosome-level impact on GS can be due to B-chromosomes, that are usually smaller than regular A-chromosomes and derive from them. They are considered as selfish genetic elements that can segregate independently at meiosis and often exhibit meiotic drive (transmission of one or more alleles favoured over another), that give them a way to quickly spread through populations causing changes in GS (Blommaert, 2020).

Of relatively fast impact are amplification of repetitive DNA classes, like satellite DNA (passive proliferation, usually through DNA polymerase slippage), ribosomal DNA and transposable elements (TE).

TE are selfish genetic elements that can actively amplify themselves throughout the genome, either via their own transposase enzymes, or by recruiting those of other elements (Blommaert, 2020). TE copy number can increase by 20-100 copies (0.1-1 Mbp) in a single generation (Petrov, 2001) and its changes among species are often the most important predictor of genome size disparities (Tenaillon et al., 2010; Chalopin et al., 2015). Different processes and dynamics govern TE activity. They spread more in species with sexual reproduction and outcrossing, while low Ne populations tend to reduce the number of active TE due to genetic drift. Their

removal can be driven not only by deletion events but also by natural selection against both GS expansions and the harmful phenotypic effects from their insertions (interruption of functional genes and loss of function, disruption of gene expression) and even by illegitimate recombination between their long terminal repeats (Wright, 2017). Furthermore, modifications in silencing mechanisms that hosts adopt to limit TE activity can result in changes in their copy number. Both hybridization and duplication have been found capable to alter this control system (Marburger et al., 2018; Dion-Cotè et al., 2014).

Spontaneous insertions and deletions are really slow mechanisms and therefore not plausible candidates to explain GS changes among closely related species or within species (Petrov, 2001).

1.2.2. Evolutionary framework

Two main theoretic hypotheses are currently debated for GS evolution, and each of them takes in account a different evolutionary force: a neutral (or nearly neutral) model that considers genetic drift as the main ruler of GS variation and an adaptive one that suggests that it is instead mostly driven by natural selection. The neutral scenario can be divided in two sub-branches that share the idea that the accumulation of new material, and so the insertion process, is regulated only by drift, but differ regarding the explanation of how it is removed. One, more strictly neutral, argues that GS reflects species specific insertion/deletion rates, and that big expansions of the genome are related with bursts in transposon activity or/and duplication events, then buffered by a constant rate of small deletions (Blommaert, 2020). Indeed, several studies observed a distinct mutational spectrum of insertions and deletions in different species (Petrov et al., 1996, 2000; Bensasson et al., 2001) but a role of natural selection in the fixation of indels cannot be ruled out (Charlesworth, 1996) and, to date, a correlation between species-specific insertion/deletion rates and GS seems not to be present (Wright, 2017). More supported is the idea that transposon bursts followed by deletions is one of the natural processes that shape GS as observed across birds and mammals by Kapusta et al. (2017). The second sub-branch of the neutral hypothesis considers insertions as slightly deleterious for the genomes, at the point that, in a scenario of low genetic drift (e.g., high effective population size – N_e), natural selection-driven deletions

occur reducing GS. This creates a clear prediction: GS negatively correlates with N_e. That seems consistent at a broad taxonomic scale (Lynch and Conery, 2003), but fails to find validation at smaller scale or when comparative phylogenetic approaches are adopted (Whitney et al., 2010; Whitney and Garland, 2010; Ai et al., 2012). Moreover, small deletions could be as deleterious or more deleterious than small insertions (Leushkin et al., 2013), raising doubts on the effective direction that GS would take in case of scarce natural selection effect (Wright, 2017).

What emerges is that even if support for neutral hypothesis is not always confirmed, neutral evolutionary forces clearly play a role in GS evolution, at least as the default starting point on which other forces may act (Arkhipova, 2018; Blommaert, 2020). On the other hand, several other theories sustain an adaptive nature of GS that can therefore be shaped diversely by natural selection depending on the ecological pressures that the organism must face. A strong positive correlation exists between GS and cell size (Tsukaya, 2013), but also were observed correlations with body size, developmental time, seed size, duration of mitosis and meiosis (Beaulieu et al., 2008; Šímová et al., 2012; Gregory et al., 2002; Bennet, 1987; Chung, J. et al., 1998) and other traits indeed linked with fitness and therefore plausible subjects of directional selection. Further proof that the phenotypic impacts of GS may be under selection comes from the fact that clines of this trait were found associated with environmental factors (Rayburn and Auger, 1990). Several studies show support for this hypothesis in different taxa, genome shrinkage driven by selection for rapid cell division and fast metabolism was for example observed in weedy plants (Bennett et al., 1998), animals with flight (Wright et al., 2014), and parasites (Cavalier-Smith, 2005).

1.2.2.1. Body size-genome size correlation

Correlation between body size and genome size was observed in different taxonomic groups, although the reasons at the base of this association could be different and bring to a different relationship between the two variables.

Positive correlations were found in copepods species (Gregory et al., 2000) probably because they exhibit determinate cell number, so, to reach bigger size, they have to increment their cell size and, since it is directly correlated, genome

size. In salamanders, the positive relationship seems to be a consequence both of physical and morphogenetic constraints associated with large cells in small species and of mutational pressure to increase genome size once such constraints are lifted in larger species (Decena-Segarra et al., 2020). It is particularly evident in species that went through miniaturization (evolution of extremely small adult body size from a larger ancestor), since the consequences of body size reduction are largely mediated through cell size because the number of cells that form tissues or organ systems affects their complexity and function (Roth et al. 1994). Other species showing a positive correlation are flatworms (Gregory et al., 2000), birds and subsets of mammals (Gregory et al., 2002), but the causes for these associations are less clear. In fish Smith & Gregory (2009) found instead a negative correlation of these two variables, that could be explained by an inverse association between genome size and developmental rate in organisms of indeterminate growth. Bigger genomes need more time for mitosis and meiosis, producing an increase in developmental time and therefore potentially smaller sizes at an adult stage.

It is interesting to notice that most of the evidence of the contrasting evolutionary hypothesis of GS are based on wide phylogenetic scale comparisons. Blommaert (2020) suggests a shift of focus on population-level differences in GS to exclude the confounding effects of large phylogenetic distances, before drawing conclusions about the importance of natural selection versus genetic drift. Few examples can be found in literature: a study on maize revealed that GS went through selection due to its effects on flowering time at different altitudes (Bilinski et al., 2018), while another found that seed beetles specific GS differences are linked with difference in reproductive fitness (Arnqvist et al., 2015). Though, to date, studies that take in account interspecific GS differences and their phenotypic effects remain scarce.

1.3. Aim of the thesis

With this study we try to contribute to filling this gap, searching for variation in GS within Lake Victoria Super Flock (LVSF) radiation and investigating if those changes can be associated to body size variations in those species, possible evidence

for an adaptive history of GS diversity. To reach this aim we build a specific bioinformatic pipeline to estimate GS starting from whole genome shotgun Illumina reads and compare 210 individuals from 155 species belonging to the different radiations of Victoria region. To understand better the evolution of the trait we use ancestral reconstruction analysis plotting it on a Maximum Likelihood (ML) tree. To explore its relationship with body size we run regression analysis using phylogeny independent and dependent methods to take in account the effect of relatedness between species. Finally, we run average subclade disparity analysis for both GS and body size to see how it behaves through time. Even if our analysis is not focused on populations, the recent origin and the high relatedness of LVSF species should be enough to rule out effects due to large phylogenetic distances.

1.4. Lake Victoria Super Flock

Lake Victoria is the second largest lake of the planet by surface and spans the equator in between the East and the West branches of the EARS. It is hydrologically open with Kagera and Katonga rivers as major inlets and Victoria Nile as primary outlet that connects it to Lake Albert. Since it is not a rift basin, it is shallower than the other Great Lakes, with a maximum depth inferior to 100 meters. This feature makes it dependent from the balance between evaporation and precipitation, and since the former is more constant, the variability of the latter has a really strong impact on the lake level. For this reason, Victoria went through a series of desiccation events during its (young) geological history. Several smaller water bodies punctuate Victoria surroundings, comprehending rivers, streams, small young lakes and even older deep lakes like Kivu, Albert, Edward, George and Kyoga.

Lake Victoria Super Flock originated between 200 and 100 thousand years ago (Verheyen et al., 2003) and comprehends around 700 species of haplochromines, spread mainly in lakes and rivers of Lake Victoria region but also in water bodies of North Africa and Israel (Danley et al., 2012). They are a monophyletic group sister of Malawi haplochromines with a common Tanganyika ancestor evolved around 4.7 million years ago (Elmer et al., 2009). Lake Victoria radiation counts

around 500 endemic species, probably diversified in the last 15 thousand years during a secondary colonization of the lake after a desiccation. Also, other big lakes of the area, Kivu, Albert, and Edward, host their own radiations, even if less copious (Meier et al., 2017). Each radiation comprises enormous diversity in habitat occupation, trophic ecology, coloration, and behaviour. The phylogeny of the group is particularly challenging and reflects the complex phylogeographic pattern due to past geological and climatic events.

Mitochondrial (Verheyen et al., 2003) and nuclear (Elmer et al., 2009) analysis support Lake Kivu as the evolutionary origin of the LVSF. Haplotype network reconstruction by Verheyen et al. (2003), shows indeed that all LVSF lakes haplotypes are directly connected with, and therefore derived from, Lake Kivu's and that the species of this radiation show a higher level of interspecific genetic diversity compared to the other lakes of the region even if counting only 15 endemic species. Apparently, from this older and deeper lake four different lineages seeded the northern rift basins of Lake Edward, George, and Albert in a stepwise manner (Figure 3), through a connection that was interrupted recently (25,000-11,000 years ago) by the uplift of Virunga Volcanoes. This event blocked the northward drainage of Lake Kivu creating the modern river systems east of it (Schmidt, 2001). At least two lineages moved from Lake Kivu to Victoria (Figure 3) through a no longer existing connection and diversified (Verheyen et al., 2003). Elmer et al.'s (2009) data show, indeed, an introgression of genes of the first lake into the basin gene pool of the second, and not the reverse, reflecting the historical direction of migration. The split between Victoria's and Kivu's faunas has been estimated between 41,500 and 30,000 years ago, really close to the one of Virunga eruption making scientists think of a correlation of the two events that though has not been confirmed yet (Danley et al., 2012).

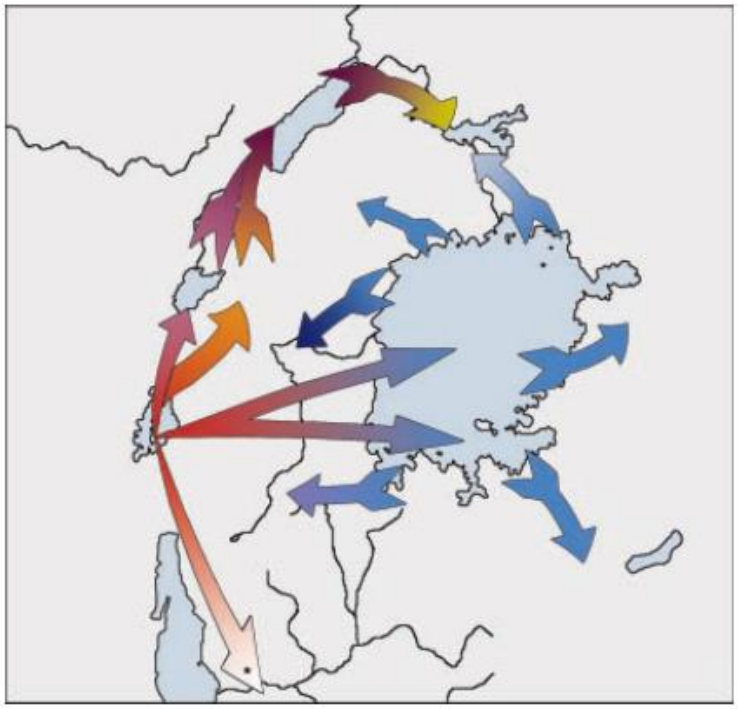


Figure 3. Possible scenario of colonization between lakes of Victoria region. Credits: Verheyen et al., 2003.

The genomes of this group show a level of genetic variation higher compared to the mean of the rest of Great African Lakes cichlids (Svardal et al., 2021) and it is surprisingly high for LVSF recent origin (Bezault et al., 2011), suggesting that large amount of standing variation must have been present at the onset of the radiation. Hybridization is now recognized as the main source of this variation and the evolutionary history of this group is now considered an example of a hybrid swarm origin (Meier et al., 2017). This hypothesis considers an ancient hybridization between distant lineages as the genetic diversity provider to start and maintain a subsequent speciation. Meier et al. (2017) found evidence of ancient admixture at the onset of the LVSF radiation between two distantly related lineages of haplochromines that evolve in isolation in two river systems for more than a million years, one from the Upper Congo and one from the Upper Nile drainage. Probably, the former entered the Victoria region during a humid phase around 145,000-200,000 years ago through the tributaries of Malgarasi river (a connection not existing now), and there it found the latter already established. Hybridization and introgression events derived from this area of contact, presumably allow these species to fully exploit the ecological opportunities offered by big lake habitats

starting the diversification process. This is supported by the fact that genomic loci likely involved in adaptation and divergence in Lake Victoria, show alternative alleles fixed in the hybridizing lineages but high divergence in Victoria species, which means that the sorting of alleles brought together by the admixture event is linked with speciation. One example of this dynamic is the LWS opsin gene, an exceptionally diverse gene in Victoria region that coded for the protein part of red-sensitive visual pigments in cones (Terai et al., 2002). It plays an important role in adaptation to different ambient light levels and in reproductive isolation because divergent colour perception is associated with divergent male nuptial coloration (Terai et al., 2006). The variety of alleles at this locus are usually grouped in two major haplotype classes that differ mainly for a substitution that shift the absorbance peak of sensitivity and are usually associated to different habitats (clearer and shallow water and deep and murkier one) and each class is shared only with one parental lineage.

2. METHODS

Short read sequence data were provided by Nathan Vranken, collected during the project KEAFish, funded by BELSPO (Belgian Federal Science Policy). Next Generation sequencing reads were obtained from Illumina whole genome sequencing of DNA extracted from ethanol-preserved fin tissue of cichlid fishes captured in lakes Kivu, Edward, Albert, and Victoria. Dataset consists in 300 samples of 180 species with a mean of 1.67 individuals per species (range 1-34). Raw reads are 150 base pairs long and they are paired end (PE). PE sequencing enables both end of the DNA fragment to be sequenced with a known distance between each paired read, an information that allow alignment algorithms to map the reads over repetitive regions more precisely. For all the species ecological data were provided and for 133 of them also the maximum known standard length (SL) is reported, either from literature or from measuring the biggest captured specimen.

2.1. Genome size estimation (GSE)

Estimating genome size is not a trivial procedure and there are different approaches to tackle this challenge. A first big division that could be made is between experimental wet lab methods and computational ones.

The former category includes techniques like Feulgen densitometry and flow cytometry that have been used for many years and were applied to tens of thousands of species, currently compiled in a GS database (Gregory et al., 2007). DNA flow cytometry is the most adopted method and involves preparation of aqueous suspensions of intact nuclei whose DNA is stained using a DNA fluorochrome. The nuclei are classified according to their relative fluorescence intensity or DNA content (Doležel and Bartoš, 2005), but this means that a set of standards is needed Even though, all of these experimental techniques rely on specific genomes that serve as internal and external size standards (Bennett et al., 2003; Doležel et al., 2007; Hardie et al., 2002), the goal of creating a set of commonly accepted standards has not yet been achieved (Doležel and Bartoš, 2005; Doležel and Greilhuber, 2010). When the same genome is analyzed in multiple laboratories,

this and additional factors such variations in sample preparation, staining/dyeing method, and stochastic drift of instruments can lead to noticeable changes in GSE for the same genome (Doležel et al., 1998).

From the point of view of our resources, experimental methods would not be an option, as they require fresh tissue from which to extract cells and nuclei. This is not possible in our situation because our samples were already collected and sequenced, and even with another expedition it would be very difficult to preserve them well during the transport from Africa.

Computational approaches, on the other hand, appear more accessible. Computational methods, indeed, are based on whole genome sequencing and bioinformatics tools. One approach of this kind involves assembling the genome whose size we want to estimate and counting the number of bases. This is usually biased because even with the latest algorithms it is difficult to fully assemble using short reads and include highly repetitive regions and non-coding ones (Torresen et al., 2019), so the genome size obtained in this way would result in an underestimation. Moreover, our coverage is too low to craft a reliable assembly.

The choice instead was to use what is called a k-mer approach. The advantage of this method is that it requires as an input whole genome shotgun reads with a remarkably lower coverage that the ones necessary for crafting an assembly. On the other end it could be computationally heavy.

Since we have a valuable number of computational resources thanks to the Vlaams Supercomputer Centrum access and good quality reads, even with low depth coverage, the kmer approach is the best choice for our GSE effort.

Furthermore, as stated by Sun et al, 2018, k-mer based GSE methods are very robust against changes in the sequencing setup, and their estimations can be highly reproducible.

In bioinformatics, k-mers are substrings of length k contained within a biological sequence (Compeau et al., 2011). In the field of computational biology, they are used for different analysis from the construction of De Brujin graph during the first step for crafting an assembly, (Compeau et al., 2011) to genomics-based taxonomy (Ounit et al., 2015), from phylogenetic purposes (Zhang, Q. et al., 2017) to genome size estimation (Pflug et al., 2020).

The k-mer approach for GSE starts from the assumption that NGS methods sequence all bases in a genome with equal probability, so if we divide the number of bases sequenced (N) by the length of the genome (GS) we obtain the mean coverage depth (C) (Sims et al., 2014).

$C = N/GS$

If we transfer this idea in a k-mer context, C is the number of times each k-mer is sequenced on average and N denotes the number of genomic k-mers in the reads. The relationship $N=C*(G–k+1)$ allows to estimate GS with $G \approx N/C$ as $G >> k$.
Both C and N can be statistically inferred from a k-mer frequency histogram, which tells us how many distinct k-mers occur at a specific frequency within a given whole-genome sequencing data set (Sun et al., 2018).
In a typical k-mer distribution of a diploid genome we have, starting from the left, a high peak at low frequencies that results from sequencing errors. They consist of many k-mers present in one or two copies. Usually other two peaks are present, the rightmost is the homozygous peak, characterized by k-mers present in both chromosomes sets, and, at half of homozygous peak frequency, the heterozygous peak with k-mers present in only one set. The higher the heterozygosity, the more dominant the heterozygous peak. The long tail at higher frequencies represents k-mers of repetitive regions, present in multiple loci. Finally, if not previously filtered out, at even higher frequencies we can find small peaks of k-mers from mitochondrial and plastidial genomes which are present in multiple hundred copies in a cell (Sun et al., 2018).
Simply dividing the number of genomic k-mers for the k-mer coverage is though quite an approximate approach due to sequencing errors and other interferences. For this reason, different software packages have implemented mathematical models to fit the distribution of distinct k-mer frequencies including mixed Poisson distributions (Li and Waterman, 2003), Bayesian estimation (Shan & Zheng, 2009) and negative binomial distributions (Vurture et al., 2017).
For our analysis we chose to test two of them that adopt two different models. Given a distribution of k-mer frequencies, findGSE first fits the distribution iteratively with a skew normal distribution model; then it calculates the total number of k-mers (N) according to both fitted and the original counts and corrects the average k-mer

coverage (C) with the skewness of the fitted curve, based on which it calculates the genome size as N/C (Sun et al., 2018). GenomeScope 2.0 takes as input the k-mer spectrum, performs a nonlinear least-squares optimization to fit a mixture of negative binomial distributions, and outputs estimates for genome size, repetitiveness, and heterozygosity rates (Ranallo-Benavidez et al., 2020).

2.2. Pipeline

2.2.1. Pre-processing

fastp -i ${name}_R1.fastq.gz -o ${name}_R1strict.fastq.gz -I ${name}_R2.fastq.gz -O ${name}_R2strict.fastq.gz -w 16 -D --dup_calc_accuracy 5 -l 40 -q 30 -5 -W 2 -M 30 -3 -W 2 -M 30

Before the GSE, Illumina reads need to be pre-processed to remove possible uninterpretable signals due to sequencing errors.

We use the package fastp v0.23.2 (Chen et al., 2018) that provides several filtering options for short reads in FASTQ format and can be run in parallel on multiple cores (-w option). Furthermore, it accepts paired ends reads as input and is able allows to correct mismatched base pairs in overlapped regions of paired end reads, if one base is with high quality while the other is with ultra-low quality.

Before applying filters to the whole dataset, we tested them on different datasets of Illumina WGS reads of Nile tilapia (*Oreochromis niloticus*) downloaded from assembly projects in NCBI to benchmark the entire process. We distinguished a good initial read quality sample (gNT) and a bad initial read quality sample (bNT). This African cichlid has relatively good assemblies but, more importantly, it has genome size estimates from independent methods, so it is the perfect candidate for our benchmarking process. The GSE range for Nile tilapia goes from 0.95 billion bases to 1.15 billion according to Feulgen densitometry method (Gregory et al, 2005).

First, we discarded reads of low quality and bases with low phred quality score. A phred quality score is a measure of the quality of the identification of

the nucleobases generated by automated DNA sequencing. (Ewing B; Hillier L; Wendl MC; Green P. (1998))

Keeping bases with a low phred quality score would lower the quality of downstream analyses (e.g., de novo and reference-based assembly), by introducing sequencing artifacts and errors that may contribute to incorrect interpretation of data (Chen et al., 2018) and incorrect k-mer count. Moreover, estimation models are more prone to not converge when fed with low quality reads (Vurture et al., 2017). Running the entire pipeline for two samples of Nile tilapia WGS reads of different starting quality (gNT, bNT) allow us to assess how the starting quality impacts the analysis and if our trimming design will be able to, at least partially, correct erroneous estimations related with sequencing uncertainties.

We performed a Fastqc analysis of the samples before and after the filtering to see how the quality had changed. We noticed that most of the low phred score (<28) bases were concentrated at the 3'and 5' positions of the reads. For this reason, we used fastp --cut tail and --cut front options to create two sliding windows of size 2 bp (-W 2), respectively starting from the front and from the tail of the read, that drops every couple of bases with mean score inferior of 30 (-M 30). When they encounter two bases with a mean quality above this threshold, they stop.

To drop low quality reads we modify the -q option of fastp, which allows us to impose a quality threshold under which the bases are considered unqualified. When a certain percentage of bases of a read is unqualified (-u option, 40% by default), the read will be dropped. We kept -u at the default value and compared a "soft" trimming regime (-q 20) with a "strict" (-q 30) one.

The downside of a more stringent filtering scenario is that a higher number of reads are dumped and that the remaining ones are characterized by a lower assembly coverage, with the consequent risk that in some cases data are too few for the estimation model to converge and to give a reliable GSE.

We proceeded following the suggested pre-processing steps presented in the findGSE publication (Sun et al., 2018). Fastp algorithm automatically search for adapters-like sequences and trim them, so we keep it by default.

Next, we need to remove PCR duplicates. PCR amplification is an important step in the preparation of DNA sequencing libraries prior to high-throughput sequencing but it introduces redundant reads in the sequence data. Clearly those kinds of

artifacts will affect the k-mer counting step and, consequently, the GSE, therefore is important to estimate the PCR duplication rate and filter them out. The -D option of fastp is designed exactly for this purpose. We chose a deduplication calculation accuracy of 5 on 6 (--dup_calc_acc), to have a good trade-off between the speed and the accuracy of the process.

Mitochondrial reads are not included in the genome size calculation and are present in many copies in a cell. Consequently, they could be problematic, creating a high frequency k-mer peak in the histogram that could produce erroneous estimates and they need to be removed. We accomplished this task for our benchmarking species Nile Tilapia aligning our reads to the mitogenome assembly of the species (previously downloaded from NCBI) and filtering out all the ones that align with it. Although this approach will remove also nuclear reads similar enough to the mitochondrial sequence, we expect that their effective number will be limited and so, not impactful in the estimation stage (Pflug et al., 2020). We used BWA (Li and Durbin, 2009) and SAMtools (Li et al., 2009) to reach this goal. The mitochondrial reads removal was performed after the filtering since, if done before it, it can remove duplication signals necessary to dump PCR duplicates.

Another issue can emerge intrinsically related with technical features of Illumina NGS. Since G bases are codified in the same way as no signals during sequencing, a variable number of polyGs (e.g., part of the sequence that consists in a repetition of G) can be identified at the level of the tail of the reads. Fastp algorithm trim them by default using a minimum length threshold of 10.

Intensive trimming produces shorter reads. Until their length remains above the chosen k-mer size, the downward analysis should work correctly. To avoid that reads too short enter the kmer counting step, we use the -l option of fastp, that filter out reads shorter than a certain threshold, in this case 40 bp.

2.2.2. K-mer counting

kmc -m100 -k21 -ci1 -cx1000000 -cs1000000 -t36 @LIST ${name}k21.res tem

kmc_tools transform ${name}k21.res histogram ${name}k21.histo

The next step of the analysis is k-mer counting that consists in determining all unique k-symbol long strings (usually with counters) in the read collection. We used the package KMC v3.2.1 for this procedure and kmc_tools for manipulation of sets produced with KMC (Kokot et al., 2017).

To obtain a single kmer spectrum for the forward and the inverse reads combined we used a list including the two of them as input for KMC.

The critical decision for this step is what k to use, however it is not obvious which one is optimal (Chikhi and Medvedev, 2014). The larger it is, more computational resources requests and smaller the total number of k-mer, with the risk to not have enough information for the model to converge. On the other hand, if it is too small, we risk having k-mers that are not unique in the genome (Liu et al. 2013).

Partially following what was performed in the findGSE paper (Sun et al., 2018) we tested different, mostly odd, ks for both Nile tilapia samples from 17 to 31. We arbitrarily chose 17, 19, 21, 24, 27, 31.

We raised the max amount of RAM (-m100) from the default 12 GB to 100GB to be sure that it's enough to process the high number of reads (and so, of k-mers) of our samples.

To be sure to include all the possible k-mers, we set the frequency lower accepted limit as one (-ci 1) and the upper limit as 1 million (-cx 1000000) while raising the maximal value of a counter to 1 million (-cs 1000000). This could have been problematic because even high frequency mitochondrial reads would have been going to enter the analysis but having already filtered them out that issue doesn't exist.

KMC can be run in parallel and we did it on 36 threads, generating a .res file with every single k-mer and the number it occurs. kmc_tools transform converts it in a k-mer frequency histogram that can be read by the estimation packages.

2.2.3. Estimation

Regarding the estimation packages, findGSE is implemented on R, while Genomescope2 can be run on Linux but also online with a user-friendly interface.

Every benchmarking analysis was performed with both bioinformatic tools to compare their results.

2.2.3.1. FindGSE

library (findGSE)

findGSE (histo ='name_21.histo', outdir = "estimation", sizek = 21)

For findGSE we kept everything in the standard mode, changing only the k-mer size according to the input histogram. Even if for gNT the heterozygous peak was clear in the histogram, it was not present in all our samples, so we didn't include the parameter for heterozygous genomes. This is not completely unexpected, given the low heterozygosity scenario that emerges from Svardal et al. (2021) estimations.

2.2.3.2. Genomescope2

Rscript genomescope.R -i name_21.histo -o output_dir -k 21 -l 8

Regarding Genomescope2 we ran it on Linux because it is possible to modify more parameters in this way. In this case, differently from what we did for findGSE, we kept the -l parameter which consists in giving the initial k-mer coverage estimate of the heterozygous peak. Without it the model returns very low estimations, usually around half the ones with the parameter included.

Even if well optimized these two algorithms cannot work with a small amount of data, because in that case the model is not able to converge. For findGSE a 10x reads coverage is still a safe data amount for correct estimations (Sun et al., 2018), while Genomescope2 requires a bit more (around 15x according to Ranallo-Benavidez et al., 2020).

To test if the GSE remains consistent even in low coverages conditions, we ran the analysis for subsamples of different number of reads, randomly extracted from the good and bad *O.niloticus* samples already strict-filtered. We chose subsamples of 10, 12 and 15 estimated coverage to simulate a range of coverage close to the declared model limit. To estimate the coverage of subsamples, we use as Nile Tilapia GS the mean value of estimations obtained by Feulgen densitometry (1.05 Gigabases).

We also tested the possibility to pool reads from 2 individuals of the same species to obtain a higher coverage and avoid the issue. We chose 11 species from the dataset with at least two individuals, expressly selecting both species where individuals have a GSE for most of k-mer sizes (enough coverage for the model to

converge) and species where the model failed to converge. For every species we selected two individuals and pooled their reads together using the “cat” command. After that we ran the analysis for both single individuals and pool samples and we compared the results.

2.3. Body size correlation and phylogenetic analysis

2.3.1. Phylogenetic tree

The phylogeny necessary for these analysis was again provided by Nathan Vranken and Hannes Svaardal group based in Antwerp. It is a Maximum Likelihood (ML) species tree built from a matrix of 1.000.000 SNPs through the package IQ-TREE v2.2.0 (Minh et al., 2020) and it includes most of our species plus some riverine species closely related. It was pruned to keep only the species of interest and through the phangorn library it was made ultra-metric, since that is a primary condition for subsequent analysis.

The tree shows that the Lake Victoria Super Flock is monophyletic with a bootstrap support of 100, while within the Lake Victoria region, all species from Lakes Albert and Victoria each form a well-supported radiation. Most species from Lakes Kivu and Edward also form a well-supported clade, except for some species from Lake Edward that seem to have more ancestral positions to all other species from the Lake Victoria region and whose placement is more uncertain. These species include H. squamipinnis, H. aeneocolor, and H. limax

2.3.2. Regression and lambda estimate

Maximum known standard length (SL) was used as proxy of body size for the species for which it is available, while the remanent species were excluded from this analysis. We calculate regression between GS and SL using both phylogenetic and non-phylogenetic approaches. Ordinary Least Squares (OLS) is the phylogeny independent method since it assumes that the points are independent. Phylogenetic Generalized Least Squares (PGLS), instead, uses knowledge of phylogenetic relationships, in this case our ML tree, to produce an estimate of expected covariance in cross-species data. Because of their shared lineage, in this model it is

expected that closely related species have more comparable features and, as a result, yield more similar residuals from the least squares regression line. (Symonds & Blomberg, 2014). Various models were proposed to predict the structure of residuals in PGLS, we tested for the Brownian motion and the Ornstein-Ulhenbeck. Akaike weight was calculated for all three regression models to choose the best one. We used R package phylolm (Ho et al., 2016) to perform this analysis.
Additionally, the phylogenetic signal contained in our data was estimated as Pagel's lambda (Pagel, 1999) through the phylosig function of phytools (Revell, 2012), for both genome size and body size.

2.3.3. Trait evolution and ancestral state reconstruction

We map our GSE on the phylogeny using the contmap function of phytools package (Revell, 2012) to understand how it changed during time across lineages. The mapping is accomplished by reconstructing ancestral states with fastAnc, and then interpolating the states along each edge as maximum likelihood estimates under a Brownian evolutionary process.
Also, for both body size and genome size a traitgram was produced, using the multirateBM function of phytools. It is again a reconstruction based on maximum likelihood and Brownian motion model that shows how the rate of evolution of a particular trait changed across different lineages during time.

2.3.4. Disparity

To explore if most of the variability in the two traits is distributed among subclades or within them and how this changed through time, disparity analysis for both traits was performed. The dtt function of geiger package (Harmon et al., 2008) was used for this analysis.

3. RESULTS

3.1. Benchmarking results (Tab 1)

Contrasting the two packages, Genomescope2, although with a slight underestimation in comparison with Feulgen densitometry estimates, remains more robust (lower SD) through different k-mer sizes for gNT on both not trimmed reads and the two filtering regimes. On the other hand, it fails to converge for every k-mer size in 12 and 10 coverage subsamples and even for k17 and k31 of the bNT strict trim regime (even though it could be argued that the quality of the reads has a role in this), while the rest of the estimates for the two filtering regimes of bNT and for both 15 coverage subsamples are evidently lower than the reference. Low coverage emerges clearly as an issue for this method. Also, Genomescope2 is quite sensitive to slight changes in the initial k-mer coverage value (-l), and since we set it approximately as half the coverage of the homozygous peak is not always easy to standardize it.

FindGSE has higher jumps in estimates between different k-mer sizes (higher SDs), but it returns reliable estimations for both gNT and bNT through the two filtering scenarios (bNT strict k31 excluded as it seems a technical problem). The model also converges for gNT all sub15 k-mer sizes and for the three smallest of sub12, even if with an overestimation of GS in comparison with densitometry results for k27 and k31 of sub15, and k17 and k19 of sub12.

Looking at these data findGSE is our choice for this analysis, because even though of good quality, our dataset have a wide range of coverage including samples with low coverage, and this method seems more robust than Genomescope2 when coverage is lower.

For what concerns the initial quality of the bases in our benchmarking process, it has an impact only at the lower coverages represented by bNT subsamples, where higher k-mer sizes struggle to converge, probably because intensive trimming reduced the mean length of the reads and consequently the amount of input data for the model. bNT seems also to produce in general lower estimates than gNT but this

could be due to individual differences in GS of the fish whose DNA composes the two samples.

We opted for “strict” trimming (-q 30) because it seems to give more consistent estimates (lower SD throughout different k-mer sizes) for both the good and the bad sample, comparing to the “soft” one (-q 20) without sacrificing the reliability.

Regarding mitochondrial filtering, at the end of the process, we noticed that a really low number of reads were filtered out. For example, for gNT only 200.000 reads (100.000 from R1 and 100.000 for R2) on 467 million aligned to the mitochondrial genome and where consequently removed. Moreover, the GSE with and without mitochondrial reads did not change and there is no evidence of high frequency peaks in the histograms of our samples, so we decided to avoid this time-consuming part for the entire dataset of Victoria region individuals.

Since the model tends to fail more frequently using larger k-mers, due to lower coverage, k31 was removed from the analysis. Furthermore, an exploratory analysis on the dataset showed that k17 estimations tend to organize themselves in a bimodal distribution when all samples are considered. This behaviour cannot be observed in the estimates obtained with the other k-mers and it is probably an artifact of findGSE, so we removed k17. This preliminary analyses also highlighted the fact that findGSE estimations tend to grow when k-mer size is increased (Figure 4). To buffer this effect, the final individual GSE is considered the arithmetic mean of the GSE of the four-remaining k-mer sizes (19, 21, 24 and 27) for that individual. Samples that do not produce reliable estimates for all k-mer sizes were discarded.

Sample	k17	k19	k21	k24	k27	k31	Mean	STD DEV	N.of reads(PE)	N.of bases	Estimated coverage
ERR7448119	958.675.930	991.657.858	1.065.663.662	1.092.730.987	1.083.646.502	1.124.325.777	1.052.783.452,67	63.914.163,41	934.987.160	140.248.074.000	133,57
ERR7448119soft	884.020.000	989.430.000	1.062.332.000	1.087.620.000	1.123.856.000	1.169.471.000	1.052.788.166,67	102.437.115,60	803.757.790	119.030.710.498	113,36
ERR7448119strict	960.642.000	981.533.000	1.073.454.000	1.079.336.000	1.108.793.000	1.139.304.000	1.057.177.000,00	71.001.462,74	751.375.440	110.827.877.400	105,55
ERR7448119sub10	-	-	-	-	-	-	-	-	76.000.000	11.210.000.000	10,68
ERR7448119sub12	1.257.860.000	1.288.266.000	1.180.474.000	-	-	-	1.242.200.000,00	55.576.123,61	90.000.000	13.275.000.000	12,64
ERR7448119sub15	1.006.650.000	1.081.466.000	1.160.661.000	1.299.208.000	1.339.291.000	1.388.656.000	1.212.655.333,33	152.876.708,55	110.000.000	16.225.000.000	15,45
SRR071614	1.070.934.757	1.015.131.544	1.027.255.776	1.007.516.995	1.012.645.571	1.020.468.786	1.025.658.904,83	23.189.183,67	275.637.646	27.839.402.246	26,51
SRR071614soft	907.821.000	959.161.000	981.083.000	1.115.645.000	1.038.903.000	1.058.960.000	1.010.262.166,67	75.172.153,75	189.543.366	16.469.302.433	15,69
SRR071614strict	914.082.000	968.152.000	1.017.998.000	1.027.919.000	1.051.738.000	-	995.977.800,00	54.989.832,18	169.917.144	16.991.714.400	16,18
SRR071614sub10	-	-	-	-	-	-	-	-	110.000.000	11.000.000.000	10,48
SRR071614sub12	1.099.336.000	-	-	-	-	-	1.099.336.000,00	-	126.000.000	12.600.000.000	12
SRR071614sub15	919.981.000	1.006.856.000	1.031.973.000	1.043.517.000	-	-	1.000.581.750,00	55.870.962,59	158.000.000	15.800.000.000	15
ERR7448119	924.221.346	941.400.729	950.892.894	960.553.889	969.178.569	978.178.025	954.070.908,67	19.560.435,14	934.987.160	140.248.074.000	133,57
ERR7448119soft	925.206.636	944.148.298	954.483.167	965.485.643	974.599.697	984.759.894	958.113.889,17	21.574.250,30	803.757.790	119.030.710.498	113,36
ERR7448119strict	917.362.537	936.393.899	947.130.706	958.500.244	967.653.672	977.999.522	950.840.096,67	22.010.033,01	751.375.440	110.827.877.400	105,55
ERR7448119sub10	-	-	-	-	-	-	-	-	76.000.000	11.210.000.000	10,68
ERR7448119sub12	-	-	-	-	-	-	-	-	90.000.000	13.275.000.000	12,64
ERR7448119sub15	733.921.649	765.297.347	744.120.379	749.213.336	723.903.482	685.406.539	733.643.788,67	27.480.540,53	110.000.000	16.225.000.000	15,45
SRR071614	868.769.066	925.810.540	924.088.130	887.907.859	872.026.130	858.251.193	889.475.486,33	29.082.369,42	275.637.646	27.839.402.246	26,51
SRR071614soft	731.018.755	752.158.794	744.741.631	704.654.481	656.330.894	573.988.233	693.815.464,67	68.209.451,06	189.543.366	16.469.302.433	15,69
SRR071614strict	-	741.693.246	716.015.040	665.701.507	604.794.883	-	682.051.169,00	60.405.758,71	169.917.144	16.991.714.400	16,18
SRR071614sub10	-	-	-	-	-	-			110.000.000	11.000.000.000	10,48
SRR071614sub12	-	-	-	-	-	-			126.000.000	12.600.000.000	12
SRR071614sub15	648.423.709	677.968.155	649.781.158	594.725.881	528.126.534	400.456.671	583.247.018,00	104.143.391,19	158.000.000	15.800.000.000	15

Tab 1. Benchmarking results of two Nile Tilapia samples, one with good initial reads quality (ERR7448119) and the other with bad initial reads quality (SRR071614). Both samples were analysed raw, after two different filtering designs and subsampled after the stricter filtering in order to show expected coverage values of 10, 12 and 15. Two estimation packages were tested: findGSE (in red) and Genomescope2 (in blue). The pipeline was run for 6 different k-mer sizes (k17, k19, k21, k24, k27, k31) and a mean of their estimates was computed with a standard deviation (STD DEV). The estimated coverage was calculated dividing the number of bases of each sample for 1,05 Gb, the average of the genome size range estimated by Feulgen densitometry in the Animal Genome Size Database (Gregory et al., 2005).

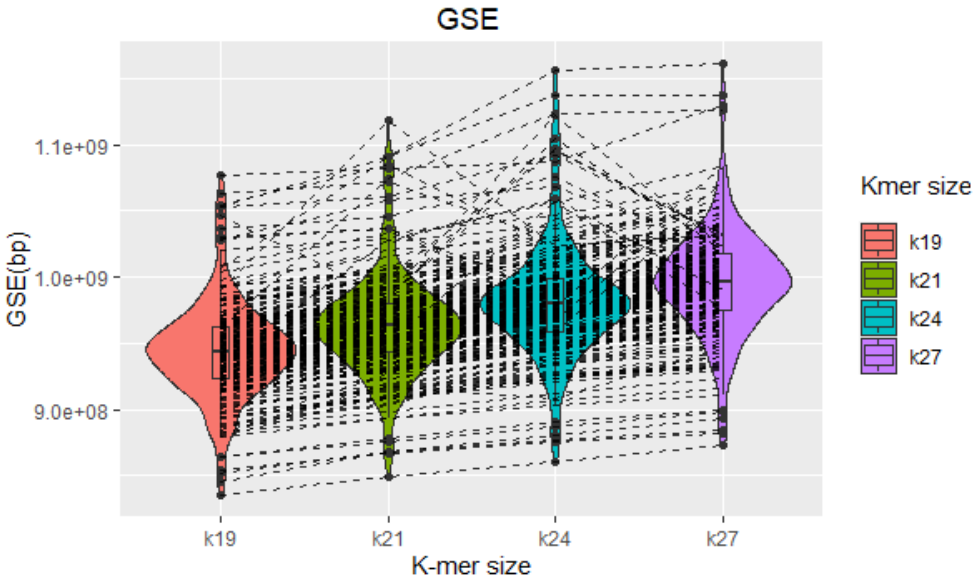


Figure 4. Violin plot comparison of individual genome size estimates for 4 different k-mer sizes (19, 21, 24, 27). Dashed lines connect estimates belonging to the same individual. Sample size 210 individuals.

The pooling test gave us quite interesting results with the mean of pooling sample that fell quite near the mean of the respective individuals (pooling estimates differ by mean of 1,06 ± 2,77 % from respective individuals mean). Moreover, all pooling sample gave a reliable estimate even if one of the samples that compose them fail to converge for one or more k-mer sizes. Despite that, we decided to not adopt this solution because most species do not have more than one individual and we do not know if pooling can possibly introduce some artificial signals in the analysis, so we prefer to keep it standard. Instead, we use the arithmetic mean of species individuals as GSE for that species.

3.2. GSE

3.2.1. Individuals

210 samples belonging to 155 species with a mean of 1.35 individuals per species (range 1-14) arrived at the end of the pipeline with a valid GSE for all four k-mer sizes and were therefore used for the subsequent analysis.

The genome size for our samples ranges from 0,855 Gb to 1,105 Gb with a mean of 0,971 Gb and a standard deviation of 0,039 Gb.

Higher variability in different individual k-mer sizes estimates could be interpreted as a signal of scarce consistency and, therefore, increasing uncertainty of the findGSE model. To explore that we calculated a coefficient of variation (CV) that is simply the individual standard deviation for different k-mer sizes estimates divided by the mean estimate to normalize it. The mean CV in our samples is 0,253, with a maximum of 0,721 and only nine samples over 0,5 (Figure 5).

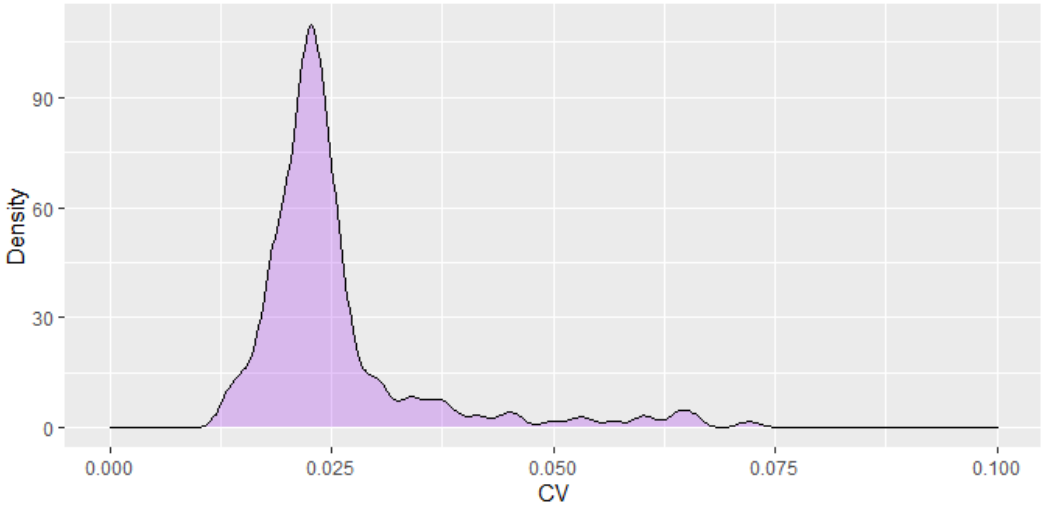


Figure 5. Density plot of the distribution of the coefficient of variation (CV). Sample size 210 individuals.

To test for possible causes of high variability, exploratory regression analyses were performed between CV and both percentage of lost reads during filtering (as proxy of initial data quality) and total number of reads after filtering (as proxy of low

amount of data). Both showed a very weak, if present, negative correlation (Figure 6).

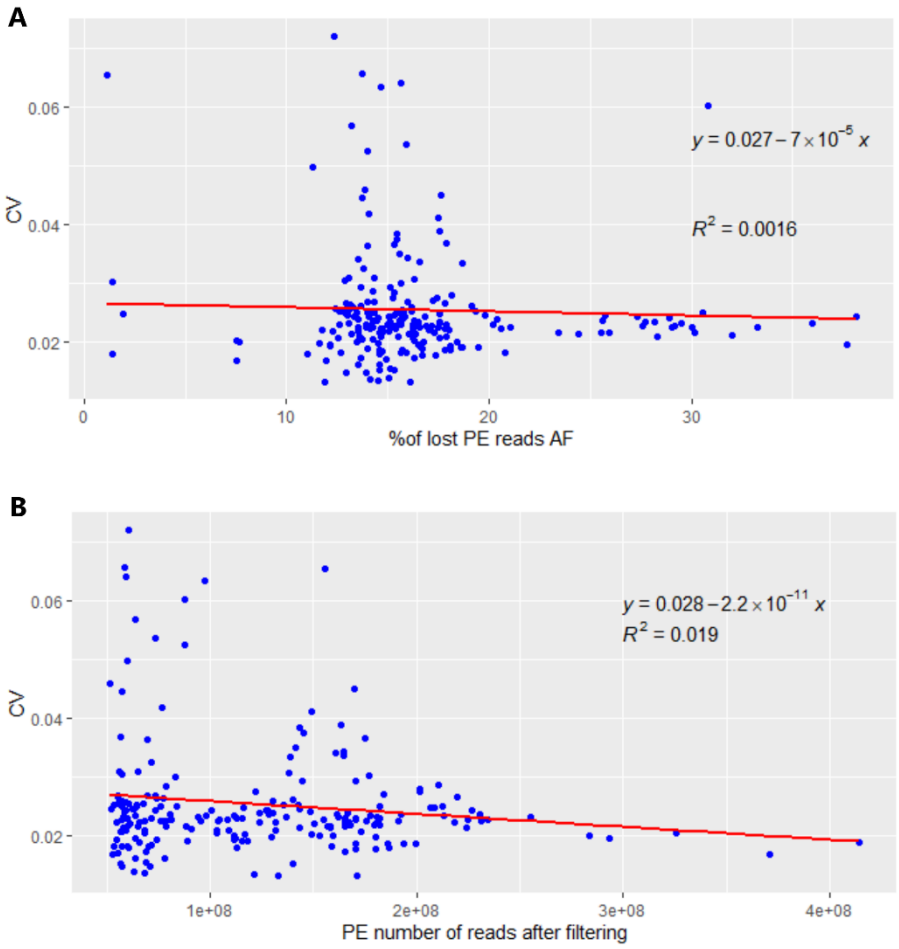


*Figure 6. Regression analysis performed between the coefficient of variation (CV) and, respectively, percentage of lost paired-end reads after filtering (**A**) and number of total paired-end reads after filtering (**B**). Sample size 211 individuals.*

3.2.2. Species

At the species level, our samples show a mean GS of 0,970 Gb with a standard deviation of 0,035 Gb. The range goes from the 0,870 Gb of *H. perrieri* to the 1,094 Gb of *H. sp_sky_blue_picker* (Figure 7). 34 species have more than one individual and the mean within species standard variation is 0,022 Gb.

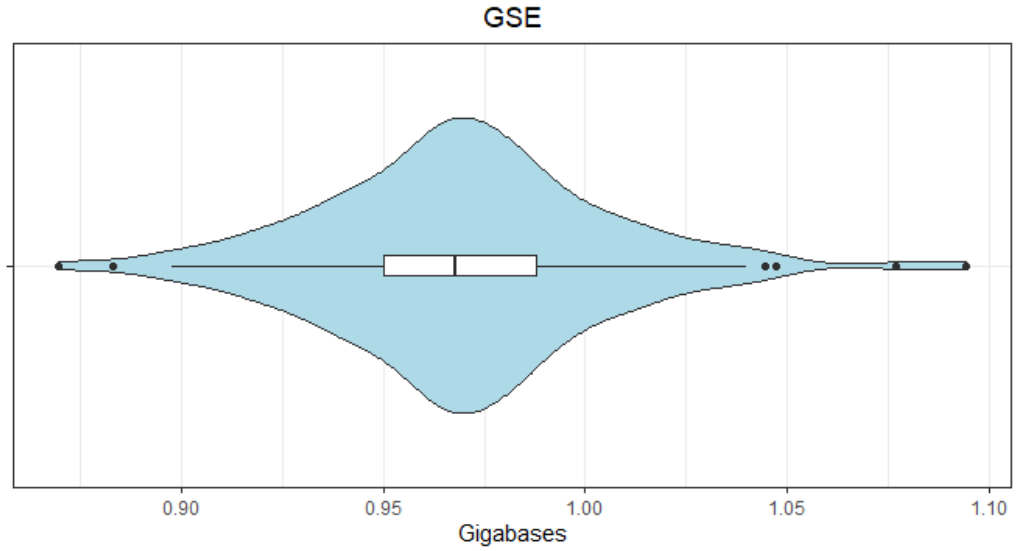


Figure 7. Violin plot and box plot of genome size (Gb) species distribution. Sample size 155 species.

The 155 remaining species are grouped in this way: 9 are from Lake Albert radiation, 32 from Lake Edward radiation, 5 from Lake Kivu radiation, 104 from Victoria radiation and 5 belong to the upper Nile lineage. A non-parametric Kruskal-Wallis test was performed using a Bonferroni correction for multiple comparisons, to test if fish from the same lakes/groups are more similar to each other than to fish from other lakes, but the p-value was not significant (Kruskal-Wallis chi-squared = 6.4226, df = 4, p-value = 0.17).

Pairwise comparison between lakes was also tested through a Dunn's test, but none of them results significant (Tab 2).

Comparison of x by group
(Bonferroni)

Col Mean- Row Mean	Albert	Edward	Kivu	Upper Ni
Edward	0.934228 1.0000			
Kivu	0.537854 1.0000	-0.109154 1.0000		
Upper Ni	-0.077216 1.0000	-0.822569 1.0000	-0.542441 1.0000	
Victoria	1.818242 0.3451	1.381483 0.8357	0.724624 1.0000	1.473951 0.7025

alpha = 0.05
Reject Ho if p <= alpha/2

Tab 2. Dunn's test for genome size between 5 groups (Albert, Edward, Victoria, Upper Nile, Kivu) with p-value corrected with Bonferroni method. Sample size 155 species (9 Albert, 32 Edward, 104 Victoria, 5 Upper Nile, 5 Kivu).

3.3. Standard length

120 of the remaining species have a SL estimation. The mean is 749,3 mm with a standard deviation of 537,1 mm, quite high but coherent with the evident bimodal distribution of the data (Figure 8).

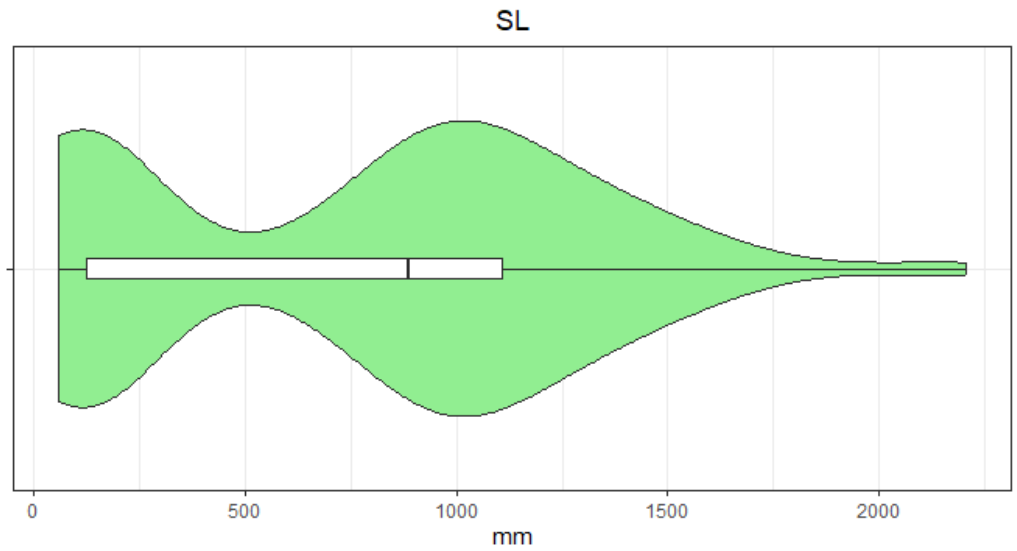


Figure 8. Violin plot and box plot of maximum standard length (mm) species distribution. Sample size 120 species.

The 120 species with an actual body size value are divided in this way: 9 belong from Lake Albert radiation, 32 from Edward radiation, 5 from Kivu radiation, 4 from Upper Nile lineage and 70 from Victoria radiation.

The Kruskal-Wallis test gave a significant p-value (Kruskal-Wallis chi-squared = 13.829, df = 4, p-value = 0.007863), evidence that variation among groups is greater than within groups.

Also, Dunn's test showed a significant difference (p-value = 0,0195) in SL between fish of Lake Edward radiation and the ones from Victoria radiation (Tab 3).

Comparison of x by group
(Bonferroni)

Col Mean- Row Mean	Albert	Edward	Kivu	Upper Ni
Edward	-0.021429 1.0000			
Kivu	1.922501 0.2727	2.246704 0.1233		
Upper Ni	-0.448504 1.0000	-0.492960 1.0000	-2.000293 0.2273	
Victoria	1.716469 0.4304	2.886305 0.0195*	-1.003421 1.0000	1.706604 0.4395

alpha = 0.05
Reject Ho if p <= alpha/2

Tab 3. Dunn's test for maximum standard length between 5 groups (Albert, Edward, Victoria, Upper Nile, Kivu) with p-value corrected with Bonferroni method. Sample size 120 species (9 Albert, 32 Edward, 70 Victoria, 4 Upper Nile, 5 Kivu).

3.4. Phylogenetic analysis

11 of the 155 initial species were not present in the ML phylogeny and other 35 did not have an actual maximum SL estimate, so these analyses were run with 104 species in total.

3.4.1. Trait mapping and ancestral reconstruction

Looking at the trait mapped on the tree (Figure 9) suggests that a phylogenetic signal for GS seems not to be present, since different values of this trait are common in close related species without a recognisable pattern. The same could be claim for

body size, except for Lake Albert species (in yellow in Figure 9) that show a similar, relatively big maximum standard length. Green and yellow colours are predominant on the tree, coherent with the data distribution that show most of the samples around the mean values.

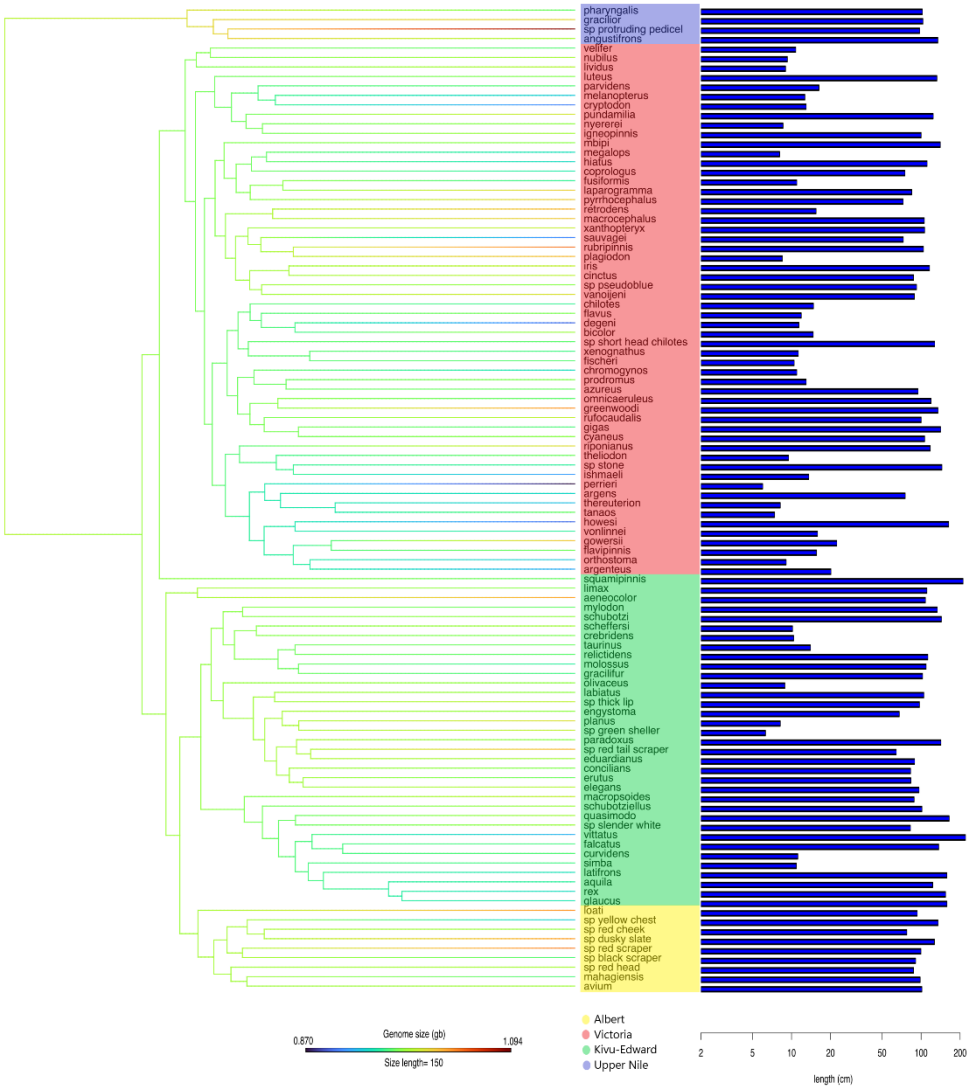


Figure 9. Tree mapping of the genome size trait (Gb) for 104 species of the dataset. Ancestral state reconstruction and branch evolution based on a Brownian motion model. Maximum standard length is also display for each individual and different colours indicate the belonging group of the species.

3.4.2. Regression

Body size shows a weak positive and significant correlation with genome size for the Ordinal Least Squares model only (p-value = 0.02178), while the two phylogenetic dependent models did not find any significant relationship (p-value PGLS-BM = 0.2206, p-value PGLS-OU = 0.07617) (Figure 10 and tab 4).

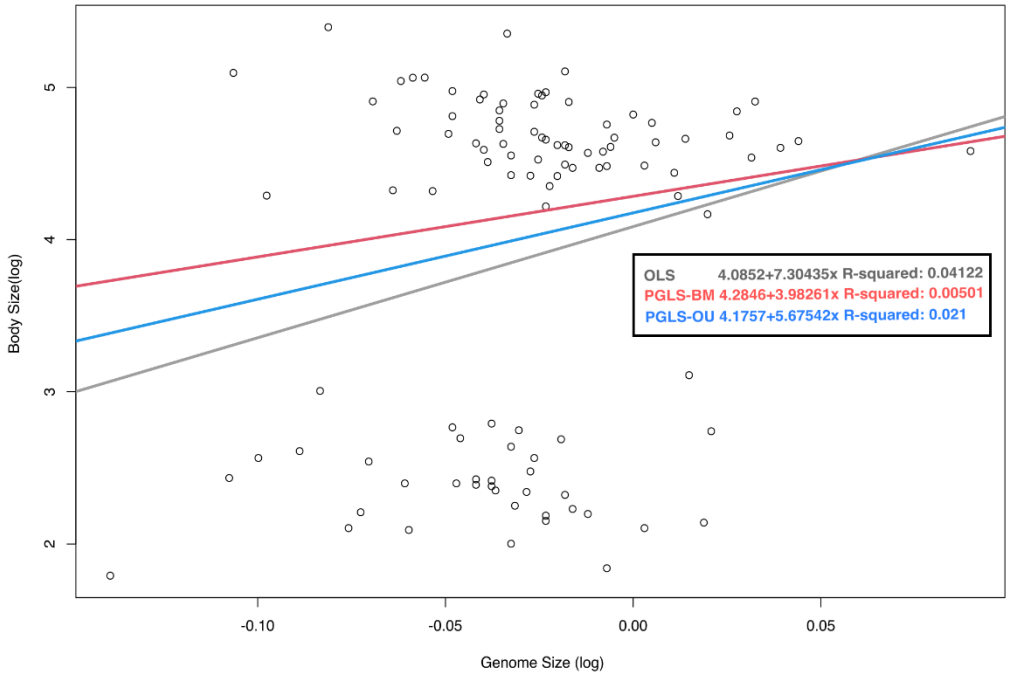


Figure 10. Regression analysis between body size (log) and genome size (log) for 104 species. Grey line represents regression computed with Ordinary Least Squares model (OLS), blue and red lines the regressions computed with Phylogenetic Generalized Least Square model (PGLS) respectively using Brownian motion (PGLS-BM) and Ornstein-Uhlenbeck (PGLS-OU) evolutionary model. Equation and adjusted R-squared are plotted.

OLS					PGLS-BM					PGLS-OU				
Coefficients:					Coefficients:					Coefficients:				
	Estimate	StdErr	t.value	p.value		Estimate	StdErr	t.value	p.value		Estimate	StdErr	t.value	p.value
(Intercept)	4.08552	0.14027	29.1255	< 2e-16 ***	(Intercept)	4.28462	0.58073	7.3780	4.425e-11 ***	(Intercept)	4.17571	0.36626	11.4008	< 2e-16 ***
log(gs)	7.30435	3.13513	2.3298	0.02178 *	log(gs)	3.98261	3.23141	1.2325	0.2206	log(gs)	5.67542	3.16787	1.7916	0.07617 .
---					---					---				
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
R-squared: 0.05053					R-squared: 0.01467					R-squared: 0.03051				
Adjusted R-squared: 0.04122					Adjusted R-squared: 0.00501					Adjusted R-squared: 0.021				
AIC	logLik				AIC	logLik				AIC	logLik			
320.3	-157.2				317.3	-155.7				318.6	-155.3			

Tab 4. Statistics regarding the regression analysis of figure 10.

Observing Akaike's weights (Tab 5), OLS has also the highest likelihood of being the best model among the set being considered for both GS (0.7043) and body size (0.72866), suggesting weak or absent phylogenetic signal in the data.

genome size						body size					
-- Akaike weights --						-- Akaike weights --					
	Rank	AIC	diff	wi	AICw		Rank	AIC	diff	wi	AICw
Ols	1	574	0.00	1.0000	0.7043	Ols	1	1130	0.0	1.00000	0.72866
OU1 2	2	572	1.97	0.3731	0.2628	OU1 2	2	1132	2.0	0.36781	0.26801
BM1	3	568	6.13	0.0466	0.0328	BM1	3	1141	10.8	0.00457	0.00333

Tab 5. Akaike weights (AICw) statistics for the three models adopted in figure xx regression for both genome size and body size. The higher the weight, the better the model.

Surprisingly lambda estimation for GS is 1,43 and strongly significant (p-value = 0,000531522), a value that indicates instead a phylogenetic effect stronger than the one predicted by a simple Brownian motion model. Body size phylogenetic signal is clearly lower (0,68) and not significant (p-value = 0.0102317) (Tab 6).

Body size		Genome size
Phylogenetic signal lambda : 0.688857		Phylogenetic signal lambda : 1.43182
logL(lambda) : -559.632		logL(lambda) : 211.083
LR(lambda=0) : 6.5941		LR(lambda=0) : 12.0017
P-value (based on LR test) : 0.0102317		P-value (based on LR test) : 0.000531522

Tab 6. Phylogenetic signal lambda statistics for genome size and body size. P-value significant < 0,05. Sample size 104 species.

3.4.3. Traitgram

Looking at the traitgram of GS through time (Figure 11) it can be noticed that, for most of the branches, the trait shows a low rate of evolution around average values that results in an overlapped topology where all the branches are placed close to each other and end with similar value at the tips. Interestingly the lineage with higher rate of evolution and that comprehend the higher estimations for the data, is

the one formed by Upper Nile species (first branch to separate from the common ancestor in the figure). The ancestral reconstructed trait appears similar to the middle point of the actual distribution.

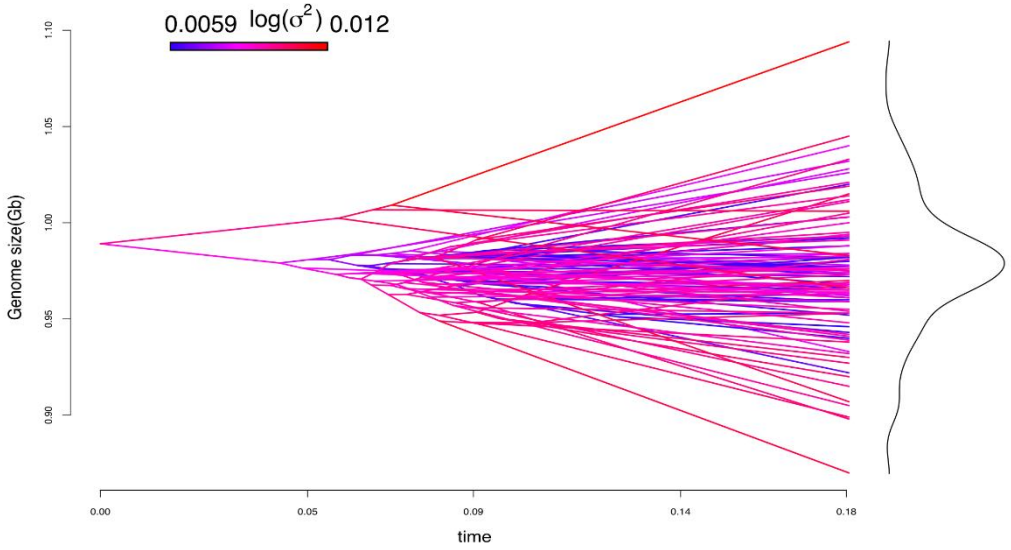


Figure 11. Traitogram showing evolution through time and across lineages of genome size (Gb). Rate of evolution (log (δ^2)) variation is displayed through a colour palette. X axis represents relative proportion of time passed from the common ancestor of the species. Sample size 104 species. Ancestral state reconstruction and branch evolution used a Brownian motion model. On the right the distribution of the dataset is plotted.

Body size, on the other hand, shows high rates of evolution for most of the lineages with a clear tendency to evolve either smaller or larger body sizes, but not intermediate ones. The Upper Nile lineage is characterized by a low rate of evolution showing an almost horizontal branch, with species of relatively big body size that stayed almost the same during time. The plotted distribution again clearly shows a bimodal nature (Figure 12).

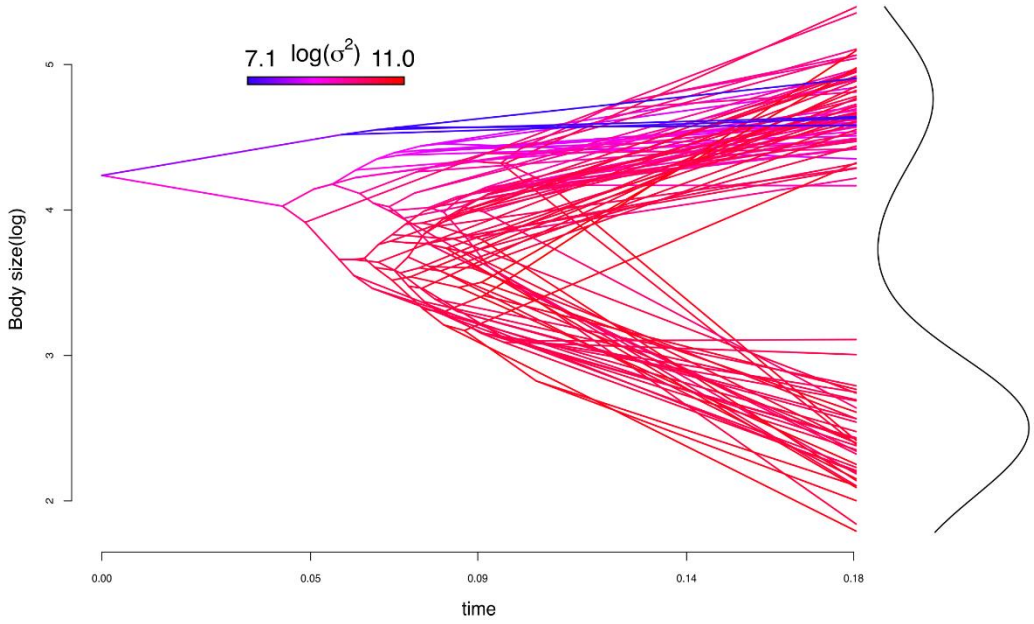


Figure 12. Traitgram showing evolution through time and across lineages of body size (log). Rate of evolution (log (δ^2)) variation is displayed through a colour palette. X axis represents relative proportion of time passed from the common ancestor of the species. Sample size 104 species. Ancestral state reconstruction and branch evolution used a Brownian motion model. On the right the distribution of the dataset is plotted.

3.4.4. Disparity

Average subclade GS disparity shows the highest values at the beginning of the radiation, then decreases gradually and constantly at the beginning and then in a more stepwise manner, until the value of 0 in most recent times. This means that, during evolution, subclades become more and more similar in genome size. The trait seems to not follow at all the trajectory of an unconstrained Brownian model trait evolution (dashed line) (Figure 13).

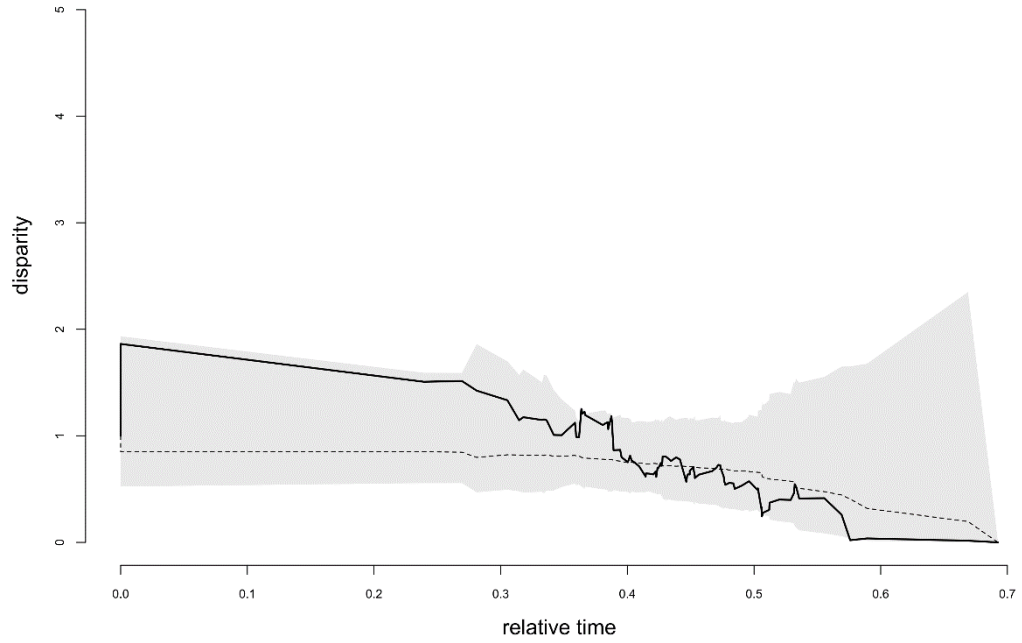


Figure 13. Average subclade disparity through time for the genome size trait (full line). X axis represents the relative proportion of time passed from the common ancestor of the species. Dashed line shows the unconstrained Brownian motion model trait evolution simulation. Sample size 104 species.

Regarding body size disparity (Figure 14), it shows low levels for most of the radiation, except for a relative recent peak where it is higher than the one predicted by Brownian motion. This more recent increase in subclade diversity is probably due to the first bifurcating lineages of the traitgram (Figure 12) that start to lead to the two different body size conditions (big and small).

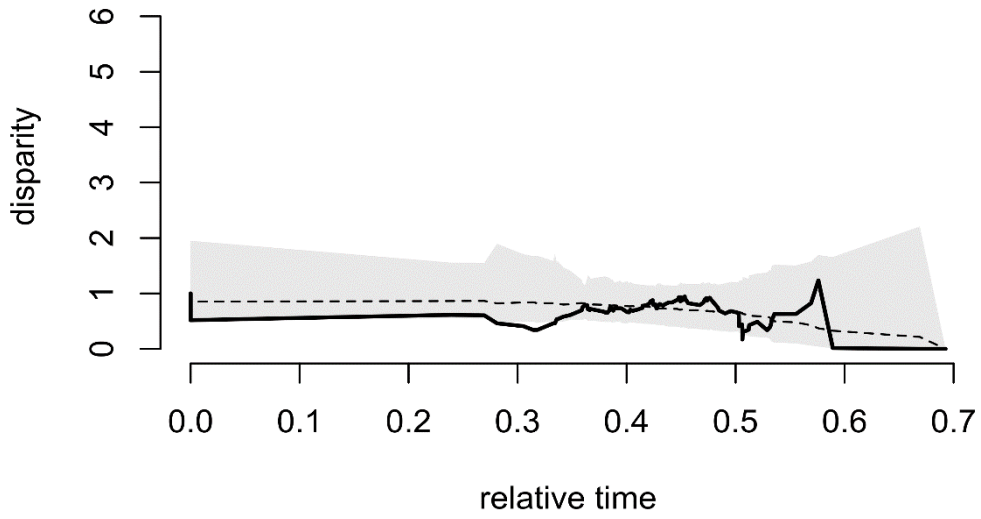


Figure 14. Average subclade disparity through time for the body size trait (full line). X axis represents the relative proportion of time passed from the common ancestor of the species. Dashed line shows the unconstrained Brownian motion model trait evolution simulation. Sample size 104 species.

4. DISCUSSION

The main purpose of this study was to estimate the interspecific variation of genome size across the Lake Victoria Super Flock through a specific-constructed bioinformatics pipeline, and to use those data to investigate the evolution of the trait and the possible relationship with body size, in this case considered as the maximum known standard length.

The results show that most interspecific variation in genome size is included in a narrow range of 1,005–0,935 Gb. Cichlids of Victoria region appear therefore similar for this trait, apart for a small number of species that occupy the extremes of the distribution, which spans 0,224 Gb. Even among lake radiations and groups, the differences in genome size are statistically comparable to the ones registered within them. Even if not statistically different, species of the Upper Nile lineage show the highest genome size (*H.sp_protruding_pedicel*) and in general relatively high values for the trait. In light of the results of Meier et al. (2017), Upper Nile lineage was one of the two hybridizing clades at the origin of the LVSF radiation, therefore, the null hypothesis is that the newly formed hybrids genome size should have somewhat values in the middle between the ones of this lineage and the ones of the other basal clade, the Congolese lineage. Departures from this value can be related with different mutational mechanisms effects as for example transposable elements expansions. Including specimens from Congolese lineage in future analysis will help us to study these dynamics.

Ancestral reconstruction methods also support the fact that genome size remained quite stable during the diversification of these groups, with low evolutionary rates registered for most of the lineages actually present in the Victoria region lakes, and an ancestral trait estimate a few megabases higher than the mean value of the current species. Clearly, interpreting these results, it should be taken into account the fact that Brownian motion reconstruction is considered the null hypothesis for trait evolution, while different evolutionary trajectories could have been shaped the current interspecific GS variation. Phylogenetic analyses seem to suggest the fact that Brownian motion is not the best way to describe genome size evolution. Although a pattern in genome size is not recognizable when the trait is mapped on

the tree, with relatively higher or lower estimations in closely related species, the estimated Pagel's lambda for genome size is 1,43. This is evidence of a strong phylogenetic signal in the data, stronger than the one predicted by a simple Brownian motion model. One possible hypothesis is that genome size could have evolved under stabilizing selection in this clade of cichlids. Stabilizing selection is a particular type of natural selection that favours an average phenotype selecting against extreme variation (Schmalhausen, 1949). The data distribution and the low rates of evolution observed across lineages are compatible with this scenario. Still, more analyses, and probably an optimality approach, are needed to confirm that and to understand what kind of underlying constraints or selective pressures are possibly limiting genome size variation in this group. Interestingly the highest rates of evolution for our data were registered within species of the Upper Nile lineage, phylogenetically older than the LVSF radiation, and therefore maybe not be affected by constraints. It is also possible that those limitations in genome size variation arose due to the hybridization event, since Runemark et al. (2018) observed that both genome and organismal function can constrain hybrid genome formation and hence its size. Future analysis will focus on including more species of the basal clades and from other riverine groups placed at the base of the radiation to investigate these hypotheses.

The trend of average subclade disparity is somewhat in line with this assumption, showing a constant decrease through time. This means that subclades become more and more similar between them regarding genome size variation, and this could be explained surely by great lineage diversification and consequent overlapping among them, but also by their possible converging towards an average optimum value. Again, the observed disparity emerges quite different to the one predicted by Brownian motion simulation.

Regarding the relationship between genome size and body size, what we found is a weak but significant positive correlation between the two variables only for the phylogenetic independent Ordinary Least Squares method, that is also the best model between the three proposed according to Akaike's weight. The positive correlation, although weak, found in this study, contradicts the general negative relationships that was found by Smith & Gregory (2019) in teleosts, but more information should be collected to understand this discrepancy. Even though

phylogenetic dependent method (PGLS) failed to return significant associations, the effect of phylogeny on this relationship cannot be ruled out. First, the evolutionary models used for PGLS analysis could be inappropriate to capture the complex pattern of trait evolution of genome size and body size in our species. The high phylogenetic signal found for genome size and the observed inadequacy of Brownian motion modelling, seems to confirm this assumption. Moreover, PGLS regression method relies on precise and unbiased trait value, so OLS regression could simply be more robust to such measurement errors and, therefore, yield better model fit.

Intriguingly, preliminary analysis limited to species of Lake Kivu, Edward and Albert (KEA region) showed a significant negative correlation between the two traits for all the three methods. One possible explanation can be that the exceptional ecological and morphological diversification that faced the younger Lake Victoria species altered the ancestral link between the two traits, still present in older and smaller lake radiations. An interesting future development could be comparing this regression results with those computed using data from Lake Malawi or Tanganyika radiation, whose species faced a wide ecological and morphological adaptation, like the Victoria one.

Focusing on body size, our analysis has highlighted some unexpected but interesting patterns. As shown by the density plot, our species body size tends to show a bimodal-like distribution, with most of the samples grouped near two peaks and a final tail composed by a few high maximum standard-length species. These fish species apparently evolved either a larger size or a smaller one, but rarely show the middle values. Moreover, looking at the traitgram, it appears evident that most of the LVSF lineages bifurcated to alternate body size peaks during their diversification, driven by high rates of evolution. On the contrary, species from Upper Nile lineage remained similar in body size through time and show the lowest rates of evolution for this trait. The observed pattern suggests that disruptive selection could be acting on body size, since it is an evolutionary force that tends to favour extreme phenotypes, while selecting against intermediate ones. Thoday (1972) claims that such selection may be expected in two contrasting types of situations. First the two or more optimal phenotypes may depend on one another, second the optima may be set by heterogeneity of the environment such as a mosaic

of ecological niches or a clinal situation. More ecological and physiological information are needed to confirm one or both of those scenarios, but the latter seems quite appropriate to describe the situation found by cichlids when they colonized Lake Victoria (Seehausen, 2015).

For body size a difference between inter- and intragroup variation was found significant, and Dunn's test highlighted a significant difference between Edward and Victoria radiation species for the trait. Our initial hypothesis was that this difference was linked to the higher diversification of Victoria species that should result in a higher within group variation. Surprisingly standard deviations for body sizes of species from Edward radiation and species from Victoria radiation are very similar (499,9 mm for Edward, 508,8 mm for Victoria) but Edward species show a mean body size around 400 mm bigger (993,1 mm against 600,5 mm of Victoria). A first literature exploration did not find an explanation for this discrepancy, so, if adding more Victoria species will confirm this trend, future analysis will be needed to tackle this evolutionary question.

It is also noticeable that Lake Albert species tend to show similar and relatively large body sizes (mean=1012,89 mm), while the other groups are characterized by both large and small species. In this lake, haplochromines are not the dominant clade and are mostly prey in a fish community dominated by predatory fish species such as *Lates*, *Hydrocynus* and *Bagrus* (Wandera et al., 2010). The evolution of bigger sizes could in this scenario be favoured to escape the high predatory pressures from species of other lineages.

4.1. Considerations on the method

The reliability of the method is clearly a fundamental part of these conjectures, since it is possible that the variation found is more technical than biological. The benchmarking process showed that our pipeline could give quite different results for the same sample under different coverage condition with gNT estimate that raised up of around 0,2 Gb when subsampled at 15x estimated coverage. Even considered that for those analysis the high k31 estimate was still considered, it is still legitimate to ask if it is appropriate to use a method whose error is comparable

to the range of the interspecific variation of the samples. Still, even with the probable presence of some noise, the signal in our data seems strong. The mean of the standard deviation of individual estimates from different k-mer sizes (0,024 Gb) and the mean standard deviation within species (0,022 Gb) are both lower than the interspecific standard deviation (0,035) and clearly lower than the whole range of variation (0,224 Gb). If results were driven by technical noise, a wider within species standard deviation would be expected. Also, considering the same underlying distribution, a standard deviation calculated for few individuals, as for most of our species, is more likely higher than if calculated for a dataset of 155 species. Looking at this data the signal-to-noise ratio seems, therefore, favourable.
Other two grey zones of our method are samples with high variation between different k-mer sizes estimates and samples with low coverage. The former ones are signal of inconsistency and show only a really weak, if present, correlation with low number and low starting quality of reads. For the latter ones instead, we are out of our benchmarking safety zone, since they have reliable estimates for all four k-mer sizes, but coverages lower than the ones tested with Nile Tilapia subsamples and for which the model failed. Both these typologies of samples were kept in the analysis. Removing them in future analyses and confronting the results, will allow us to assess their effective impact on the conclusions.

Bibliography

Aguillon SM, Dodge TO, Preising GA, Schumer M. Introgression. *Curr Biol*. 2022;32(16):R865-R868. doi:10.1016/j.cub.2022.07.004

Ai B, Wang ZS, Ge S. Genome size is not correlated with effective population size in the oryza species. *Evolution (N Y)*. 2012;66(10):3302-3310. doi:10.1111/j.1558-5646.2012.01674.x

Arkhipova IR. Neutral theory, transposable elements, and eukaryotic genome evolution. *Mol Biol Evol*. 2018;35(6):1332-1337. doi:10.1093/molbev/msy083

Arnqvist G, Sayadi A, Immonen E, et al. Genome size correlates with reproductive fitness in seed beetles. *Proc R Soc B Biol Sci*. 2015;282(1815):11-14. doi:10.1098/rspb.2015.1421

Beaulieu JM, Leitch IJ, Patel S, Pendharkar A, Knight CA. Genome size is a strong predictor of cell size and stomatal density in angiosperms. *New Phytol*. 2008;179(4):975-986. doi:10.1111/j.1469-8137.2008.02528.x

Bennett MD, Leitch IJ, Price HJ, Johnston JS. Comparisons with Caenorhabditis (~100 Mb) and Drosophila (~175 Mb) using flow cytometry show genome size in Arabidopsis to be ~157 Mb and thus ~25% larger than the Arabidopsis genome initiative estimate of ~125 Mb. *Ann Bot*. 2003;91(5):547-557. doi:10.1093/aob/mcg057

Bensasson D, Petrov DA, Zhang DX, Hartl DL, Hewitt GM. Genomic gigantism: DNA loss is slow in mountain grasshoppers. *Mol Biol Evol*. 2001;18(2):246-253. doi:10.1093/oxfordjournals.molbev.a003798

Bezault E, Mwaiko S, Seehausen O. Population genomic tests of models of adaptive radiation in Lake Victoria region cichlid fish. *Evolution (N Y)*. 2011;65(12):3381-3397. doi:10.1111/j.1558-5646.2011.01417.x

Bilinski P, Albert PS, Berg JJ, et al. Parallel altitudinal clines reveal trends in adaptive evolution of genome size in Zea mays. *PLoS Genet*. 2018;14(5):1-19. doi:10.1371/journal.pgen.1007162

Birch MC, Keenlyside JJ. Tapping behavior is a rhythmic communication in the death-watch beetle, Xestobium rufovillosum (Coleoptera: Anobiidae). *J Insect Behav*. 1991;4(2):257-263. doi:10.1007/BF01054618

Blommaert J. Genome size evolution: towards new model systems for old questions. *Proc R Soc B Biol Sci*. 2020;287(1933). doi:10.1098/rspb.2020.1441

Brawand D, Wagner CE, Li YI, et al. The genomic substrate for adaptive radiation in African cichlid fish. *Nature*. 2015;513(7518):375-381. doi:10.1038/nature13726

Chalopin D, Naville M, Plard F, Galiana D, Volff JN. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biol Evol*. 2015;7(2):567-580. doi:10.1093/gbe/evv005

Charlesworth B (1996) The changing sizes of genes. Nature 384 (6607): 315–316.

Chenuil A, Sauc T, Hemery LG, et al. Understanding processes at the origin of species flocks with a focus on the marine Antarctic fauna. 2018;93:481-504. doi:10.1111/brv.12354

Chung J, Lee JH, Arumuganathan K, Graef GL, Specht JE. Relationships between nuclear DNA content and seed and leaf size in soybean. *Theor Appl Genet*. 1998;96(8):1064-1068. doi:10.1007/s001220050840

Compeau PEC, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol*. 2011;29(11):987-991. doi:10.1038/nbt.2023

Danley PD, Husemann M, Ding B, DiPietro LM, Beverly EJ, Peppe DJ. The Impact of the Geologic History and Paleoclimate on the Diversification of East African Cichlids. *Int J Evol Biol*. 2012;2012:1-20. doi:10.1155/2012/574851

Dion-Côté AM, Renaut S, Normandeau E, Bernatchez L. RNA-seq reveals transcriptomic shock involving transposable elements reactivation in hybrids of young lake whitefish species. *Mol Biol Evol*. 2014;31(5):1188-1199. doi:10.1093/molbev/msu069

Doležel J, Bartoš J. Plant DNA flow cytometry and estimation of nuclear genome size. *Ann Bot*. 2005;95(1):99-110. doi:10.1093/aob/mci005

Doležel J, Greilhuber J. Nuclear genome size: Are we getting closer? *Cytom Part A*. 2010;77(7):635-642. doi:10.1002/cyto.a.20915

Doležel J, Greilhuber J, Suda J. Estimation of nuclear DNA content in plants using flow cytometry. *Nat Protoc*. 2007;2(9):2233-2244. doi:10.1038/nprot.2007.310

Elliott TA, Gregory TR. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philos Trans R Soc B Biol Sci*. 2015;370(1678). doi:10.1098/rstb.2014.0331

Elmer KR, Reggio C, Wirth T, Verheyen E, Salzburger W, Meyer A. Pleistocene desiccation in East Africa bottlenecked but did not extirpate the adaptive radiation of Lake Victoria haplochromine cichlid fishes. *Proc Natl Acad Sci U S A*. 2009;106(32):13404-13409. doi:10.1073/pnas.0902299106

Gante HF, Matschiner M, Malmstrøm M, Jakobsen KS, Jentoft S, Salzburger W. Genomics of speciation and introgression in Princess cichlid fishes from Lake Tanganyika. *Mol Ecol*. 2016;25(24):6143-6161. doi:10.1111/mec.13767

Gregory TR. Genome size and developmental complexity. *Genetica*. 2002;115(1):131-146. doi:10.1023/A:1016032400147

Gregory TR. Genome size and developmental parameters in the homeothermic vertebrates. *Genome*. 2002;45(5):833-838. doi:10.1139/g02-050

Gregory TR, Hebert PDN, Kolasa J. Evolutionary implications of the relationship between genome size and body size in flatworms and copepods. *Heredity (Edinb)*. 2000;84(2):201-208. doi:10.1046/j.1365-2540.2000.00661.x

Gregory TR. 2005 Animal genome size database. http://www.genomesize.com

Gregory TR, Nicol JA, Tamm H, et al. Eukaryotic genome size databases. *Nucleic Acids Res*. 2007;35(SUPPL. 1):332-338. doi:10.1093/nar/gkl828

Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W. GEIGER: Investigating evolutionary radiations. *Bioinformatics*. 2008;24(1):129-131. doi:10.1093/bioinformatics/btm538

Ho, L. S. T., Ane, C., Lachlan, R., Tarpinian, K., Feldman, R., Yu, Q., ... & Ho, M. L. S. T. (2016). Package 'phylolm'. *See http://cran. r-project. org/web/packages/phylolm/index. Html.*

Irisarri I, Singh P, Koblmüller S, et al. Phylogenomics uncovers early hybridization and adaptive loci shaping the radiation of Lake Tanganyika cichlid fishes. *Nat Commun*. 2018;9(1). doi:10.1038/s41467-018-05479-9

Jeffery NW, Hultgren K, Chak STC, Gregory TR, Rubenstein DR, Katju V. Patterns of genome size variation in snapping shrimp. *Genome*. 2016;59(6):393-402. doi:10.1139/gen-2015-0206

Kapusta A, Suh A, Feschotte C. Dynamics of genome size evolution in birds and mammals. *Proc Natl Acad Sci U S A*. 2017;114(8):E1460-E1469. doi:10.1073/pnas.1616702114

Keller I, Wagner CE, Greuter L, et al. Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. *Mol Ecol*. 2013;22(11):2848-2863. doi:10.1111/mec.12083

Kokot M, Dlugosz M, Deorowicz S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*. 2017;33(17):2759-2761. doi:10.1093/bioinformatics/btx304

Leushkin E V., Bazykin GA, Kondrashov AS. Strong mutational bias toward deletions in the Drosophila melanogaster genome is compensated by selection. *Genome Biol Evol*. 2013;5(3):514-524. doi:10.1093/gbe/evt021

Li X. Why are de Brujin graphs useful for genome assembly? *Physiol Behav*. 2016;176(3):139-148. doi:10.1038/nbt.2023.Why

Lynch M, Conery JS. The Origins of Genome Complexity. *Science (80-)*. 2003;302(5649):1401-1404. doi:10.1126/science.1089370

M. Pagel. Inferring the historical patterns of biological evolution. *Nature*. 1999;401(October):877-884.

Maddison W, Knowles L. Inferring phylogeny despite incomplete lineage sorting. *Syst Biol*. 2006;55(1):21-30. doi:10.1080/10635150500354928

Malinsky M, Svardal H, Tyers AM, et al. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nat Ecol Evol*. 2018;2(12):1940-1955. doi:10.1038/s41559-018-0717-x

Mallet J. Hybrid speciation. *Nature*. 2007;446(7133):279-283. doi:10.1038/nature05706

Marburger S, Alexandrou MA, Taggart JB, et al. Whole genome duplication and transposable element proliferation drive genome expansion in corydoradinae catfishes. *Proc R Soc B Biol Sci*. 2018;285(1872). doi:10.1098/rspb.2017.2732

Marescalchi O, Scali V, Zuccotti M. Flow-cytometric analyses of intraspecific genome size variations in Bacillus atticus (insecta, phasmatodea). *Genome*. 1998;41(5):629-635. doi:10.1139/gen-41-5-629

Meier JI, Marques DA, Mwaiko S, Wagner CE, Excoffier L, Seehausen O. Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nat Commun*. 2017;8(May 2016):1-11. doi:10.1038/ncomms14363

Meyer A. Explaining Exuberant Diversification. *Science (80-)*. 2001;294(5540):64-65. doi:10.1126/science.1062185

Neiman M, Paczesniak D, Soper DM, Baldwin AT, Hehman G. Wide variation in ploidy level and genome size in a new zealand freshwater snail with coexisting sexual and asexual lineages. *Evolution (N Y)*. 2011;65(11):3202-3216. doi:10.1111/j.1558-5646.2011.01360.x

Nichols P, Genner MJ, van Oosterhout C, et al. Secondary contact seeds phenotypic novelty in cichlid fishes. *Proc R Soc B Biol Sci*. 2014;282(1798). doi:10.1098/rspb.2014.2272

Paul Decena-Segarra L, Bizjak-Mali L, Kladnik A, Sessions SK, Rovito SM. Miniaturization, Genome Size, and Biological Size in a Diverse Clade of Salamanders. doi:10.5061/dryad.ht76hdrcg

Petrov 2001. *Opinion Relationships between Nuclear DNA Content and Seed and Leaf Size in Soybean*. http://tig.trends.com.

Petrov DA, Lozovskaya ER, Harti DL. High intrinsic rate of DNA loss in Drosophila. *Nature*. 1996;384(6607):346-349. doi:10.1038/384346a0

Petrov D. *Opinion Relationships between Nuclear DNA Content and Seed and Leaf Size in Soybean*. http://tig.trends.com.

Pflug JM, Holmes VR, Burrus C, Spencer Johnston J, Maddison DR. Measuring genome sizes using read-depth, k-mers, and flow cytometry: Methodological comparisons in beetles (Coleoptera). *G3 Genes, Genomes, Genet*. 2020;10(9):3047-3060. doi:10.1534/g3.120.401028

Rayburn AL, Auger JA. Genome size variation in Zea mays ssp. mays adapted to different altitudes. *Theor Appl Genet*. 1990;79(4):470-474. doi:10.1007/BF00226155

Revell LJ. phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol*. 2012;3(2):217-223. doi:10.1111/j.2041-210X.2011.00169.x

Romero-Soriano V, Burlet N, Vela D, Fontdevila A, Vieira C, Guerreiro MPG. Drosophila females undergo genome expansion after interspecific hybridization. *Genome Biol Evol*. 2016;8(3):556-561. doi:10.1093/gbe/evw024

Ronco F, Matschiner M, Böhne A, et al. Drivers and dynamics of a massive adaptive radiation in cichlid fishes. *Nature*. 2021;589(7840):76-81. doi:10.1038/s41586-020-2930-4

Roth G, Blanke J, Wake DB. Cell size predicts morphological complexity in the brains of frogs and salamanders. *Proc Natl Acad Sci U S A*. 1994;91(11):4796-4800. doi:10.1073/pnas.91.11.4796

Runemark A, Trier CN, Eroukhmanoff F, et al. Variation and constraints in hybrid genome formation. *Nat Ecol Evol*. 2018;2(3):549-556. doi:10.1038/s41559-017-0437-7

Salzburger W. Understanding explosive diversification through cichlid fish genomics. *Nat Rev Genet*. 2018;19(11):705-717. doi:10.1038/s41576-018-0043-9

Salzburger W, Bocxlaer B Van, Cohen AS. Ecology and evolution of the African great lakes and their faunas. *Annu Rev Ecol Evol Syst*. 2014;45:519-545.

doi:10.1146/annurev-ecolsys-120213-091804

Santos ME, Braasch I, Boileau N, et al. The evolution of cichlid fish egg-spots is linked with a cis-regulatory change. *Nat Commun*. 2014;5. doi:10.1038/ncomms6149

Schmalhausen, I. I. (1949). *Factors of evolution: the theory of stabilizing selection.* Blakiston.

U. Schmidt, in Palaeoecology of Africa and the Surrounding Islands, J. Runge, Ed. (Balkema, Lisse, Tokyo, 2001), pp. 51–62.

Seehausen O. Process and pattern in cichlid radiations - inferences for understanding unusually high rates of evolutionary diversification. *New Phytol*. 2015;207(2):304-312. doi:10.1111/nph.13450

Selz OM, Seehausen O. Interspecific hybridization can generate functional novelty in cichlid fish. *Proc R Soc B Biol Sci*. 2019;286(1913). doi:10.1098/rspb.2019.1621

Serrato-Capuchina A, Matute DR. The role of transposable elements in speciation. *Genes (Basel)*. 2018;9(5). doi:10.3390/genes9050254

Šímová I, Herben T. Geometrical constraints in the scaling relationships between genome size, cell size and cell cycle length in herbaceous plants. *Proc R Soc B Biol Sci*. 2012;279(1730):867-875. doi:10.1098/rspb.2011.1284

Smith EM, Gregory TR. Patterns of genome size diversity in the ray-finned fishes. *Hydrobiologia*. 2009;625(1):1-25. doi:10.1007/s10750-009-9724-x

Stager JC, Day JJ, Santini S. Comment on "Origin of the superflock of cichlid fishes from Lake Victoria, East Africa". *Science (80-)*. 2004;304(5673):325-330. doi:10.1126/science.1091978

Stelkens RB, Schmid C, Selz O, Seehausen O. Phenotypic novelty in experimental hybrids is predicted by the genetic distance between species of cichlid fish. *BMC Evol Biol*. 2009;9(1):1-13. doi:10.1186/1471-2148-9-283

Stelkens RB, Schmid C, Seehausen O. Hybrid breakdown in cichlid fish. *PLoS One*. 2015;10(5):1-11. doi:10.1371/journal.pone.0127207

Sun H, DIng J, Piednoël M, Schneeberger K. FindGSE: Estimating genome size variation within human and Arabidopsis using k -mer frequencies. *Bioinformatics*. 2018;34(4):550-557. doi:10.1093/bioinformatics/btx637

Svardal H, Salzburger W, Malinsky M. Genetic Variation and Hybridization in Evolutionary Radiations of Cichlid Fishes. *Annu Rev Anim Biosci*. 2021;9:55-79. doi:10.1146/annurev-animal-061220-023129

Tenaillon MI, Hollister JD, Gaut BS. A triptych of the evolution of plant transposable elements. *Trends Plant Sci*. 2010;15(8):471-478. doi:10.1016/j.tplants.2010.05.003

Terai Y, Mayer WE, Klein J, Tichy H, Okada N. The effect of selection on a long wavelength-sensitive (LWS) opsin gene of Lake Victoria cichlid fishes. *Proc Natl Acad Sci U S A*. 2002;99(24):15501-15506. doi:10.1073/pnas.232561099

Terai Y, Seehausen O, Sasaki T, et al. Divergent selection on opsins drives incipient speciation in Lake Victoria cichlids. *PLoS Biol*. 2006;4(12):2244-2251. doi:10.1371/journal.pbio.0040433

Thoday JM. Disruptive selection. *Proc R Soc London Ser B Biol Sci*. 1972;182(67):109-143. doi:10.1007/978-3-319-19650-3_2114

Thomas CA. The genetic organization of chromosomes. *Annu Rev Genet*. 1971;5:237-256. doi:10.1146/annurev.ge.05.120171.001321

Tørresen OK, Star B, Mier P, et al. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Res*. 2019;47(21):10994-11006. doi:10.1093/nar/gkz841

Tsukaya H. Does ploidy level directly control cell size? Counterevidence from arabidopsis genetics. *PLoS One*. 2013;8(12):1-7. doi:10.1371/journal.pone.0083729

Wandera SB, Balirwa JS. Fish species diversity and relative abundance in Lake Albert-Uganda. *Aquat Ecosyst Heal Manag*. 2010;13(3):284-293. doi:10.1080/14634988.2010.507120

Whitney KD, Baack EJ, Hamrick JL, et al. A role for nonadaptive processes in plant genome size evolution? *Evolution (N Y)*. 2010;64(7):2097-2109. doi:10.1111/j.1558-5646.2010.00967.x

Whitney KD, Garland T. Did genetic drift drive increases in genome complexity? *PLoS Genet*. 2010;6(8):1-6. doi:10.1371/journal.pgen.1001080

Wright NA, Gregory TR, Witt CC. Metabolic "engines" of flight drive genome size reduction in birds. *Proc R Soc B Biol Sci*. 2014;281(1779). doi:10.1098/rspb.2013.2780

Wright SI. Evolution of Genome Size. In: *ELS*. Wiley; 2017:1-6. doi:10.1002/9780470015902.a0023983

Zhang J. Evolution by gene duplication: An update. *Trends Ecol Evol*. 2003;18(6):292-298. doi:10.1016/S0169-5347(03)00033-8