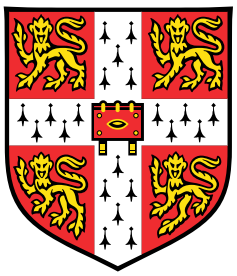
Epigenetic variability and inheritance in East African cichlid fishes



Grégoire M.L.P.G. Vernaz

Department of Genetics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Wolfson College

September 2019

I would like to dedicate this PhD thesis to my family. Their constant guidance and profound curiosity about the processes of Life have continuously shaped me and led me to carry out this doctoral research.

Je souhaite dédier cette thèse à ma famille. Leur direction aimante et avisée, ainsi que leur avide curiosité pour les processus de la Vie ont constamment participé à mon développement et sont à la base même de cette recherche doctorale.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 60,000 words exclusive of tables, footnotes, bibliography, and appendices.

Grégoire M.L.P.G. Vernaz
September 2019

Acknowledgements

Firstly, I would like to thank Professor Eric Miska who trusted me and gave me the opportunity to carry out this doctoral study in his laboratory. His enthusiasm and curiosity for Science have been greatly inspirational.

Importantly, I would like to express my gratitude to the funding bodies that financially supported my research, which includes Wellcome, the Genetics Society and Wolfson College, University of Cambridge. My thanks also go to the essential staff at the Gurdon and Sanger Institutes.

My thanks go to Tomás di Domenico for shedding light on the new *in silico* world that was lying before me at the very beginning of this journey; to Hannes Svardal and Milan Malinsky for their patience and for enthusiastically sharing their knowledge of population genomics and cichlid ecology; to Prof. Richard Durbin and the cichlid community in Cambridge for the many fruitful meetings and discussions.

Two field trips to Malawi and Tanzania were part of this thesis. I wish to thank the people that helped me during the field trips in Malawi and Tanzania: Richard, Bosco, Joseph, Everest and many others in the departments of fisheries in Malawi and Tanzania, as well as the many fishermen of the shores of Lake Malawi. Their expertise and tremendous help were essential to the success of both trips.

Parts of this thesis present collaborative work. I am deeply grateful to Prof. George Turner for immense help during field trip, insightful discussions and also for setting up and providing cichlid hybrids, and to Prof. Martin Genner for being an enthusiastic collaborator.

I am greatly indebted to Dr. Emília Santos, Prof. Eric Miska, Navin Ramakrishna and Prof. Martin Genner for their precious comments on the manuscript.

My beloved friends, Valérian, Linda, Annabelle, Joël, Isabela, Navin, Fabian, Jakob, Asia, Nadia, Maximilian, Milena for their immeasurable contributions to my life. Thanks guys.

My dear and amazing super postdoc, now PI in Zoology, Emília Santos for being a precious mentor and friend. To this memorable boat trip in Malawi. Your motivation and dedication have greatly inspired me.

Finally and most importantly, for their continuous, meaningful and unconditional support and love that made me. For their sensible guidance. This thesis is dedicated to you. To my father, mother, brothers Audren and Guillaume. To you, my dearest Brigitte.

Cichlids of Lake Malawi
The five cichlid species studied in this thesis.
Painting by Soviet – Nkhata Bay, Malawi. April 2018

Abstract

The hundreds of cichlid species forming part of the radiation of East African Lakes show a remarkable diversity of phenotypic and ecological adaptations. Despite this, recent studies highlighted that genetic diversity within the radiation is among the lowest ever observed in vertebrates.

Such a high phenotype/genotype diversity ratio makes cichlids a promising system to investigate the role of genetics and, for the first time at a species level, epigenetics in the context of adaptation and convergent evolution. Yet, the molecular mechanisms and, in particular, any epigenetic aspects underlying such phenotypic diversity and speciation success remain largely unknown.

Here, I focus on whole-genome DNA methylation (methylome), a heritable and dynamic epigenetic mark that has been reported to be responsible for rapid and transmissible changes in phenotype in plants and mammals. In light with phenotypic plasticity related to diet adaptation, I hypothesise that the liver methylome may affect liver function and thus be related to diet. I thus performed sequencing of liver tissues of different cichlid species presenting distinct eco-morphological and trophic adaptation from both Lake Malawi and crater lake Massoko, Tanzania.

The main results reveal striking differences in methylome at conserved underlying DNA sequences – some variation shared in cichlids of both lakes. Furthermore, I observe an enrichment for methylome variation in transposable elements (TE) and promoter regions. Remarkably, most of the variation (ca 80%) common to fishes from both lakes are located in TEs and is, in part, correlated with differential expression levels at some key metabolic and developmental genes in liver. This suggests a possible conserved role of TE-related methylome in the adaptation of liver function.

Furthermore, I generated inter-species hybrids to investigate the inheritance of DNAme variation in cichlids. The liver methylome of F1 hybrids, although mostly resembling parental methylomes, exhibited some level of divergence, suggesting unique DNAme patterns in hybrid offspring and possible transgressive segregation.

I conclude there might be a conserved crosstalk between the local environment and methylome in different natural populations of cichlids. The results presented in this thesis postulate an important role of natural DNA methylation variation in promoting adaptive phenotypic diversification in divergent habitats during the early stages of speciation.

Table of contents

List of figures

List of tables

Nomenclature

Species

AC	*Astatotilapia calliptera* (endemic to the Lake Malawi catchment)
AS	*Aulonocara stuartgranti* (Lake Malawi)
Benthic	AC spp. benthic Lake Massoko, also known as the deep/blue ecomorph
Itupi	AC spp. Itupi (river part of the Lake Malawi catchment)
Littoral	AC spp. littoral Lake Massoko, also known as the shallow/yellow ecomorph
Mbaka	AC spp. Mbaka (river part of the Lake Malawi catchment)
MZ	*Maylandia zebra* (Lake Malawi)
PG	*Petrotilapia genalutea* (Lake Malawi)
PN	*Pundamilia nyererei* (Lake Tanganyika)
RL	*Rhamphochromis longiceps* (Lake Malawi)
Usisya	AS spp. Usisya (Usisya refers to the name of a town on the shore of Lake Malawi)

Greek Symbols

μ	mean of a population
σ	standard deviation

Acronyms / Abbreviations

A	Adenine

bp	Base pair
BSeq	Bisulfite sequencing
C	Cytosine
CGI	CpG island
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeat DNA sequences
dCas9	Catalytically inactive dead CRISPR-associated (Cas) 9
DEG	Differentially expressed gene
DMR	Differentially methylated region
DNA	Deoxyribonucleic acid
DNAme	DNA methylation, namely DNA 5-methylcytosine (5mC)
G	Guanine
GO	Gene ontology
HMW-gDNA	High molecular-weight genomic DNA
IAP	Intra-cisternal A particle
IC95	95% confidence interval
indels	Insertion or deletion of bases, from 1 to 10,000 base pairs in length
kbp	Kilo base pair, 1000 bp
KRAB-ZFPs	Krüppel-associated box (KRAB) domain-containing zinc-finger proteins (ZFP)
LINE	Long interspersed nuclear element
LTR	Long terminal repeats
mCpG	Dinucleotide 5-methylcytosine followed by guanine, linked by phosphate group

NGS	Next-generation sequencing
O/E	Observed over Expected ratio
padj	Statistical p-value adjusted for multiple testing comparisons
PC	Principal component; by extensions, PCA refers to PC analysis
PCR	Polymerase chain reaction
PGC	Primordial germ cell
piRNA	Piwi-interacting RNA
PTGS	Post-transcriptional gene silencing
RNA	Ribonucleic acid
RNAseq	RNA sequencing
RRBS	Reduced Representation Bisulfite Sequencing
rRNA	ribosomal RNA
SEM	Standard error of the mean
SINE	Short interspersed nuclear element
SNP	Single-nucleotide polymorphism
spp.	Species sub-population
TES	Transcription End Site
TE	Transposable elements and repeats
TPM	Transcript Per Kilobase Million
tRNA	transfer RNA
TSS	Transcription Start Site
T	Thymine
C	Uracil

WGBS	Whole Genome Bisulfite Sequencing
ZGA	Zygotic genome activation

Chapter 1

Introduction

1.1 Epigenetic and phenotypic evolution

1.1.1 Historical overview

The inheritance of traits over generations, from parents to their offspring is mainly explained by the transmission of genetic information contained in chromosomes. Yet, some heritable phenotypes have been observed to occur in the absence of changes in the underlying DNA sequences, defying the Mendelian principles of inheritance [1–3].

In the Western world, the first observations of the transmission of acquired phenotypes in response to environmental or physiological stimuli have their roots in Ancient Greece. In ca. 400 B.C., Hippocrates, along with Aristotle, observed and formulated the Darwinian concept of 'pangenesis' or the inheritance of acquired traits [4]. The study of the inheritance of acquired traits was later expanded on by the French naturalist Jean-Baptiste de Lamarck in his theory of evolution [5]. During that same period, the seminal work of his contemporary, Charles Darwin, postulated that evolutionary adaptation may occur through natural selection [6]. In 1883, the theories of August Weismann rejected the concepts formulated by Lamarck, in stating that somatic and germ cells were separated and thus any environmentally-induced modifications may not be passed on to gametes, which set the basis for the Weismann's barrier [7]. It is only much recently that some researchers have highlighted a number of heritable phenotypic variation that cannot be fully explained by Mendel's principles of heredity, that is, via DNA-sequence based mechanisms alone [8–10]. Such mechanisms could therefore have evolved in the context natural selection as promoting phenotypic plasticity and thus possibly facilitating adaptation, speciation and adaptive radiation [11].

The definition of epigenetics has been a long-standing debate. In 1942, Conrad Waddington defined the study of 'epigenesis' as being "the processes by which the genotype brings the phenotype into being". Since Waddington, epigenetic studies have been at the centre of intense research carried out in plants, animals and other non-model organisms. In 1996, in light with the accumulation of experimental evidence, A. Riggs and colleagues [12] offered a revised definition of the epigenetic processes as being "mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence". This has laid the concept of non-DNA sequence-based inheritance of molecular states and traits [10]. The adaptive role of epigenetic processes was later on proposed by Adrian Bird in 2007, pioneer for his research on DNA methylation, defining epigenetics as any "structural adaptations of chromosomal regions so as to register, signal or perpetuate altered activity states" [13].

In brief, epigenetic mechanisms encompass any traits or molecular states that can be not only inherited from one dividing cell to the next (mitosis), but also between generations (meiosis), without changes in the DNA sequence, resulting in heritable and reversible phenotypes (such as altered transcriptional activity). Epigenetic processes might provide a way to transmit molecular states transgenerationally without any genetic implication, and in response to environmental or physiological stimuli [13, 7, 14]. Moreover, such epigenetic memories could possibly drive evolutionary changes through phenotypic variation, and, just like mutations, could be part of the natural selection process in one population if resulting in increased organismal fitness via the emergence of advantageous adaptive traits [2, 11].

The following sections introduce and discuss the current knowledge on epigenetic mechanisms, their functions in cellular identity, genome defence and stability, and during organismal development. In the second part of the introduction, I summarise the potential crosstalk between epigenetics and the environment, possibly promoting phenotypic diversity.

1.1.2 Epigenetic mechanisms

Three major carriers of epigenetic information have been identified and recognised thus far [15, 16, 1, 17] and are the following:

DNA cytosine methylation

Covalent modification of DNA cytosine, namely, the addition of one methyl group onto the 5th carbon of the nucleotide cytosine is a conserved feature of many genomes, from prokaryotes to eukaryotes [18, 19]. Extensive work has been carried out and has revealed

essential regulatory functions pertaining to immune defence mechanisms, cell differentiation and embryogenesis, as well as a possible role in adaptation. In vertebrate genomes, cytosine methylation is considered a major regulatory innovation [17, 20, 2, 21]. 5-methylcytosine (5mC), or sometimes referred to as DNA methylation or simply DNAme, will be extensively introduced below.

Histone modification

Approximatively 147 base pairs of DNA are wrapped around the highly conserved histone proteins. In addition to the covalent addition of a methyl group to DNA cytosine, more than 100 histone modifications have been catalogued as well as many chromatin modifying enzymes [22]. These reversible and heritable chemical modifications are mainly located on the amino acid tails of the histone proteins and can participate in the remodelling of chromatin architecture (biophysical interactions between histone marks impacting chromatin compaction, such as histone acetylation) and in the recruitment of DNA-interacting enzyme complexes [23, 24]. Different genomic regions are marked with different histone modifications: active promoters are marked by histone H3 lysine 4 dimethylation (H3K4me2), H3K4me3, acetylation, while enhancers are tagged with H3K4me1, H3K4me2, H3K27ac, for example [22]. There is strong evidence for extensive functional interactions between histone marks and DNA methylation in many eukaryotes, often acting in concert to modify chromatin accessibility and regulate gene expression [25]. Few examples are given and described throughout his introduction.

Non-coding RNAs

Many classes of non-coding RNAs (ncRNAs) exist, each of them exerting distinct regulatory functions in general [26]. Historically, ncRNAs have been arbitrarily grouped into two groups based on their length: the long non-coding RNAs (lncRNAs) and the small non-coding RNAs (sncRNAs).

lncRNA are defined as being longer than 200 nucleotides and have been reported to exert a wealth of biological roles. A famous of example of lncRNA is the X-inactive specific transcript (*XIST*) [27]. In female therian mammals, one X chromosome is randomly silenced via the activity in *cis* of *XIST*. The accumulation of Xist RNA over the chromosome represses X-linked genes and is followed by the deposition of repressive histone marks and, finally, DNAme, ensuring a stable repression [17, 27]. Thus, X-chromosome inactivation requires

the cooperation of ncRNAs, DNAme and histone marks to ensure dosage compensation as a way to equalise X-linked gene expression.

Small non-coding RNAs are equally as diverse in functions and lengths (<200 nt-long). Many classes have been identified, such as PIWI-interacting RNA (piRNA) involved in transposable elements (TE) repression and post-transcriptional gene silencing, microRNA participating in post-transcriptional silencing, tRNA fragments possibly involved in TE repression, and small-interfering RNA (siRNAs), just to cite a few.

The activity of sncRNA are intertwined with the ones of DNAme and histone modifications, in that ncRNAs can direct the deposition of mC and/or histone marks at particular genomic loci, participating in the regulation of host and transposon gene expression [28]. Furthermore, ncRNAs have been reported to play important roles in the transgenerational epigenetic inheritance of certain traits [1, 16, 29, 30]

1.1.3 Epigenetics and cellular fate in multicellular organisms

Although this thesis mostly focuses on the variability and possible roles of the epigenetic mark 5mC in promoting phenotypic diversity in natural populations of cichlids, a general overview of the roles of epigenetic mechanisms in cellular identity is given below.

During animal development, the crucial roles of epigenetic mechanisms have been characterised very early on in 1950s. It is now well established that epigenetic processes act in concert to ensure both the establishment and the maintenance of cellular identities from early developmental stages and in fully differentiated somatic cells [1] (see the section, 1.3). This role is best reflected in Waddington's definition of epigenetics, in that epigenetics pertain to the mechanisms underlying the processes by which the genotype produces the phenotypes in the context of development.

Seminal work performed by John Gurdon and colleagues more than 50 years ago revealed that reprogramming of a somatic cell back to totipotency by transferring the nucleus of a fully differentiated somatic cell into an enucleated oocyte was very inefficient, although not impossible [31]. Very early on during vertebrate development, cellular identity is established and then maintained in somatic cells, as a result of intense interplay between epigenetic mechanisms and transcription factors. Such epigenetic marks carry memories of one somatic cellular state, which would in turn impede cellular reprogramming, unless such epigenetic memory is erased or reset [32, 33]. Cell-type transition during development is mostly determined and put in place by finely tuned epigenetic processes, which ensure

the correct establishment of chromatin states and transcriptional patterns pertaining to each cell fate [17, 18]. Many conserved molecular actors (characterised later on) participate in defining the epigenetic landscape in somatic and germinal cells; any loss of function generally results in severe defects or embryonic lethality, highlighting the organismal dependence on such mechanisms [17]. The coordinated deployment of many transcription factors during animal development is crucial to ensure cellular differentiation and their activity is tightly linked to epigenetic mechanisms [33]. Such epigenetic patterns are then maintained and transmitted throughout cell divisions even without the initial modifying actors and in a non-DNA sequence based manner, possibly also maintained through meiosis. Such observations are reflected in the definitions of epigenetics put forward by Riggs, Holiday, and Bird [13, 12, 1], that takes into account the mitotic/meiotic inheritance of such epigenetic marks without the initial trigger.

In summary, and before characterising the different roles of DNA methylation in detail, epigenetic processes are major actors in defining cellular identity, resulting in transcriptional expression distinct to each cell type. Many epigenetic processes can act in concert to regulate chromatin states, directly impacting chromatin accessibility and the ability of DNA-binding factors to interact in *cis*. Moreover, epigenetic processes have an important ancestral role in defending the host genome against selfish elements, dicing nascent TE transcripts and repressing their transcription [34, 20], which is discussed further below.

1.2 DNA methylation - 5-methylcytosine

The following sections deal with the metabolism and specific functions of 5-methylcytosine in eukaryotic genomes. Although many features of DNAme are widely conserved in eukaryotes, distinct functions have evolved independently and are restricted to some taxa and even sometimes to some species. Here, I summarise the current knowledge on DNA methylation in a comparative manner. A great emphasis is placed on teleost fishes.

1.2.1 Conservation of 5mC across kingdoms

The presence of methylated cytosine in DNA is a widely conserved genomic feature in many organisms across kingdoms and might have its origins in bacteria [35]. The addition of the methyl group onto the fifth carbon of the nucleotide cytosine is mediated, from bacteria to plants and vertebrates, by the conserved enzyme family of methyltransferases.

Bacteria have evolved different methyltransferase enzymes to protect their genomes from bacteriophage infections. Some bacteria extensively methylate their own endogenous genome (both adenine and cytosine methylation), allowing for the specific degradation of exogenous, parasitic DNA fragments. Such processes are carried out by the restriction-modification system that degrades unmethylated DNA fragments. In some bacteria however, probably in the microbial ancestors of the eukaryotes, such a process changed to not only specifically methylate cytosines of exogenous DNA fragments, but also to participate in host transcription regulation. This suggests that DNAme as a molecular tool to control gene activity as well as to suppress exogenous DNA fragments has evolved in some bacteria before eukaryotic life [34]. This has major implications, in that exogenous DNAs are not degraded any longer, but become part of the host genome and are replicated in their repressed state mostly. There is *de facto* a direct correlation between eukaryotic genome length and genomic TE load [36]. This has also given rise to an arms race evolution between these exogenous DNA sequences and host defence mechanisms. Furthermore, possible co-options and domestications of some functions derived from such genomic parasites to their host fitness advantage have been observed and are discussed later – in particular in sexual organisms [19, 37, 38]. Strikingly, the catalytic domain of DNA methyltransferases in vertebrates shows high sequence conservation with their bacterial homologues, suggesting a common evolutionary origin.

Here below are summarised the degree of conservation of 5mC in eukaryotic genomes and the associated metabolic pathways.

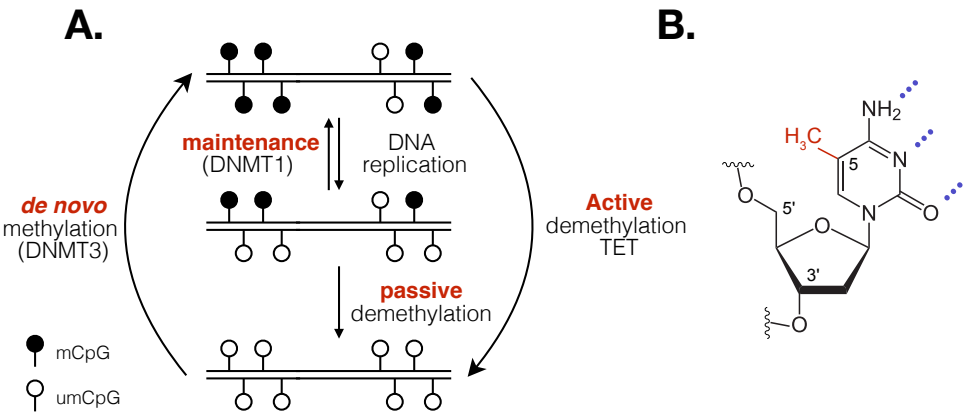


Fig. 1.1 Structure and metabolism of DNA 5-methylcytosine. **A.** Metabolic pathways of CpG methylation. The enzyme family DNA methyltransferases are primarily involved in the maintenance of 5mC upon DNA replication (DNMT1 recognises hemimethylated CpG), as well as *de novo* methylation (DNMT3 family). Active demethylation process is catalysed by the TET enzyme family or passively through DNA replication-mediated dilution. **B.** Chemical structure of 5-methylcytosine. One methyl group (in red) is added onto the 5^{th} carbon of the nucleotide cytosine by the DNMT enzymes. Blue dots indicate non-covalent hydrogen bonds with guanine (Watson-Crick base pairing). Wiggly lines represent the start of the phosphodiester bond with the next base.

5mC is a conserved feature of most eukaryote genomes

Despite its broad conservation, DNAme has been lost in many organisms [39]. Some important model organisms, such as the worm *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster* as well as yeasts and many fungi are virtually devoid of any DNA 5mC. These are usually exceptions. In other nematodes, DNA methylation of repeat sequences is particularly conserved and is likely to be a feature present in the common ancestor of the nematodes [40], unlike methylation in bodies of genes, which has been lost in many nematode lineages [40] but seems to be highly conserved in many organisms, such as plants, vertebrates and fungus [41].

The enrichment for cytosine methylation in transposon sequences is thought to have evolved independently in plants and animal [35, 41, 40] and can be mediated in both cases by small non-coding RNAs. However the exact mechanisms are slightly different and are discussed further in the introduction (in brief, TE-targeted DNA methylation is thought to be mediated by PIWI-interacting small non-coding RNAs [piRNAs] in most animals via still unclear interactions with the DNAme machinery, and by the well-established RNA-depend DNA methylation [RdDM] system via the production of small-interfering RNAs [siRNAs] in most plants [20, 28]).

Methylation mostly occurs in a sequence context-specific manner. In animals, cytosine methylation mostly occurs in a CpG dinucleotide context and is symmetrical. Cytosine methylation in other sequence contexts, such as CHG and CHH (where H is either A, C, or T) have been extensively observed in plants (outside gene bodies, principally in transposon-related sequences – probably recognised by distinct mechanisms) and in brain and pluripotent cells in human and mouse at least, but usually at much lower levels genome-wide than CG methylation [41, 42, 17]. In addition, while CG methylation is mostly symmetrical in most organisms, methylation of cytosine in other sequence contexts can be asymmetrical, with one strand only harbouring methylation.

Interestingly, metazoan genomes feature a remarkable CpG depletion [43, 44]. In fact, cytosine methylation might come at a cost: the inherently mutagenic nature of mC. Methylated cytosines are more liable to deaminate spontaneously [45, 46] – a C-T transition that could lead to deleterious mutations. Any C-to-T transition might not always be efficiently repaired by the cellular machinery, in particular in plants [45, 47]. This has resulted in genomic depletion of CG dinucleotides in organisms harbouring CpG methylation (see Fig.2.11 for detailed comparison of genomic CpG depletion). However, recent work making use of thousands of sequenced genomes has highlighted that although methylated cytosines were associated with increased mutability not only at the focal mC site but also in neighbouring nucleotides in plants (*A. thaliana* and rice), this observation might not be true in animals. In humans, mCG and neighbouring nucleotides have been reported to be less likely to mutate, in stark contrast with observations in plants. This highlights that the repair machinery has differently evolved in plants and mammals, ensuring more reliable repair mechanisms of spontaneous deamination in mammals. This might also reflect the different biological functions of DNAme between plants and animals [47].

In addition to spontaneous deamination leading to DNA mutations, the activity of DNMTs have been reported to generate alkylation DNA damage (through production of 3-methylcytosine; see Fig. 1.2a). Interestingly, mechanisms involved in the repair of DNMT-related DNA alkylation, in particular the DNA oxidative demethylase ALKB2/3, have been shown to have co-evolved together with DNA methylation machinery in animals [40]. 5-methylcytosine produced by off-target activity of DNMTs, if not repaired, causes the blockage of DNA polymerase upon DNA replication, leading to double-strand DNA breaks.

In conclusion, DNA methylation poses numerous threats to the cells, from potential deleterious mutations to DNA damage leading to impeded DNA replication. Such high costs could explain the loss of DNA methylation in several organisms, where the diverse functions of DNA methylation might not have balanced the costs.

Importantly, most vertebrate genomes display high levels of CG methylation genome-wide (70-80% mCG/CG), in contrast to invertebrate and plant genomes that present usually much lower methylation levels (4-50%), with DNA methylation often localised to some genomic regions (gene bodies and repeats in particular) [35, 41].

Finally, the presence of DNAme and the "complexity" of one organism does not seem to correlate. One of the most conserved feature of DNA methylation across kingdoms seems to be related to TE silencing and to gene bodies. Zemach and Zilberman as well as Bestor [19, 37] have proposed that DNA methylation might be especially relevant in animals with sexual reproduction.

1.3 Metabolism of DNA methylation in vertebrates

In the following sections, most of the studies and knowledge of DNA methylation is based on mammalian work, where the different pathways associated with DNA cytosine methylation have been thoroughly investigated. A comparative approach will be used to understand the conservation of biological functions and protein homology.

In the 1970s, seminal works by R. Holliday [49] and A. Riggs [50] described DNA methylation as a heritable mark, actively maintained in cells with important regulatory functions, such as X chromosome inactivation. In addition, they postulated that DNA methylation could alter DNA-protein interaction and participate in cell-type differentiation.

The metabolic pathways involved in DNA methylation are widely conserved across animals and plants. Overall, two distinct families of enzymes are thought to participate in the processes leading to the addition or erasure of DNA 5-methylcytosine. On the one hand, the enzyme family of DNA methyltransferases (DNMTs) maintain or create new patterns of mC, while, on the other hand, the enzymes ten-eleven translocation methylcytosine dioxygenase (TETs) catalyse active demethylation [51].

Interestingly, within teleost fishes, many paralogues of the genes involved in DNA methylation metabolism have been identified. This could be due to the genome-wide duplication event observed in teleost fish [52] in addition to independent duplications. Lake Malawi cichlid fishes in particular exhibit a very high number of gene duplicates, which might have promoted their successful adaptative radiation [53]. Gene duplication could allow for novel gene function (neofunctionalization), while preserving the original, ancestral gene function [52].

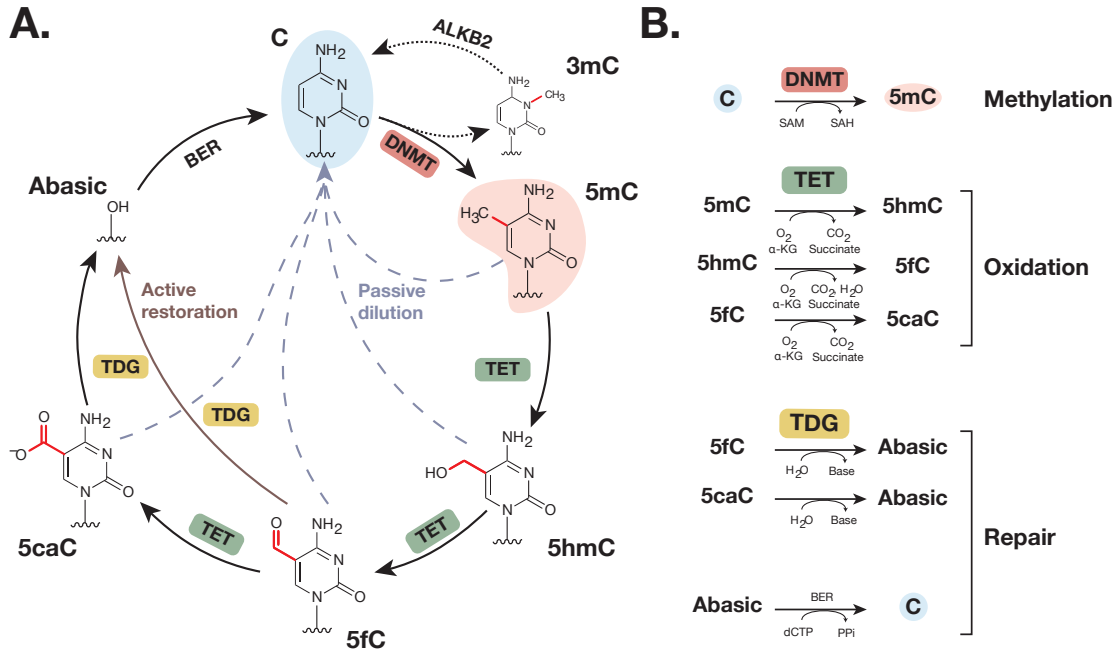


Fig. 1.2 Pathways for dynamic DNA cytosine methylation.
A. The addition of one methyl group is catalysed by methyltransferase enzymes (DNMTs), converting cytosine (C) into 5-methylcytosine (5mC). In some instances, 3-methylcytosine (3mC) can also be produced, and is be repaired by ALBH enzymes. There are two distinct pathways participating in the cytosine demethylation process: an active oxidative pathway, involving several enzymatic reactions, and a passive pathway via replication-dependent dilution of modified C. The active removal of 5mC is carried out by TET enzymes. First, mC will be oxidised into 5-hydroxymethylcytosine (5hmC), then into 5-formylcytosine (5fC) and finally into 5-carboxylcytosine. Both 5fC and 5caC can then be excised by TDG, thus producing an abasic site, which will then removed and replaced with unmodified C via BER. The passive demethylation process occurs upon DNA replication, when methylated and oxidised forms of cytosine are not maintained, therefore leading to passive dilution of modified C. **B.** The different enzymatic reactions part of DNA methylation metabolism are shown. DNMT, DNA methyltransferase; TET, ten-eleven translocation methylcytosine dioxygenase; TDG, G/T mismatch-specific thymine DNA glycosylase; ALKBH, alpha-ketoglutarate-dependent hydroxylase; BER, the cellular mechanism of base excision repair; α-KG, *alpha*-Ketoglutaric acid. SAM, *S*-Adenosyl methionine; SAH, *S*-adenosyl homocysteine; dCTP, deoxycytidine triphosphate; PP_i, pyrophosphate. Figure adapted from [48, 40].

1.3.1 DNMTs - 5mC writers

In most animals, all DNMT enzymes use the co-factor S-adenosylmethionine (SAM) as a methyl group donor in order to covalently add one methyl group onto the 5th carbon of cytosine [54] (see Figs. 1.1b and 1.2).

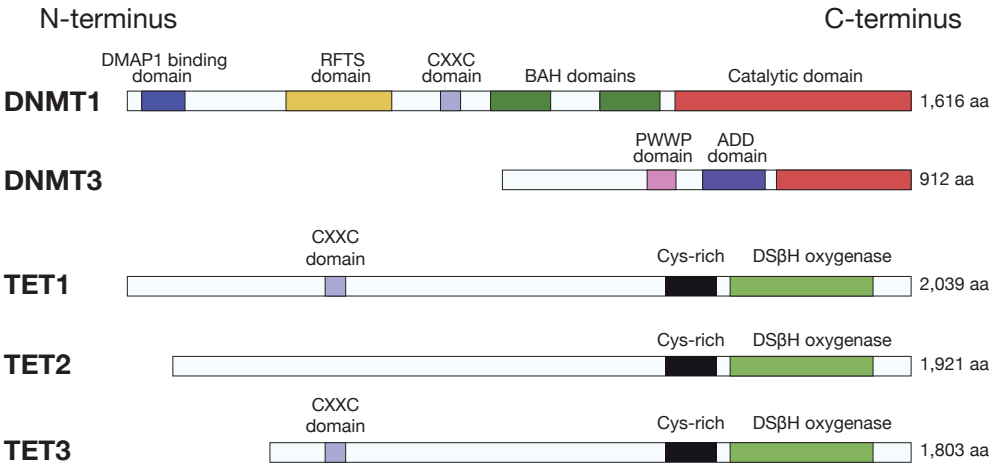


Fig. 1.3 DNA methylation writers and erasers in animals. In most animals, all DNMT proteins possess a catalytic domain (methyltransferase activity, in red) – which is also the only domain of the RNA methyltransferase DNMT2 (not shown). In addition, the N-terminus of DNMT1 contains four conserved regulatory domains. DNMT3 contain one PWWP domain (conserved in the six zebrafish paralogues) and ADD domains (partly conserved in mbuna as well). TET enzymes all contain a double-stranded β-helix (DSβH) fold core oxygenase domain and a cysteine (Cys)-rich domain. TET1 and TET3 contain an additional CXXC domain in their N-terminus. Pro-Trp-Trp-Pro (PWWP) domain; ADD, ATRX–DNMT3–DNMT3L; BAH, bromo-adjacent homology; DMAP1, DNA methyltransferase 1-associated protein 1; RFTS, replication foci targeting sequence. The number of amino acid is for human and mouse homologues for DNMTs and TET, respectively. Not to scale. Figures adapted from [54, 48, 25].

DNMT1

Although DNMT1 is thought to be primarily involved in the maintenance of DNA-cytosine methylation patterns and DNMT3 enzymes in the *de novo* methylation process (Fig. 1.1a), their roles are not exclusive and might in certain instances overlap. DNMT1 has been shown to have *de novo* methylation roles at certain repetitive elements for example *in vivo* in mouse [59]. Vice versa, DNMT3 enzymes also participate in DNA methylation maintenance. Therefore, the functions of DNMTs are broader than originally thought, and might also include roles in transcription activation and post-transcriptional regulation via molecular interactions, which are discussed in detail in the following sections [54].

Table 1.1 Conservation of DNMT and TET genes in vertebrates

mouse	**human**	**zebrafish**[1]
Dnmt3a	*DNMT3A*	*dnmt6* (*dnmt3ab*, *dnmt3a1*) *dnmt8* (*dnmt3aa*, *dnmt3a2*)
Dnmt3b	*DNMT3B*	*dnmt3* (*dnmt3bb*.2, *dnmt3b3*) *dnmt4* (*dnmt3bb.1*, *dnmt3b1*) *dnmt5* (*dnmt3bb.2*, *dnmt3b4*) *dnmt7* (*dnmt3ba*, *dnmt3b2*)
Dnmt3c	-	-
Dnmt3l	*DNMT3L*	-
Dnmt1	*DNMT1*	*dnmt1*
Tet1	*TET1*	*tet1*
Tet2	*TET2*	*tet2*
Tet3	*TET3*	*tet3*

alternative gene names in brackets
[1] Ref. [55–58]

At the protein level, all DNMTs in most animals possess the methyltransferase domain in their C-terminus (Fig. 1.3), catalysing the transfer of the methyl group from SAM to the 5th carbon of the cytosine ring [54] (Fig. 1.2).

The N-terminus of DNMT1 exerts regulatory functions, and is composed of four conserved domains, allowing for molecular interactions (Fig. 1.3). The DNMT1-associated protein 1 (DMAP1) binding domain is required for the interaction between DNMT1 and the histone modifying enzyme HDAC2 (histone deacetylase 2) via the transcriptional repressor DNMAP1 (DNA methyltransferase 1 associated protein 1). The CXXC (a zinc-finger domain composed of eight conserved cysteine residues) domain of DNMT1 specifically recognises unmethylated CG dinucleotides, while the RFTS (replication foci targeting sequence) domain brings the enzyme to the replicating forks, promoting maintenance of DNAme patterns upon DNA replication [54, 25].

Crosstalk between DNMT1 and specific modified histones at the replication forks have been described, mainly via the interaction of DNMT1 with the adaptor protein UHRF1 (Ubiquitin-like, containing PHD and RING finger domains 1). The latter might play a crucial role in DNA methylation maintenance by bringing together histones harbouring specific modifications (H3K9me2 and H3K9me3, via the TUDOR TTD domain), hemimethylated CpG DNA and DNMT1 at the replication forks [54, 17]. UHRF1, via its UBL (ubiquitin-like) domain, recruits DNMT1 to the replication forks, structurally modifying DNMT1 from an

auto-inhibitory configuration to an active form. The RFTS domain can then bind histone H3 tails: DNMT1 is now able to methylate the daughter strand [25]. Loss of UHRF1, in zebrafish, mouse and *Arabidopsis* (VIM, plant homologue of UHRF1) has been linked to drastic reduction in global CG methylation in gene bodies, repeats and transposons [41, 17]. Moreover, DNMT1 is required for the terminal differentiation of normal tissues involving important interactions with histone modifying enzymes [60]. Lastly, the function of the two tandem BAH (Bromo-adjacent homology) domains remains elusive [25, 54].

DNMT1 is widely conserved across animals and only a single copy of the gene *dnmt1* is found across invertebrates, mammals, zebrafish and the cichlid mbuna in general [61, 41, 40] (Fig. 2.3). Yet, possible paralogues of the *dnmt1* gene have been identified on different chromosomes in some teleost fish (salmon and trout) and could have emerged from a lineage-specific genomic event. Such duplication of the gene *dnmt1* might promote novel protein function and is in line with a rapidly evolving DNA methylation machinery [61]. The facilitator UHRF1, is widely conserved as well, from mammals to mbuna (orthologues), and appears to have co-evolved with DNMT1 in animals [41, 28].

Loss of function of DNMT1 has been associated in mammals with low global methylation, transcriptional de-repression of IAP transposons and ultimately early embryonic lethality [17]. Similarly, in zebrafish, *DNMT1*-knock-down is in general associated with 40% lethality during gastrulation or severe developmental defects of specific organs (retina, exocrine pancreas, intestine) [60].

In plants (in particular characterised in *A. thaliana*), the maintenance of DNA methylation patterns at CG sites is carried out by DNMT1 homologues: symmetrical CG methylation at TE and gene sequences is maintained by DNA methyltransferase, MET1, via its interaction with VIM (VARIANT IN METHYLATION; plant homologues of UHRF1) proteins binding to hemimethylated DNA stretches [28, 2]. Cytosine methylation in other contexts, in particular over TE sequences (where most of DNA methylation is localised in *Arabidopsis*), is maintained by DOMAINS REARRANGED METHYLTRANSFERASEs, DRM1/2, via the RNA dependent methylation pathway (RdDM, for CHH asymmetrical methylation) involving siRNAs, and by CHROMOMETHYLASE 2 and 3 (CMT2/3), a plant-specific DNA methyltransferase (for CHG and sometimes CHH contexts) via a self-reinforcing loop implicating histone marks [20, 30, 2].

Of important note, the protein DNMT2, despite its name, is associated with RNA methylation.

DNMT3

Like DNMT1, the C-terminal regions of the DNMT3 family contain the methyltransferase catalytic domain (Fig. 2.3). The N-terminal part of DNMT3A and DNMT3B in mammals however contain distinct domains: one PWWP (Pro-Trp-Trp-Pro) that mediates interaction with specific histone marks (H3K36me2 and H3K36me3) and one ADD (ATRX-DNMT3-DNMT3L) domain, a zinc-finger protein domain important for promoting the interaction of DNMT3 complexes with unmethylated H3K36 [17].

In zebrafish, six DNMT3 paralogues have been identified and have been shown to be transcribed in a tissue-specific manner and at different developmental times, which might reflect specific and novel functions for DNMTs [61, 58]. *dnmt6* and *dnmt8* are related to vertebrate *DNMT3a*, while *dnmt3*, *dnmt4*, *dnmt* 5 and *dnmt7* are more similar to vertebrate *DNMT3b* in terms of sequence similarity (Fig. 1.3 and Table 1.1). All vertebrate DNMT3A, DNMT3B and the six zebrafish DNMT3 paralogues contain both the methyltransferase catalytic domain in their C-terminus and the PWWP domains in the N-terminus [54, 57].

In addition to the different molecular interactions of *DNMT3* that regulate their functions, DNMT3 genes have been lost and duplicated in the genome of many organisms, leading to lineage- or species-specific gain or loss of function related to DNMT3 [54]. In rodents, the emergence of the paralogue DNMT3C (duplication) seems to be associated with the silencing of rodent-specific evolutionary young transposable elements [62]. The mammal-specific DNMT3L, is a DNMT3A cofactor lacking the PWWP domain and containing a truncated version of the catalytic C-terminal domain. DNMT3L has been linked to mammalian imprinting. Both DNMT3L and DNMT3A are not found in teleost fishes (Fig. 1.3). In fish, the extreme diversity of DNMT3 paralogues, expressed at different developmental times and in a tissue-specific manner, probably reflects some hitherto unknown distinct functions of DNA methylation restricted to teleost fish [61, 58, 63].

Importantly, *de novo* methylation of transposon promoters has been shown to be mediated by a class of small-non coding RNA, the PIWI-interacting RNA in the germ line of animals, via the recruitment of the DNAme machinery to TE sequences, however the exact mechanism remains unknown [17]. Similar mechanisms exist in plants and involve the RdDM (RNA-dependent DNA methylation) system, leading to siRNAs-mediated methylation of TE sequences. In higher vertebrates, another mechanism leading to DNAme-mediated TE silencing exist and involve KRAB zinc finger proteins (ZFP). Both sncRNAs- and KRAB-ZFP-driven silencing of transposon activity through DNAme are discussed in detail in section 1.4.5.

Knockout (KO) studies in mouse have revealed severe phenotypes following loss of DNMT3-related functions. In fact, loss-of-function of any of the *Dnmt3* is generally associated with embryonic (*Dnmt3B*) or postnatal lethality (*Dnmt3A*) and with deleterious activity of some then de-repressed transposon families. In addition, DNMT3C-KO in mouse leads to male sterility probably due to lack of DNA methylation-mediated suppression of TEs during spermatogenesis, while a loss of function associated with the mammal-specific *Dnmt3L* leads to male sterility and mid-gastrulation lethality in females [17]. Knock-down studies in zebrafish have revealed essential functions of *dnmt3* in regulating tissue differentiation, leading to embryonic lethality at 96 hours post fertilisation [63].

In conclusion, the DNA methylation machinery, in particular associated with DNMTs, seems to be evolutionary fast evolving, with many events of genic losses and duplication, leading to lineage-specific gain of function. This could also reflect the evolutionary arms race between fast evolving TE sequences and host defence mechanisms mediated in part by DNMT3. In addition, loss of function in genes involved in DNA methylation maintenance and establishment are generally embryologically lethal in most vertebrates, and are therefore required for tissue development and homeostasis [61, 64, 17]. Such mechanisms confer to the cells regulatory plasticity and stable epigenetic heritability in response to physiological stimuli [1].

1.3.2 TETs - 5mC erasers

In mammals as well as in zebrafish and the cichlid mbuna, three *tet* genes have been described (Table 1.1). All the TET family members harbour a C-terminal catalytic domain (Fig. 1.3), containing metal binding residues essential for the oxidative reactions, part of the active demethylation of modified cytosine nucleobases [48, 61]. TET enzymes catalyse the oxidation of 5mC into 5hmC, 5hmC into 5fC and 5fC into 5caC, using oxygen and αKG as main substrates, together with a ferric cofactor (Fig. 1.2 and ref.[65]). Some oxidised forms of cytosine might be stable epigenetic marks *per se* (at very low concentration genome-wide), with possible distinct biological functions [66]. Furthermore, TET enzymes are known to interact with many cellular actors, including metabolite- and nutrient-sensing factors, and have important functions during development, cell-type transition (differentiation) comprising somatic cell reprogramming, and also has neural functions [65, 67–70]. These functions are usually achieved through active demethylation at key enhancers and other regulatory regions that may promote transcriptional changes to happen. Loss of function related to *TET* genes are a hallmark of many cancers and many diseases [65].

During zebrafish development, TET enzymes are extremely lowly expressed [71, 72] and 5hmC levels could not be detected in one study via immunohistochemistry in embryos [72]. Demethylation processes, in particular of the maternal methylome, are thought to primarily happen through passive cytosine methylation dilution upon DNA replication in zebrafish [73, 71]. However, a small subset of highly conserved enhancers of key developmental genes in zebrafish, have been reported to be actively demethylated in zebrafish through TET-dependent mechanisms [74], explaining the detection of low global levels of 5hmC. Indeed, single TET mutants only display mild phenotypic defects in zebrafish [74], similar to single TET mutant-related phenotypes in mouse [17]. However, double (*tet2-tet3*) and triple *tet* mutants exhibit severe defects and mostly embryonic lethality in zebrafish [75] and mouse (*Tet1;Tet2* double mutant instead) [17], and completely abrogate the low methylated state at key developmental enhancers [74]. Although studies on 5hmC levels and Tet expression in zebrafish embryos are somewhat contradictory, preliminary results suggest the conserved importance of TET enzymes during vertebrate embryogenesis in possibly priming DNA demethylation in mouse and in removing DNA methylation at key developmental enhancers, in vertebrates.

Plants, however, do not seem to possess the machinery to oxidize 5mC. Instead, DNA glycosylases and other proteins directly participate in the active removal of 5mC bases in a DNA replication-independent manner. Active DNA demethylation in plants has also been associated with developmental processes (pollen, fruit ripening, pollen tube formation, among others) [76, 2].

1.4 DNA methylation and transcriptional control

1.4.1 5mC readers

In addition to DNA methylation "writers" (DNMTs) and "erasers" (TETs), 5-methylcytosine is sensed by "readers", the cellular actors capable of interacting with methylated or unmethylated CG dinucleotides.

The establishment and maintenance of distinct cellular identities rely on the transmission of specific mC patterns upon cell division and differentiation, as well as on the transcription factor (TF) repertoire [77]. Furthermore, methylated cytosines can affect the binding of TF themselves, which can as well, upon DNA binding, modify methylation levels at TF binding sites via the recruitment of DNAme erasers/writers (Fig. 1.4 and ref.[78]).

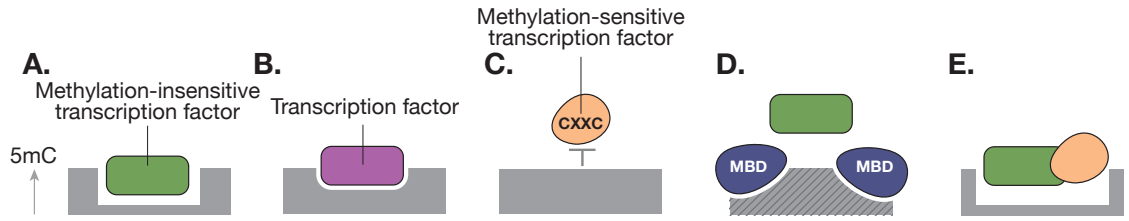


Fig. 1.4 Interaction between cytosine methylation and DNA-binding factors. Interplay between DNA-binding proteins (such as transcription factors, TFs) and DNA methylation levels. **A.** A methyl-insensitive TF binds to DNA in a sequence-specific manner, resulting in lower methylation levels. **B.** A TF specifically binds to methylated binding site. **C.** A methyl-sensitive TF (that contain CXXC-domain in some instances) is blocked from binding by high levels of mC. **D.** A methyl-CpG-binding domain-containing protein (such as MBD1 or MeCP2) specifically recognises methylated CpG-rich DNA sequence (shading), as part of DNAme-mediated repression. **E.** A methyl-insensitive TF binds to methylated binding site, resulting in lowered mC levels, therefore allowing methyl-sensitive proteins to bind. Figure adapted from ref.[78].

In humans, a recent large-scale study investigating the impact of DNAme on TF binding affinity highlighted three major classes of TFs: the first class of TFs preferentially binds methylated CpG (34% of all TF studied), the second class interacts with unmethylated CpG binding sites (23%) and the third class is not affected by DNAme at binding sites (33%). Interestingly, TFs that bind methylated CpG sites (such as homeodomain TFs) are particularly associated with developmental processes (Fig. 1.4b), while TFs targeting unmethylated DNA sequences are enriched for processes related to cell differentiation and proliferation (Fig. 1.4c). Other TFs, such as forkhead box proteins (FOX), appear to be unaffected by methylation state at binding sites (Fig. 1.4a) [77]. Similar work has been carried out in plants, revealing that more than two-thirds of TFs are methyl-sensitive [79].

Methyl-binding-domain (MBD) containing proteins can specifically bind methylated CpG sequences, thus competing off TFs and participating in the repressive state (Fig. 1.4d) [78, 80]. Inversely, some proteins preferentially binding unmethylated CpG can harbour a conserved CXXC (Cys–X–X–Cys) domain and have been reported to interact with histone modifying enzymes, ensuring a stable unmethylated state (Fig. 1.4c) [78, 18].

Furthermore, some TFs are known to participate in locally modifying methylation levels at binding sites. For example, the transcription repressor CTCF, upon binding to methylated CpG, can reduce the methylation level at binding sites via the recruitment of the TET machinery (Fig. 1.4a). Inversely, the binding of other TFs (such as NR6A1) to unmethylated CpGs can bring DNMTs to increase methylation levels at binding sites [80].

Variable TF-DNA interactions are thought to be an important source of phenotypic variation [81]. Changes in methylation levels at TF binding sites can participate in such differential interactions. Genetic variation could sometimes contribute to altered DNA methylation levels at *cis*-regulatory element sequences [81]. A single nucleotide mutation in a particular TF gene or at a TF binding site can result in the TF not being expressed or in impeded TF-DNA interaction, both possibly resulting in altered methylation levels at one particular locus [78]. One recent study highlighted that SNPs disrupting TF binding affinity to bind were significantly associated with altered DNAme levels, suggesting that TFs could participate in shaping the methylome themselves [82]. In turn, this might greatly affect gene expression and possibly such genetically driven methylation patterns could be mitotically and meiotically transmitted. Therefore, the genetic basis of DNAme variation, especially at *cis*-regulatory elements, can participate in global methylome divergence [83]. Whole genome bisulfite sequencing could highlight differentially methylated TF binding sites that may lead to phenotypic variation.

Additionally, DNMTs have also been reported to be important for neural crest development, finely orchestrating transcription silencing of key genes (mainly TFs) in a spatially and temporally specific manner [84].

In summary, the binding of many TFs can be affected by differential methylation levels at CpG sequences. Some proteins preferentially bind to either unmethylated or methylated binding sites, in either a sequence-specific or non-specific manner. Interestingly, upon binding, TF can also participate in modifying the methylome landscape by recruiting other eraser/writer proteins. The complex interplay between cytosine methylation and TF warrants future and exciting work.

To conclude, DNA methylation erasers, writers and readers all play a major role in establishing and maintaining methylome patterns essential for organismal development and cellular identity.

Many transcription factors show methyl-specific binding specificity and activate transcription upon promoter demethylation [78] – processes particularly relevant during embryogenesis to prime DNA methylation-mediated cell-type transitions and are therefore involved in cell differentiation [17, 57]. In mammals, many pluripotency factors, homeobox proteins and cellular patterning factors contains methyl-specific binding (MBD) motif, whose activity is tightly orchestrated by TET- and DNMT-mediated CG methylation changes [17, 80]. TET enzymes might be required as well to establish a pluripotent state, leading to cellular preprogramming [85, 70]. Moreover, some transcription factors have been shown to recruit TET enzymes to enhancers for demethylation during cell-type reprogramming [17]. This role for

TFs to mediate methylation and demethylation at genomic loci has been reported extensively – this suggests that not only are TFs able to bind DNA sequence in a methylated matter, they can also participate in the establishment of DNA methylation landscape themselves [80].

1.4.2 DNA methylation and alternative splicing

Cytosine methylation in other genomic contexts, such as over gene bodies, might possibly exert biological functions, however more experimental evidence is required to support this [78]. Meanwhile, a growing amount of experimental evidence has shed light on a possible implication of cytosine methylation in alternative splicing [86]. In particular, exon junctions show increased methylation load in vertebrates. Methylation over a gene body is a very conserved feature of DNA methylation across kingdoms and has been observed in most vertebrates, invertebrates, plants and fungi at variable levels [78, 41], possibly exerting conserved biological functions.

At least two main DNAme-mediated mechanisms have been reported to be involved in the regulation of alternative splicing. The first one involves a tight interaction between the methyl-sensitive CCCTC-binding factor (CTCF) and methyl-CpG binding protein 2 (MeCP2) with RNA polymerase II (Pol II), resulting in altered kinetics of Pol II and consequently in alternative splicing [86]. In the case of CTCF, its binding interaction with unmethylated CG dinucleotides at the exon-intron junction leads to the inclusion of exons on the elongating nascent mRNA through the physical pausing of Pol II, independently of histone marks [87]. The second mechanism involves MeCP2, a methyl-sensitive reader, binding specifically methylated CG dinucleotides. Binding of MeCP2 to methylated exon-intron junctions promotes the inclusion of particular exons, through tight interactions with histone mark modifiers HDACs (HDAC).

In summary, intragenic methylation can either cause the inclusion or skipping of specific exons in the nascent mRNA by mostly altering Pol II kinetics and recruiting histone modifiers. Even at the level of gene bodies, the roles of DNA methylation are contradictory, as higher methylation levels can lead to alternative splicing depending on the cellular machinery expressed and present at one particular genomic locus. Altered DNA methylation patterns might therefore have important regulatory functions, in particular during development, ensuring cell-specific transcription of distinct isoforms [86, 78]. Yet many unknown mechanisms might exist, involving an interplay between DNA methylation and histone marks. Further experimental work, including comparative analysis in different organisms, is required to fully comprehend the potential roles of DNA methylation in gene bodies.

1.4.3 Mammalian-specific functions of DNAme

Another striking example of the role of DNA methylation seen in mammals, is sex-dosage compensation through the silencing of one X chromosome in mammalian females [27]. This is brought about by the expression of the long non-coding RNA *Xist*, which in turn coats one X chromosome, resulting in long-lasting epigenetic silencing subsequently via the recruitment of chromatin modifying enzymes and the DNA methylation machinery at the X chromosome (see section 1.1.2).

Another mechanism in mammals mediated by cytosine methylation is genomic imprinting, whereby certain genes are expressed in a parent-of-origin-specific manner via the specific epigenetic silencing of one parental allele [88]. Some loci are differentially methylated in the two parental germ lines and are not reprogrammed during the first wave of demethylation taking place in early embryogenesis after fertilisation (see Fig. 1.5 and section 1.5). This results in the mono-allelic expression of genes in their vicinity. During primordial germ cell (PGC) differentiation, imprinted promoters are then reset (during the second wave of demethylation) to be then re-established in a oocyte- or sperm-specific manner during germ cell development. In the developing oocytes, DNAme at imprinted promoters are thought to be established by both DNMT3A and mammal-specific DNMT3L [17, 88]. While maternal imprints are all located in promoter regions, the three paternal imprints are intergenic. Mammalian imprinting could represent an intragenomic evolutionary conflict between parental alleles, when some traits can differentially affect the inclusive fitness of the parents [88]. For example, the expression of the gene *Igf2* (insulin-like growth factor 2) promotes embryo growth, which leads to increased paternal inclusive fitness (bigger offspring), however a larger embryo might be disadvantageous to female mammals, which results in this gene being maternally imprinted and repressed to negatively regulate growth [88].

1.4.4 Cichlid-specific functions of DNAme

Knowledge of DNA methylation or epigenetic mechanisms in cichlids is poor, with only two studies reported so far.

First, Chen and colleagues first provided the characterisation of the methylome of the gonads of the riverine cichlid Nile Tilapia, using methylated DNA immunoprecipitation (MeDIP) [89]. In addition to characterising the methylome at a reduced representation of loci genome-wide, the authors could correlate low methylation levels at promoters with higher expression levels of respective genes. Although this study revealed some conserved roles of

DNAme in cichlid fish, it lacks a genome-wide, single-base resolution and was limited to only one tissue.

Interestingly, the second study investigated the role of DNAme in social dominance in the cichlid species *A. burtoni* [90]. In cichlids, male social status can rapidly shift in response to environmental stimuli (such as social interactions), resulting in striking and fast phenotypic alterations, such as behavioural changes, pigmentation and differential transcription of genes in the brain and reproductive system. Strikingly, the authors could identify differential changes in DNAme in the brains of male dominant vs non-dominant cichlids, suggesting an implication of DNAme in mediating or being correlated to social dominance, in particular in genes responsible for such social shift. It is tempting to put forward an epigenetic basis of social dominance, likely to promote rapid and reversible changes in gene expression. However, functional analysis is required to fully characterise and test this hypothesis, and to investigate the inheritance of such modified patterns.

Altogether, little is known about the role and function of DNA methylation in cichlids. A high degree of conservation with other vertebrates of the mechanisms and roles associated with cytosine methylation is to be expected. However, and in particular, any epigenetic basis for the phenotypic diversity among cichlid species and populations is understudied, especially in the light of their extraordinary low sequence divergence coupled with high phenotypic plasticity.

To conclude, the roles of DNA methylation are multi-faceted and, in some cases, paradoxical and contradictory. The genomic context of DNA methylation seems to mostly dictate the functions. Although DNA methylation has been lost in some organisms, TE repression mediated by DNA cytosine methylation seems to be one of its most conserved features across kingdoms, suggesting a possible ancestral common mechanism against genomic exogenous DNA [34]. In vertebrates, DNA methylation at promoters, enhancers and gene bodies has been reported to exert important regulatory functions, in particular pertaining to gene expression and alternative splicing [17, 67]. The high conservation of factors involved in DNA methylation as well as any loss of functions generally resulting in embryonic lethality attest to the importance of such mechanisms in vertebrates.

In vertebrates, DNA methylation as well as demethylation exert essential regulatory functions during development, required for the regulation of chromatin states of key tissue-specific regulatory of terminal differentiations [17, 1, 78, 51]. From embryonic cells to fully differentiated cells, roughly 20% of all CpGs will undergo significant modification of their methylation levels. Furthermore, there is also an important crosstalk between cytosine methylation and histone modifications, participating in cell-specific chromatin states and

gene transcription [25]. In addition to its role during development, a tight regulation of DNA methylation is required for cell reprogramming, pluripotency and homeostasis [70]. The activity of TET and DNMT enzymes are known to be regulated in response to environmental stimuli, which might in turn alter transcriptional networks.

1.4.5 DNA methylation and silencing of transposable elements activity

Overview

Another important biological process highly regulated by epigenetic mechanisms, in particular DNA methylation, pertain to transposable elements (TE) and repeats, or exogenous, parasitic DNA elements [34, 20]. TEs compose a large fraction of eukaryotic genomes [36]. Although more complex organisms do not necessarily possess larger genomes (the so-called C paradox), there is a direct correlation between genome size and the genomic load of TE [36, 91], in particular in sexual organisms [91, 19]. In bacteria, most exogenous DNA, such a viral fragments, are left unmethylated to be selectively recognised by the host and degraded. In some modern bacteria and eukaryotes, this system has evolved differently, where both host genome and TE sequences are heavily methylated resulting in gene regulation and genome expansion due to TE sequence accumulation. This might have come at a cost for the host as the genomic retention of TE sequences in the host, instead of their degradation, has resulted in an arms race evolution to control their deleterious activity. In some instances, TE expansion might be advantageous to the host as it can allow for some degree of genome 'instability' and possible co-option of TE functions, such as novel genomic regulatory elements [91, 92].

Diversity of TE sequences

Transposable elements are very diverse in their sequences, yet they all share the ability to transpose in the host genome, exploiting host cellular functions to replicate [36, 92]. Two major classes of TEs have been identified, primarily based on their distinct mode of transposition. While DNA transposons (class I) employ a cut-and-paste mechanism via a DNA intermediate, retrotransposons (class II) are inserted in the host genome through an RNA intermediate in a copy-paste fashion [36, 20].

Interestingly, the TE landscape is usually distinct in different taxa or even between species, reflecting the high evolvability of such exogenous elements. In mammals, retrotransposons compose most of the repeat genome (which makes up 40-50% of total genome), in particular with long and short interspersed nuclear element (LINE and SINE, respectively). In other vertebrates, such as in bony fishes, DNA transposons are predominant [93]. LTR (long terminal repeats), part of the retrotransposon class, are the most numerous TEs in plants [94]. The overall TE content in eukaryotic genomes is very variable: from 40-50% in mammals, 25-50% in teleost fishes, <10% in birds, <20% in invertebrates and to 10-80% in plants [93, 20, 94]. Furthermore, the genome size in eukaryotes is directly linked to the TE content [36].

Host defence mechanisms against transposons

Despite the high propensity of eukaryotic genomes to contain TE elements, only few transposons are reported to be active in the host in general. In fact, the activity of most TEs are either under the tight control of the host or have become inactive due to "natural" mutations (resulting from spontaneous mutations, genetic drift, or as a consequence of host recombination events) [20].

Metazoa have evolved powerful tools to counteract the deleterious mutational activity of TEs, giving rise to an evolutionary arms race between TE sequences and the host genome. In the 1990s, Bestor and colleagues observed that in eukaryotic cells cytosine methylation was primarily localised in genomic TE sequences and thus postulated that DNA methylation, originally derived from a bacterial immune mechanism, might have evolved to primarily promote TE silencing [34, 95]. The wide conservation of cytosine methylation in TE sequences [19, 41], from plants, invertebrates to mammals, hints at conserved mechanisms of genome defence against these 'intragenomic parasites'.

Nevertheless, genomic insertions of TEs are thought to be usually fixed in one population by genetic drift, as most of them have a neutral and mildly deleterious impact on host's biological functions [20] – the principle of the host-virus interaction, where killing the host would be as detrimental to TE fitness. Any major deleterious impact resulting from TE activity, such as a *de novo* TE insertion disrupting a gene, would be eliminated from one population by purifying selection [92]. Interestingly, TE insertions in the genome are not random, rather they tend to occur in genomic regions of the host that would be favourable to their replication, while minimising deleterious effects in the host - for example, upstream regions of genes transcribed by RNApolIII (Pol III), while coding regions will be depleted of TE sequences [96].

DNA methylation-mediated silencing of TE activity

One other crucial function of DNA methylation, which might be one of its most conserved features, is to methylate transposon promoters, thus mediating their silencing. There is a constant evolutionary arm race between transposons and mechanisms of defence in the hosts [20, 38].

Vertebrates have developed efficient mechanisms to ensure transposon silencing via 5mC. One important line of host defence is mediated by the class of small non-coding RNAs, the piRNAs, mostly expressed in the germ line of most animals [97, 98]. In male mice, nascent TE mRNAs are recognised and diced by primary genome-encoded piRNAs

into secondary piRNAs by the PIWI argonautes MILI and MIWI, which are then loaded onto another argonaute protein (MIWI2) which is thought to interact with key proteins, such as DNMTs, to promote nuclear TE repression, in addition to their main role in post-transcriptional gene silencing via the degradation of TE mRNAs [20, 17, 98]. The exact mechanism remains unclear. piRNAs are specifically expressed in germ cells in mammals and fish [98]. Both male and female germ cells of zebrafish and the cichlid Nile tilapia have been shown to express 26-30nt long piRNAs together with the piRNA-interacting proteins, Piwil1 and Piwil2, homologues of mammalian PIWIs [99, 100]. Zebrafish piRNAs preferentially target transposon sequences in the genome [101] and might participate in the repression of TE activity through DNA methylation (via a direct or indirect recruitment of the DNAme machinery). In Lake Malawi cichlids, piRNAs are also expressed in the gametes of both sexes (Malinsky *et al.*, unpublished). Interestingly, piRNA-related metabolic processes seem to be under positive selection in Lake Malawi cichlids[1], reflecting fast evolving pathways, with lineage- or species-specific novel functions [102].

Furthermore, in higher vertebrates, a second and important line of genome defence against transposon is mediated by Krüppel-associated box domain zinc finger proteins (KRAB-ZFPs) [103]. KRAB-ZFPs together with the co-factor KAP1/TRIM28 participate in the silencing of sequences derived from TE and possibly in TE domestication as well [20, 104]. During embryogenesis, KRAB-ZFPs recognise TE sequences in a sequence-specific manner, which results in the recruitment of repressive histone modifiers and possibly DNMT3A/B/L, leading to a repressive chromatin state [20, 104]. Importantly, KRAB-ZFPs have evolved in higher vertebrates but are absent in teleost fishes, although present in coelacanth and lumpfish [103] – possible different mechanisms may have evolved in fish, yet they remain unknown.

In summary, in mammals, evolutionary young – and even species-specific – transposons are thought to be silenced first by piRNA-mediated processes, rapidly co-evolving with TE sequences. Meanwhile, the KRAB-ZFP machinery offers a second line of defence to further prevent any transposition and possibly promoting the regulatory potential of some TEs, resulting in gene expression in a tissue-specific manner at specific developmental time points [17, 104]. While *de novo* methylation of TE sequences in mammals is carried out by piRNA and the KRAB-ZFP machinery, the maintenance of methylation at TE sequences is ensured by DNMT1-UHRF1 at the replication forks (see section 1.3.1 and ref. [20]).

As a comparison, in plants (both in somatic and sperm cells in pollen), *de novo* and maintenance of methylation at TE sequences can happen in any sequence contexts (CG,

[1] non-synonymous mutations fixed in a given population, altering the amino acid sequence of a protein, possibly accompanied with changes in functions or molecular interactions.

CHH, CHG, where H = C, A, or T)) via the RNA-dependent DNA methylation (RdDM) pathway. In brief, transcribed TEs are cleaved into 21-22 or 24-nt-long small RNAs by dicer-like proteins that will be then loaded onto two argonaute proteins (AGO1/2 and 4/6), resulting in TE methylation by DOMAINS REARRANGED METHYLTRANSFERASES 1/2, DRM1/2 (homologues of the mammalian DNMT3A) in all sequence contexts and histone modifications [20, 28, 30].

Co-option of TEs and genome evolution

Strikingly, while there is a clear co-evolution of TE sequences and host genomic defence mechanisms, in some instances, novel insertions of TE have led to advantages to the host, through what is called TE domestication [92]. Such genomic instability arising from TE expansion could have played a role in the genomic TE retention observed in eukaryotes, rather than pure TE excision seen in some prokaryotes [37]. The co-option of some TE insertions have been reported to generate novel biological functions that have been selected for in some populations, resulting in increased organismal fitness [92, 2]. This is mostly due to the propensity of TEs to carry regulatory sequences (promoters, splicing sites, poly-A signals) [105, 106].

The mutational impact of TEs on gene expression and phenotypic diversity dates back to the first observations of mobile elements in maize [107]. Active TE transposition has usually a neutral impact on the host, however, they can be responsible for many heritable phenotypes, either beneficial or detrimental to the host [96].

For example, aberrant recombination events leading to large-scale restructuring of the host genome can arise from the repetitive nature of some TE sequences, mimicking host recombination hotspots. Furthermore, DNA transposons, due to their imprecise 'cut-and-paste' transposition mechanism, can lead to gene duplication or re-shuffling [96] and may play a significant role in genome evolutions.

Furthermore, some striking examples of TE-derived genes exist in vertebrates. The recombination activating genes 1 and 2 (*rag1*, *rag2*), essential in generating the massive diversity of antibodies and antigen receptors by somatic recombinations in vertebrate adaptive immunity are derived from DNA transposon sequences [108]. Another example of co-option of TE function is exemplified with the gene *arc*, specifically expressed in mammalian brains and involved in inter-neuronal communication and long-lasting memory formation. *Arc* encodes for viral-like capsids, derived from the *gag* gene present in LTR retrotransposons, enabling intercellular transfer of *Arc* mRNA [109].

In addition to modulating the rate and localisation of recombination events and to duplicating or generating novel genes in eukaryotic genomes, TE insertions can underlie the creation of new regulatory elements, resulting in modified transcription patterns. In cichlids, for example, some unique pigmentation patterns involved in sexual traits in haplochromines have been linked to the specific expression of some genes in the iridophores (pigment cells). This specific gain in expression might have resulted from a TE insertion (namely, AFC-SINE) in the *cis*-regulatory region upstream of these genes [110]. Interestingly, recent genomic analysis of East African cichlids have observed that most TE insertions have occurred upstream of genes, which might potentially hint at novel regulatory functions derived from transposons [53]. In rodents, the insertion of one IAP (intra-cisternal A particle) retrotransposon in the upstream *cis*-regulatory region of the agouti gene is associated with considerable phenotypic variation of coat colours and metabolic changes. Interestingly, differential methylation levels at this TE-derived ectopic promoter directly impact the activity of the agouti gene [111] - such epigenetic patterns of methylation are transmitted in the offspring along with the altered phenotypes in a non-DNA sequence-based manner. As a final example, in plants, the flower symmetry of toad flax is linked to the variable and heritable methylation patterns in the TE-derived promoter of the *Lcyc* gene, resulting in symmetrical or asymmetrical flowers [112].

Altogether, these examples postulate TEs as being major actors in the evolution of eukaryotic genomes, when under tight regulation by their hosts.

Escaping host defence

The constant arms race evolution between host defence mechanisms and transposons has led to the evolution of some extraordinary counter-mechanisms. However such examples are rare, probably due the overall disastrous effects of such counter-defence mechanisms for transposon fitness, as well as for the host, resulting in sterility and embryonic lethality in the latter upon global TE de-repression.

For example, young, lineage-specific TE sequences have been reported to escape the host DNAme-mediated control system [17, 113]. This is seen in rodent-specific TEs which are under tight repression by the protein DNMT3C, exclusively expressed in male mice [62], suggesting fast evolution mechanisms to counteract TE evolvability (see section 1.3.1).

In plants, some transposable elements (*VANDAL21* and *VANDAL6*) have evolved mechanisms to bypass host defence by encoding a protein that demethylates their DNA sequence in a sequence-specific manner, thus enabling their transposition activity [114]. Other bypasses in plants include the retrotransposon *Athila6* which expresses *trans*-acting small-interfering

RNAs (tasiRNAs) that specifically target the 3′UTR of some host proteins involved in TE repression, which in turn results in TE de-repression [115].

Other strategies involved in bypassing defence mechanisms include the retrotransposon EVADÉ in *Arabidopsis*, which encodes a gag nucleocapsid to protect EVADÉ mRNA from host siRNA-based degradation [116].

1.5 DNA methylation reprogramming

1.5.1 Mammals

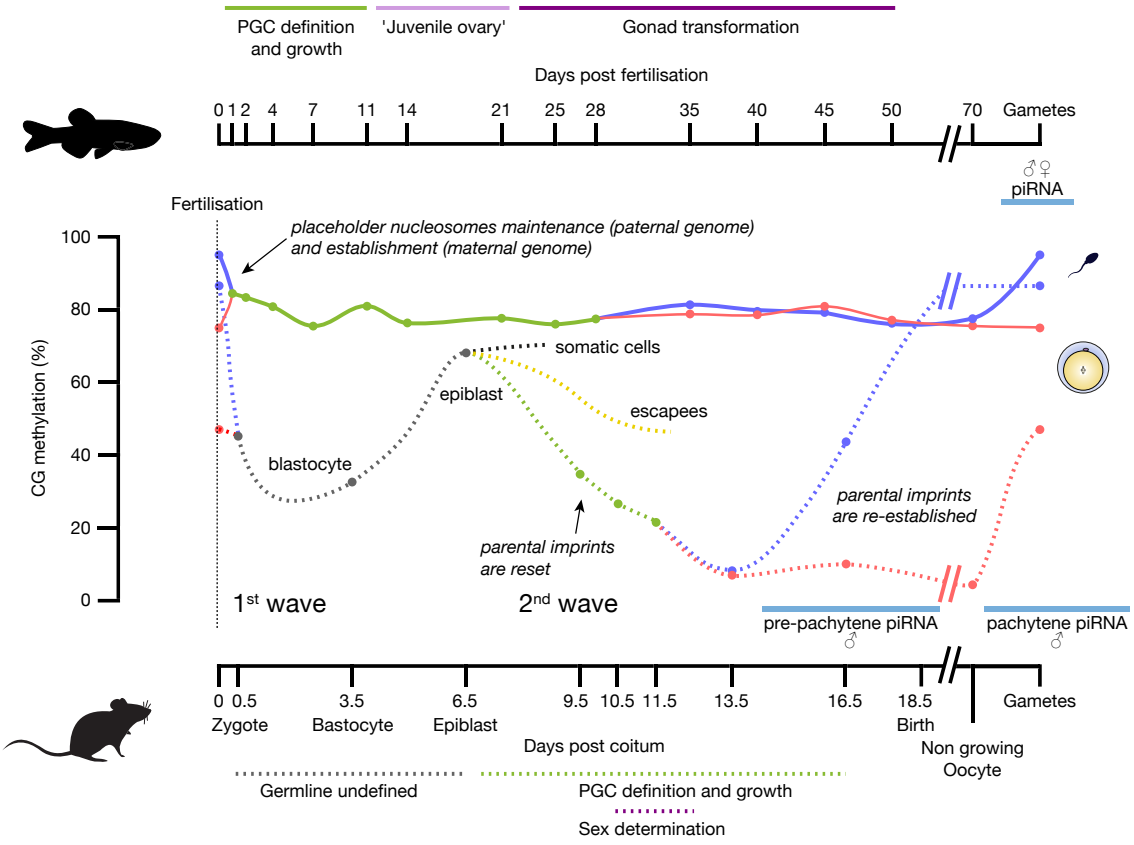


Fig. 1.5 Dynamics of DNA methylation in zebrafish and mouse during development. In striking contrast to mammals, there is no evidence for a global DNAme reprogramming in early embryo and germ lines of zebrafish. Methylomes of primordial germ (PGCs) and somatic cells upon fertilisation resemble sperm patterns of methylation. The oocyte methylome is reprogrammed to match sperm methylome, both of which are virtually similar by mid-blastula stage [73, 3]. In mammals, two waves of global demethylation takes place: first, in embryonic stem cells before reaching blastocyst state, and then, in migrating PGCs. Figure adapted from Ref. [3, 117, 17].

In mouse early development (Fig.1.5), it is well established that two waves of demethylation occur [118, 17]. In brief, the first wave happens in the early embryo, after fertilisation. Both the paternal, first, and later the maternal nuclei are actively demethylated (TET3-mediated conversion of 5mC into 5hmC) [119]. Then, while embryonic cells are dividing to reach the blastocyst stage (embryonic day 3.5 or 6 in mouse and human, respectively), passive demethylation takes place, and the methylome is not maintained upon cell division (passive dilution of DNAme), to reach overall low DNA methylation levels. During that time, the enzyme DNMT1 is excluded from the nuclei to prevent any methyltransferase activity [118, 17]. Imprinted loci, however, withstand this first wave of DNAme reprogramming [88].

The second wave of demethylation in mammals takes place in a subset of cells (the primordial germ cells, PGCs) of the epiblast that will become the developing germ line. First, a brief passive demethylation phase (loss through dilution) is followed by an active and global erasure of DNAme orchestrated by the enzymes TET1 and TET2. Global methylation levels reach their lowest point in PGCs (6-8% mCG), before entering the sex determination phase. This ensures genomic imprinting resetting and X-chromosome reactivation. Interestingly, some genomic loci escape demethylation, in particular young, species-specific transposons, and retain some levels of DNAme, which could give rise to epigenetic inheritance [113, 17]. The primary function of TET enzymes during global demethylation might be to prevent any spurious *de novo* methylation, rather than an active role in the demethylation process *per se* [17, 85].

Finally, the methylome of male gametes is then re-established to a very high global level (ca.80%), following the activation of the DNA methyltransferases DNMT3A, DNMT3B and DNMT3L. This coincides with the production of pre-pachytene piRNAs in male mouse germline [120], a process most probably involved in the repression of transposon activity via the DNMT3 family-mediated *de novo* methylation of their promoters and post-transcriptional silencing of nascent TE mRNA [17]. The genome of female gametes remain lowly methylated (ca.50%), almost exclusively located in gene bodies, until ovulation [17]. Sex-specific imprints are then re-established at that time.

1.5.2 Fishes

In zebrafish, there is no evidence of global DNAme reprogramming in early embryo or developing PGCs (Fig. 1.5). Instead, the methylome is retained in PGCs and somatic cells in its paternal configuration (global high levels of methylation, >95%) in the early embryo after fertilisation [73, 71, 3]. The methylome of the maternal chromosomes (lower

methylation levels, ca.80%) is reset to match the paternal methylome upon fertilisation via active *de novo* methylation and passive demethylation, and both parental methylomes are virtually identical by mid-blastula stage (at the time of zygotic genome activation, ZGA) [73, 71]. Interestingly, in parthenogenic embryos lacking a replicating paternal genome, the reprogramming of the maternal methylome takes place nevertheless, suggesting that the paternal genome might not act as the template *per se* for reprogramming [71]. Rather, the sperm genome bears 'placeholder nucleosomes' (containing histone H2A variant H2A.Z(FV) and H3K4me1 namely), inhibiting DNMT-mediated *de novo* methylation, which ensures the maintenance of hypomethylated state at key promoter regions. The maternal methylome remodelling is a transcription factor-driven mechanism, in that some TFs bind specifically to the placeholder nucleosomes, thus maintaining placeholder nucleosomes on the paternal genome and establishing them on the maternal genome, resulting in maternal methylome reprogramming prior to ZGA [121]. This licences for the maternal contribution to the transcriptional activity needed for the developing embryo. Sperm cells are highly methylated (>95%), in contrast to lower levels of DNAme in fully differentiated somatic cells and female gametes. As PGCs divide, their methylomes become more cell-specific with reduced overall levels of DNAme, in particular at enhancer regions, resembling methylome levels of somatic cells. This is in stark contrast with mammalian reprogramming where totipotency is achieved by global DNAme erasure. It remains however unclear whether such reprogramming processes are conserved in other teleost fishes, which warrants further comparative studies.

To conclude, the degree of conservation of DNAme reprogramming in early embryo and PCGs in vertebrates appears variable. It seems that the process leading to cellular totipotency and epigenetic resetting has been fast evolving and might be unique to some organisms, even at the vertebrate level.

While mammalian genomes (mouse and human) undergo two waves of DNA methylation, globally resetting DNAme patterns to extremely low levels, zebrafish and Xenopus genomes exhibit very high methylome levels, similar to these of male gametes, from fertilisation until cellular differentiation into different cell types. This lack of global DNAme erasure is in total contrast with their amniotic relatives and does not impede with cellular pluripotency. X-chromosome inactivation and parental imprinting are reset upon demethylation in mammals; two processes that are not conserved in fish. On the other hand, the piRNA machinery seems to be conserved in both amniotic and non-amniotic vertebrates and active in PGCs and mature germ cells, with the conserved function of transcriptionally silencing nascent TE mRNA and repressing transposon genes through DNAme. Interestingly, some loci have been reported to escape erasure in mouse and human germ line, and could offer a way to pass DNAme patterns

on to subsequent generation. That process might occur in zebrafish, due to the retention of sperm methylome patterns in somatic cells [122]. Yet, nothing is known about epigenetic reprogramming in cichlid fishes.

1.6 Environmental epigenetics and transgenerational inheritance

1.6.1 Overview

Transgenerational epigenetic inheritance (TEI) can be defined as the transmission of traits or phenotypes to subsequent generations without changes in DNA sequence and even in the absence of the initial trigger [10].

Since Mendel's first observations, genetic information has been thought to be the only process underlying the inheritance of traits across generations, leading to adaptation through natural selection [10]. Interestingly, although genetic differences are the primary source of most of heritable phenotypic variation, recent studies have highlighted the transmission of traits that could not be explained by Mendelian principles of inheritance. Diverse epigenetic factors, discussed above, might also mediate this non-DNA sequence-based transmission of traits. Such mechanisms could also reflect an adaptive response to external stimuli (such as environmental stresses, for example constraints to colonise new ecological habitats), which might impact the phenotypes in subsequent generations and eventually lead to higher fitness [2].

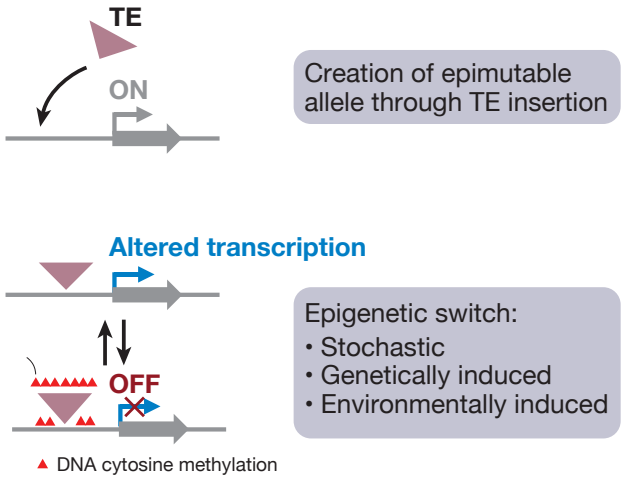


Fig. 1.6 Possible example of epiallele creation. Epialleles are defined as alleles divergent in their epigenetic state while being identical in terms of DNA sequence. The insertion of one TE upstream of one gene might affect its transcription. DNAme levels associated to this TE can modulate the gene expression levels. Such epigenetic patterns can be inherited, possibly promoting phenotypic variation in one population. Altered methylation levels can be the result of environmental stimuli or be stochastic or genetically induced. Figures adapted from Quadrana and Colot [2]

Although the role of methylome variation in the context of organismal development is well established, large-scale population studies investigating DNAme variation and, in particular, in non-model organisms in the context of adaptation and phenotypic diversity remain scarce. It is therefore important to explore the role of the environment in inducing heritable epigenetic variation, and to investigate whether such variation can be independent of *cis*- or *trans*-acting DNA sequence changes.

In the following section, I discuss some examples of TEIs in plants and animals. A great emphasis is put on examples of TEIs mediated by DNA methylation, although other epigenetic mechanisms have been shown to be involved in TEIs as well (in particular in worms and flies, both which lack DNA methylation) [7, 10]. Epialleles, that is, alleles divergent in their epigenetic state while being identical in terms of DNA sequence, can provide a fast and heritable way to change gene expression patterns and chromatin states, while preserving the underlying DNA sequence. The creation of epialleles could follow from TE insertions in the genome (Fig. 1.6), resulting in possible novel *cis*-regulatory regions [92]. As epialleles can modify the transcriptional activity of genes in their direct vicinity [2], they might participate in stable and heritable phenotypic divergence in natural populations. Epialleles, if selected for in one population, could represent a powerful evolutionary device, encompassing a heritable and adaptive response to changing environmental states [123].

1.6.2 TEIs in plants

In plants, there is strong evidence that environmental cues can shape the methylome, possibly influencing developmental processes transgenerationally. In 2014, a study induced DNAme variation in isogenic lines of *Arabidopsis* by using mutant parental lines for methyltransferase genes. The authors could observe many differentially methylated regions (DMRs) that were stably transmitted to subsequent generations without changes in the underlying DNA sequence. Some of these DMRs were associated with adaptive phenotypes, such as root length and flowering time. Furthermore, the authors highlighted that some of these DMRs were also found in natural populations of *Arabidopsis*, suggesting that such DNAme variation could represent good candidates for natural selection [124].

In 2016, the first population-scale study investigating DNAme variation in more than a thousand natural *Arabidopsis* accessions revealed that some epigenetic variation was associated with differential environmental conditions (in particular in TE-related sequences) [125]. The authors observed that some of this natural variation was significantly correlated with changes in gene expression associated with flowering times in particular. This suggests

DNAme may be an actor in generating phenotypic variability in the context of adaptation. Therefore, methylome divergence could possibly underlie the emergence of phenotypic diversification, in particular associated with adaptive traits. Furthermore, some retrotransposons in plants, in particular LTRs, might be especially sensitive to environmental alterations. For example, the LTR *Rider* in tomatoes, and possibly in other plants, has been identified as being an environmental-responsive element, potentially representing a source of genetic and epigenetic variation [126]. Aforementioned, the flower symmetry in toad flax is linked to the expression of one gene controlled by the methylation levels at its TE-derived ectopic promoter. Such methylation patterns, as well as the flower-related traits is transmitted to subsequent generations. However, it is unclear whether such traits can participate in adaptation (increased fitness outcome) or/and could be under selection.

1.6.3 TEIs in vertebrates

Even though clear examples of transgenerational epigenetic inheritance have been demonstrated in plants, extensive work is still required to fully comprehend the relevance of such processes in vertebrates. Nevertheless, whether epigenetic variation is subject to selection and whether it can contribute to adaptation remains elusive. It is important to note that mammals, and possibly other vertebrates, undergo global DNA methylation erasure during germ line reprogramming, which might represent an obstacle to the intergenerational transmission of specific methylome patterns (see section 1.5).

In addition, much of the current work in vertebrates has focused on TEIs induced in laboratory conditions (diet perturbation, early life stress, etc...) using laboratory reared, inbred, model organisms. Although this comes with some advantages (isogenic animals, controlled environments), there are also clear experimental limitations, in that TEIs might be particularly relevant in a natural adaptive context in populations of non-model organisms. It is therefore crucial to understand first if the transmission of acquired traits can be induced (via environmental perturbation) and/or whether some adaptive traits have been selected in some populations and have an exclusive epigenetic basis, independent of DNA sequence variation.

Examples of induced TEIs in response to environmental perturbations (diet, early life stress) have been linked to changes in epigenetic mechanisms (including DNA methylation and sperm ncRNAs), resulting in the transmission of aberrant phenotypes in subsequent generations [127–130, 16, 131, 132]. However, the relevance and occurrence of TEIs in a natural context remain to be explored.

In one species of coral reef fish, transgenerational DNA methylation changes have been observed in response to increased sea water temperatures in the context of climate change, correlated with differential transcription patterns of certain adaptive traits, suggesting physical acclimatisation to heat stress across generations [133].

In rodents, the murine *Agouti viable yellow* locus is a classic and recognised form of TEI, whereby differential methylation states at a TE-derived ectopic promoter result in heritable and reversible changes in mouse coat colours and in metabolic processes transgenerationally, independently of the underlying DNA sequence [134] (see section 1.4.5). In light of this observation, a recent study has explored other possible cases of functional DNAme variation at TE-derived sequences, in particular at IAP elements, an evolutionary recent class of long terminal repeats (LTR) [135]. The authors could identify many IAP regions showing stable DNAme variation, although only very few were associated with changes in gene expression and thus were not functional ectopic promoters, as opposed to the one described in the Agouti mouse model. Furthermore, the variation in DNAme in almost all of these variable IAP elements were found to be reprogrammed during germ line development, in that, the parental DNAme states were reset in the offspring. This indicates that most IAP-related epialleles in the offspring showed DNAme variation regardless of parental methylation levels (i.e. no inheritance of methylation states). This is in contrast with the observation of inherited DNAme levels in the Agouti mouse model, suggesting that the epigenetic inheritance of parental DNAme associated with IAPs might be exceptional, but does not rule out the involvement of other TEs [135].

During human evolution, epialleles could have played a role in phenotypic divergence. Indeed, some differentially methylated regions associated with developmental and neurological processes among others, have been found between modern and extinct humans, as well as between humans and primates, suggesting a possible implication of DNA methylation variation in the evolution of modern human-specific traits [136–138].

To conclude, examples of non-Mendelian inheritance in plants and worms are well established, however such examples in vertebrates remain scarce. Moreover it remains unclear whether such TEI would bear any adaptive advantages in a natural context, and whether they could be subject to natural selection. The examples of flower symmetry in toad flax and the Agouti mouse pigmentation have not been shown to exert any adaptive advantages. It is therefore crucial to investigate the possible epigenetic mechanisms underlying the variation and inheritance of adaptive traits in natural populations of non-model organisms, which are more likely to rely on TEIs to promote phenotypic adaptation in response to environmental cues.

Due to the extraordinary high phenotypic variation and low genetic polymorphism, East African cichlids offer a unique opportunity to investigate the possible roles of genetic and epigenetic variation in the context of phenotypic diversification and species radiation. This is discussed in the following sections.

1.7 East African cichlids

Cichlids (*Cichlidae*) are a family of modern bony fish (*Teleostei*), belonging to the *Labroidei* suborder [139]. Today, cichlid fish are found in Southern India, Madagascar, across Sub-Saharan Africa (Fig. 1.7a) and along the Nile and in the tropics of the Americas. In addition to the geographical distribution, phylogenetic relationships revealed that the cichlid family originated on the Gondwana supercontinent about 150 million years ago at least, prior to Gondwana landmass fragmentation [140, 141].

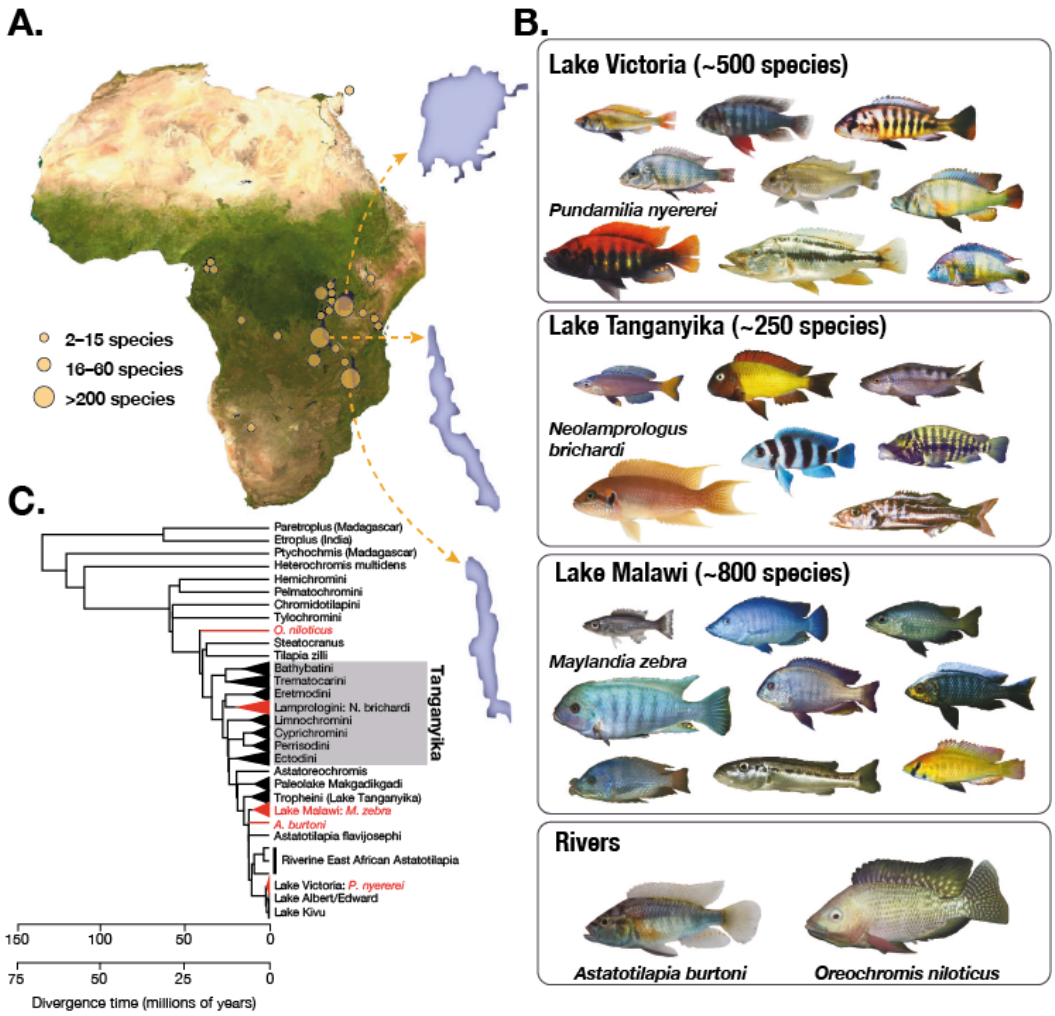


Fig. 1.7 East African radiation of cichlid fishes. A. A map showing the location of East African Lakes. **B.** Examples of major cichlid ecotypes found in East African Lakes. The species with first whole-genome sequenced have their full names shown. Estimation of the number of species shown. **C.** Phylogenetic tree showing the relationships between East African cichlids from different lakes, with timescales representing two different estimates. Figures modified from Ref. [53].

1.7.1 Ecology and anatomy of cichlid fishes

Although East African cichlids display great phenotypic divergence (Fig. 1.7b), common morphological features exist, such as a similar body plan in terms of jaws and fins arrangements, an interrupted lateral line and more importantly a unique arrangement of pharyngeal jaw, functioning as a unitary tooth plate and producing a higher bite force (Fig.1.8a,b and ref[139]). Two sets of jaws (lower and pharyngeal ones) unique to cichlids (and to some *Labrodei* families [independent evolution]) is thought to have participated in new feeding specialisation and in giving access to previously unoccupied niches, possibly contributing to the evolutionary success of cichlids in particular [139, 142].

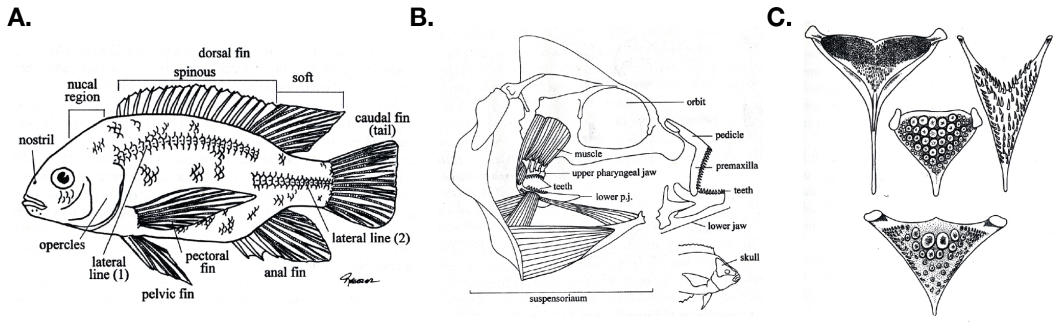


Fig. 1.8 Anatomy of East African cichlid fishes and evolution of the feeding apparatus. A. Body plan of East African cichlids. Note the single nostril and the interrupted lateral line, distinctive features of African cichlids. **B.** Cross section of the head of Cichlidae fishes. Note the unique pharyngeal jaws system. **C.** Detailed anatomy of the upper pharyngeal jaws (bearing specific set of teeth) of an algae eater (top left), piscivore (top right), a specialised molluscivore (middle) and a more generalist molluscivore (bottom). Figures from Ref. [139].

1.7.2 East African cichlid explosive diversification

Speciation is the process by which new species emerge. What dictates and governs the rate of speciation observed in one population remains unclear and has been in the centre of many evolutionary studies. Why some lineages of organisms exhibit explosive species radiation, while other sympatric species do not remains a mystery.

Remarkably, East African cichlids, in particular part of the haplochromine tribes, present a unique rate of explosive speciation, successfully colonising most of the habitats of the Great Lakes and characterised by fast diversification rates [143, 140, 144]. This has led to the emergence of myriads of phenotypes in relatively short evolutionary period of time (15,000 to 100,000 years for Lake Victoria, less than 5 million years for Malawi [see Fig.1.9], 10–12 million years for Lake Tanganyika), making the cichlids one of the most

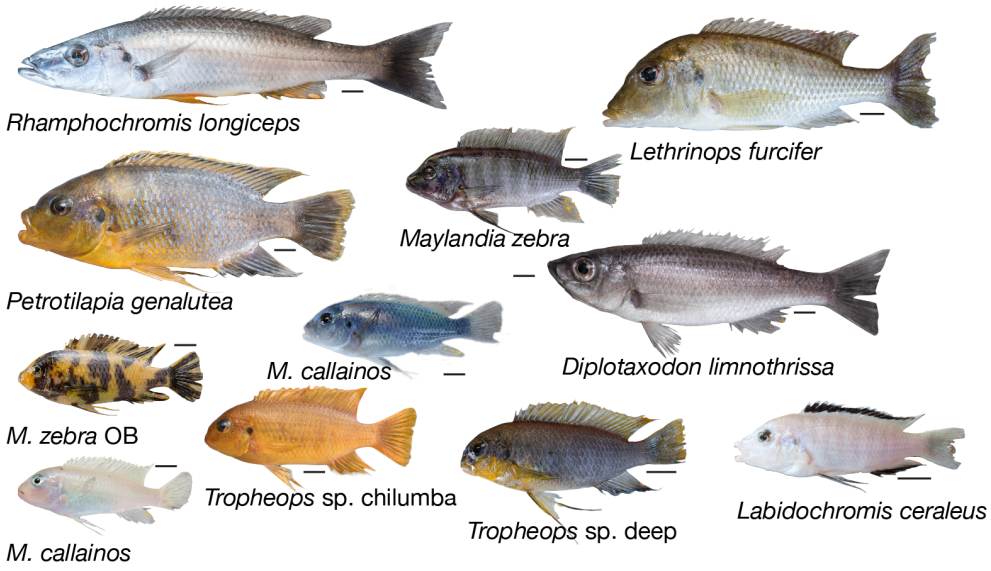


Fig. 1.9 Phenotypic diversity in Lake Malawi cichlid fishes. Cichlids of Lake Malawi show a considerable diversity of phenotypes, in particular in terms of male breeding colours, shapes of the feeding apparatus and overall body length. Only males in full breeding colours are shown (except for *M. zebra* OB). Scale bar, 1cm. Photographs taken by Dr. Hannes Svardal upon collection in the wild. Prof. G. Turner, Richard Zatha and Dr. Bosco Rusuwa helped with species identification. *M.* stand for *Maylandia*.

species-rich populations with no known precedent in the history of vertebrates (Fig. 1.7a-c). Furthermore, this renders East African cichlids extremely valuable as a model to investigate the mechanisms, both molecular and ecological, leading to such an adaptive radiation. Other iconic examples of adaptive radiation have been reported in Darwin's finches, *Anolis* lizards and stickleback fish [144].

More than >1500 different cichlid species have been identified in Lakes Victoria, Tanganyika and Malawi, as well as satellite lakes (Fig. 1.7a,b). Diversification of cichlids is thought to be due the successful adaptation to numerous ecological opportunities provided by these unique ecosystems [145, 143]. Lake Tanganyika, geologically the oldest of the East African Lakes, has multiple major cichlid lineages (tribes), with diversification taking place in several of these (most notably the *Lamprologini*, *Haplochromini*, *Ectodini*). By contrast, the cichlid diversity in both Lakes Victoria and Malawi is exclusively dominated by the *Haplochromini* cichlid tribes (Fig. 1.7a-c and Ref. [144]). Other organisms (such as non-cichlid fish, gastropods and ostracods) have also populated the lake and share similar

ecological habits. However the number of non-cichlid genera is usually 10 times smaller and exhibit a very low level of endemicity in comparison to cichlids (more than 90% of all cichlid species are endemic to Great African Lakes [145])). This intrinsic propensity of cichlids for diversification remains largely unexplained and has motivated a copious amount of studies, with the ambition to decipher the molecular and genomic mechanisms underlying this adaptive speciation unique to cichlids.

Of interesting note, although extremely evolutionary successful, cichlids have also failed to evolve some forms. Other fish families have colonised particular ecological niches where no cichlid species are found (e.g. no fast moving, shoal forming sardine-like cichlid species in any of the East African lakes), which could be a trade-off of the evolutionary innovation of pharyngeal jaws of cichlids [140, 142]. Yet, the high level of endemicity and speciation rate are not observed in most of the non-cichlid fish or other animals of the Lakes [145].

Another characteristic of East African Lakes cichlids is their tendency to evolve similar traits in parallel in different lakes (convergent evolution) [140]. This is of particular interest as similar habitats and natural selection would tend to generate similar phenotypic diversity, possibly suggesting that convergent processes might come into play in parallel, possibly facilitating adaptation in similar ways (Fig. 1.10).

The process of speciation

The emergence of new species, or speciation, in East African Lakes is thought to be gradual and primarily driven by a combination of both ecological and sexual selections [143, 140, 144]. Speciation in East African lakes can happen due to geographic barriers (allopatric speciation), and even in sympatry, when different species coexist in the same habitat [144]. Kocher has suggested three major steps in the cichlids' explosive radiation. The first one involves adaptation to novel distinct ecological habitats, such as rocky vs sandy shores or deep, dimly-lit areas of the lake. This is then followed by specific morphological adaptations, such as the diversification of trophic apparatus (e.g. lips, teeth and jaws). Finally, the last step pertains to sexual selection, such as male breeding colours [143]. This important last stage might contribute to sexual isolation and the emergence of phenotypically distinct populations from the ancestral lineage [147]. Interestingly, many sympatric cichlid species show strong assortative mating – females will tend to mate with related males. Many sexual traits might be important in the early stages of reproductive isolation in one species, such as different courting behaviours, male pigmentations, and even more cryptic phenotypes, such as courting sounds or hormones [140]. This suggests that many traits, both sexual and

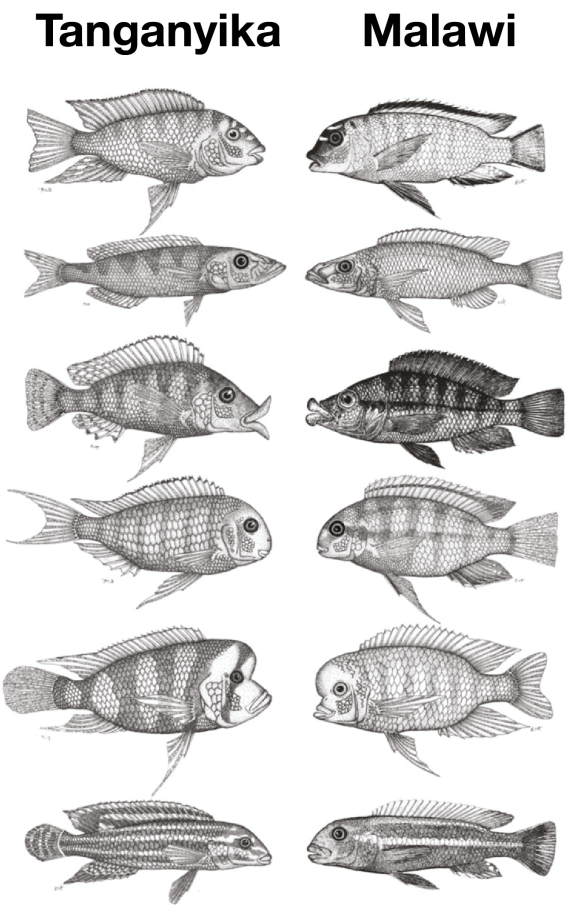


Fig. 1.10 Convergence of traits in cichlids of different East African Lakes. Cichlids of two different East African Lakes (Tanganyika and Malawi) have evolved similar phenotypes in parallel. This evolutionary convergence (or parallel evolution) of traits is still mechanistically not well understood. The six species pairs share specific similar morphologies (top to bottom): rasping jaw morphology, fusiform body, fleshy lips, mbuna habit, nuchal hump and horizontal striping. Figure from Ref. [146].

ecological-related (diet, habitats), may participate in the sexual isolation of one species, thus driving haplochromine divergence.

Moreover, species hybridisation in cichlids is common, with viable and fertile offspring [140]. Hybridisation might have greatly participated in the evolutionary success of East African cichlids, resulting in transgressive phenotypes, not observed in the parental taxa [147]. Hybridisation rate, as well as species diversification, might have fluctuated considerably together with the history of water level fluctuation rises and falls [148, 140, 149], creating possible hybrid zones during periods of drought [147]. Distinct cichlid species which might still be radiating (i.e. have not reached the end of the speciation continuum), can sometimes inter-breed, which makes cichlid taxonomy difficult to undertake and also challenges the very definition of a species. Yet, in general, sister species evolving in sympatry might show

strong assortative mating, suggesting clear species distinction [144]. In this thesis, I am interested in morphologically distinct populations of cichlids, or ecomorphs, having adapted to different ecological habitats, showing sometimes incomplete reproductive isolation, in order to investigate adaptation via phenotypic plasticity – a broad definition of the term species is employed and accepted in this thesis, in that hybridisation might be an actor in the phenotypic diversification process and that reproductive isolation does not solely constitute the basis for the creation of distinct species [11].

1.7.3 Genomic basis of cichlid adaptive radiation

East African cichlids are a text book example of rapid and explosive diversification with successful adaptation to many different habitats. Strikingly, recent studies using genome deep sequencing have revealed particularly low genetic polymorphism among the species of different East African Lakes [53], in particular for the >800 species of Lake Malawi, where the average sequence divergence ranges from as low as 0.1 to 0.25% [102] (see Chapter 2 for detailed description of the evolution of Lake Malawi cichlid flock). Using parents-offspring trio sequencing, it has been estimated that the mutation rate in Lake Malawi cichlids is extremely low (3.5×10^{-3}, or 3-4 times lower than in humans). Moreover, the sequence divergence within a species (heterozygosity) and between species sometimes overlap, suggesting very low sequence divergence and old genetic variation shared between species [102]. The high rate of speciation along with the emergence of many phenotypes is therefore observed in a context of low sequence divergence and mutation rate [102]. This somewhat paradoxical observation appears to be in conflict with the idea that mutations solely would provide the substrate for species adaptation. Recent genomic advances using the unique case of cichlid rapid adaptive radiation have unravelled a more complex picture. Ancient hybridisation is thought to play a crucial role in speciation, whereby old genetic variation (or standing genetic variation) might provide an important substrate for diversification at a higher rate than *de novo* mutation [150, 102]. Furthermore, other molecular events might have also facilitated cichlid speciation, in particular cichlid-specific gene duplication events [53, 52], potentially leading to numerous novel gene functions – this high level of neofunctionalization of duplicated genes might participate in the evolutionary speciating success [52]. Interestingly, another source of genetic diversity observed in East African cichlids pertains to transposable elements (TE): evolutionary recent insertions of TEs have reported to be preferentially located in the vicinity of genomic coding regions, thus potentially

impacting gene transcription [53, 38] – the potential roles of TE in the evolution of eukaryotic genomes are discussed in section 1.4.5.

This thesis investigates the potential roles of epigenetic mechanisms in facilitating heritable and phenotypic diversity in cichlids. Furthermore, this work aims at expanding our comprehension of the processes involved in species diversification and adaptation.

1.8 Overall aims of this thesis

Cichlid species of Lake Malawi represent one of the most extensive recent vertebrate adaptive radiation. Over a relatively short evolutionary time, hundreds of cichlid species have emerged, colonising most of the ecological habitats of the lake. Interestingly, recent population-scale genetic studies have revealed that the extremely phenotypically diverse cichlids of Lake Malawi were characterised by overall low sequence divergence combined with low mutation rate, suggesting that other molecular mechanisms might act in concert and participate in phenotypic plasticity and adaptation.

This thesis studies the natural epigenetic variability and plasticity in response to different environmental conditions in different wild-caught and tank-reared populations of East African cichlids using whole-genome bisulfite sequencing and RNA sequencing of liver tissues. Adaptation to different diets in Lake Malawi cichlids may manifest as distinct hepatic functions and thus different liver methylome may facilitate adaptation to different sources of food.

This thesis is organised in five chapters: the introduction (**chapter 1**), three result chapters and finally the conclusion (**chapter 5**). Each result chapter contains its own discussion section. The chapters dealing with the results are divided as follows:

- **Chapter 2** aims at identifying, quantifying and localising liver DNA methylation variation at conserved underlying DNA sequences in three biological replicates of five Lake Malawi species characterised by distinct trophic adaptations. To address this aim, whole-genome bisulfite sequencing of liver and muscle tissues was performed and sequencing data were mapped to the same reference genome. Once characterised, the inter-species variation in liver methylomes was correlated with gene expression in the same individuals in order to investigate the epigenetic basis for phenotypic variation (i.e. differential gene expression). Furthermore, this section describes species-specific DNAme variation that is tissue-independent, which might reflect considerable methylome divergence established very early on during embryogenesis with possible important species-specific developmental differences.

- **Chapter 3** explores and quantifies the epigenetic diversity in the satellite crater lake of Lake Massoko, home to a population of cichlid species, *A. calliptera* sp. Massoko, composed of two distinct ecomorphs in early stages of speciation. Radio-isotope labelling has confirmed adaptation to different sources of food, which might play an

important role in phenotypic diversification. Liver methylomes of a large natural population of the two ecomorphs are extensively characterised. Finally, the methylomes of Lake Massoko cichlids are compared with these of Lake Malawi cichlids to investigate any common basis for epigenetic variation and identify genomic regions showing high epigenetic variability in both lake systems. Shared variation could participate in the divergence of adaptive traits in the two different lakes in parallel.

- **Chapter 4** deals with the plasticity of epigenetic variation in response to environmental perturbation. To address this aim, a common garden experiment was performed, whereby wild-caught *A. calliptera* sp. Massoko and wild-caught riverine *A. calliptera* (closely related to the ancestral cichlid species of Lake Massoko) cichlid fishes were reared and bred in tanks under the same controlled environmental conditions. This experiment identifies and characterises patterns of DNA methylation that are unique to the wild specimens of Lake Massoko, thus possibly underlying phenotypic diversity observed upon colonisation of the deeper ecological habitats of the lake. Finally, the second part of chapter 4 describes an initial and preliminary analysis of the inheritance of liver methylomes in cichlid inter-species hybrids. It also identifies possible transgressive DNA methylation patterns, unique to hybrid individuals, which might lead to transgressive segregation.

In **chapter 5**, I first summarise the main contributions and findings of the research presented in this thesis and then discuss ongoing work and future possible directions. **Appendix A** contains two additional supporting figures. Finally, all the published work done in parallel to this thesis are listed in the **appendix B**.

Chapter 2

The methylome of Lake Malawi cichlids

2.1 Background

2.1.1 Cichlids of Lake Malawi

Lake Malawi is located in the East African Rift in sub-Saharan Africa (Figs. 1.7 and 2.1). The formation of this deep-water lake, main characteristic of Lake Malawi today, dates back to ca.4.5 million years ago, however it was followed by a long period of drought (between 1.6-1 million years ago), resulting in the mass extinction of most of the fauna [151, 149].

The explosive radiation of Lake Malawi cichlids is estimated to have started over the last 800 thousand years [152, 144], coinciding with a wetter climate and reduced lake-level fluctuations [148, 149]. This more favourable climate (still present today) has probably resulted in increased species diversification, likely due to the wealth of ecological opportunities that such a deep-water lake has been able to provide until now. Fossil records have revealed that the first trace of cichlid life dates to 3.75 to 2 million years ago, although extant cichlid species at that time may have been morphologically very different to present-day cichlids, without the characteristic species diversity observed in the lake nowadays [153].

The lake is home to >800 distinct species of cichlids, the majority of them showing a high degree of endemicity [144]. Lake Malawi cichlids can be grouped into seven different eco-morphological clades based on their ecology, morphology and genetic differences: (1) shallow benthic, (2) deep benthic, (3) deep pelagic zooplanktivorous/piscivorous *Diplotaxodon*, (4) the rock-dwelling 'mbuna', (5) zooplanktivore 'utaka', (6) the generalist *Astatotilapia calliptera* and finally (7) the midwater pelagic piscivores *Rhamphochromis* [102, 53].

Interestingly, apart from the unique explosive and adaptive radiation of cichlids, other organisms are present in high numbers and are thriving in Lake Malawi, however most of

them are characterised by a low degree of endemicity. Approximately, 19 non-cichlid fish genera, 21 ostracod genera, 10 gastropod genera and 2 bivalve genera were identified [145]. Cichlids are unique in their propensity to evolve rapidly, resulting in their characteristic high endemicity in the lake.

2.1.2 Genomic basis of Lake Malawi cichlid radiation

Recent advances in high-throughput whole-genome sequencing techniques have allowed for large-scale studies investigating the genetic basis for such an extraordinary phenotypic diversification and successful adaptation.

In 2018, whole genome sequencing of 73 cichlids species of Lake Malawi covering all major lineages has revealed very low sequence divergence overall. From parent-offspring trios studies, Malinsky and colleagues have highlighted a very low mutation rate (in laboratory conditions), which is 3-4 fold lower than the rate in humans [102]. Moreover, in terms of sequence divergence, the cichlids of Lake Malawi are extraordinary genetically closely related [102], with on average 2.0 SNP per kbp (range of 0.1-0.25% sequence divergence – strikingly, the sequence divergence within one species, or heterozygosity, sometimes overlaps with the one observed between species, that is the sequence divergence). This represents a fifth of the sequence divergence observed between human and chimpanzee. Another interesting point is that most of the genetic variation seems to be shared between species of Lake Malawi, in that 82% of heterozygous sites are observed in other species, which is a hallmark of incomplete lineage sorting, e.g. retention of large amount of ancestral genetic polymorphism [102]. Altogether, this suggests that reshuffling of existing genetic variation and introgressive hybridisation even between different eco-morphological groups, more than mutation rate and fast sequence diversification, may participate in the phenotype diversification and explosive radiation [144, 102]. This hints as well at a role of non-DNA sequence-based mechanisms in phenotypic diversification.

Defining Lake Malawi cichlids based solely on sequence divergence has therefore turned out to be difficult due to the low sequence divergence, high rate of gene flow between taxa (hybridisation) and general incomplete lineage sorting [102, 144]. Interestingly, the putative ancestor of the cichlid radiation of Lake Malawi might be a riverine generalist *Astatotilapia calliptera*-like cichlid (ecologically and morphologically), inferred from whole genome sequencing of several *A. calliptera* populations from outside Lake Malawi as well [102]. This suggests that the riverine ancestor of Lake Malawi fish could have represented an reservoir,

populating the river systems and Lake Malawi along the active history of fluctuating water levels. The cichlid flock of Lake Malawi is therefore thought to be monophyletic.

Furthermore, the rate of non-synonymous and synonymous mutations, a signature of positive selection, in the coding sequences of key genes is important in Lake Malawi cichlid flock. Genes involved in transposon repression (via piRNA mechanisms), oxygen transport, photo-transduction/visual perception and immune system might in particular be under positive selection, exhibiting high rate of sequence evolution in one population [102]. These mechanisms may be playing an important role in the successful adaptation of cichlids to many ecological niches, promoting phenotypic plasticity. In different cichlid species of Lake Malawi, shared mechanisms of adaptation to deeper, dimly-lit part of the lakes seem to have involved the same genes related to vision and oxygen transport, therefore highlighting some convergence in the molecular processes possibly facilitating adaptation [102].

In conclusion, defining species in Lake Malawi based on the biological species concept (reproductive isolation), on the sequence divergence (phylogenetic species concept) or even on the different ecological habitats as many species evolve in sympatry (ecological species concept) are mostly not applicable to Lake Malawi cichlids [144].

This therefore represents a unique system to expand our comprehension of the molecular basis underlying adaptation by investigating the epigenetic mechanisms possibly facilitating this exceptional species radiation process and successful phenotypic diversification.

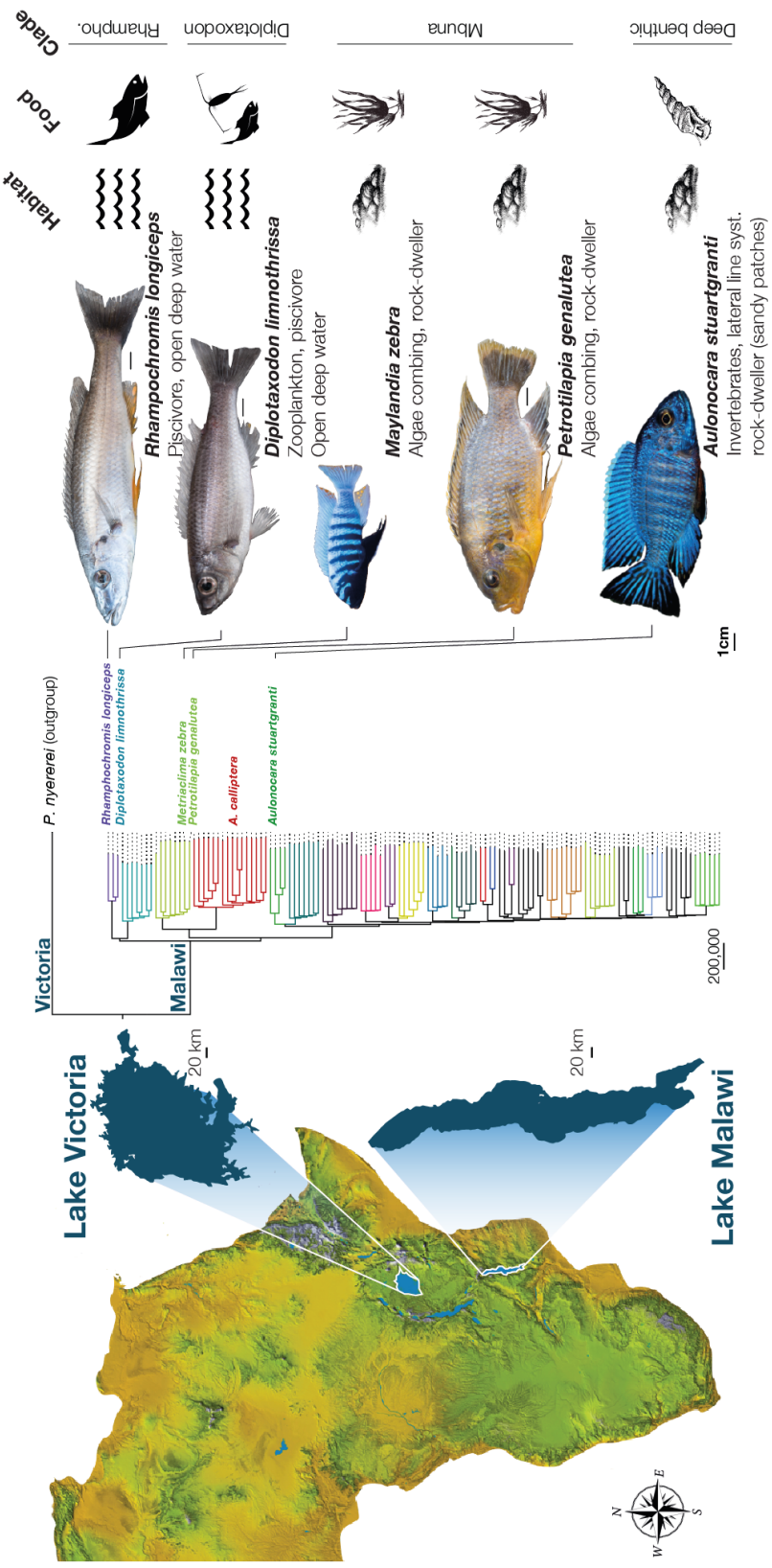


Fig. 2.1 Experimental design - The methylome of Lake Malawi cichlids. The five cichlid species of Lake Malawi used to generate genome-wide liver methylomes at a single-base resolution. These species cover four of the five major ecological clades, and thrive typical and distinct ecological habitats (pelagic or sand- or rock-dwelling/littoral, represented by the wave and rock symbols, respectively) and diet (piscivore, zooplantivore, algivore, invertebrate eater). *Pundamilia nyererei*, a species from Lake Victoria, is used as an outgroup. Phylogenetic tree (left) has been generated from genome-wide genetic differences (produced by Dr. H. Svardal). All photographs of fish to scale, taken directly upon collection in the field (credits: Dr. H. Svardal, except for *Maylandia zebra* and *Aulonocara stuartgranti* [Google]). Satellite map, NASA.

2.2 Sampling and Experimental designs

In order to quantify and characterise the methylome of Lake Malawi cichlids, three wild male specimens of five different species of cichlids, endemic to Lake Malawi, were collected (collection by Prof. G.F. Turner, Dr. M. Du, Dr. M. Malinsky and Dr. H. Svardal). Liver and muscle tissues were placed in RNA*later* [1] upon field collection to reduce fragmentation and degradation of RNA/DNA molecules. Methylome (WGBS) and transcriptome (RNA-sequencing) sequencing data were generated for both liver and muscle tissues (see detailed sampling design, Table 2.4). Detailed methodology is provided at the end of the chapter (see section 2.10).

In this thesis, the primary aim was to characterise the methylome variation in liver of Lake Malawi fish in order to first quantify epigenetic variation in a context of low genetic diversity and explosive radiation, and then to investigate whether some traits related to diet could be explained and transmitted in a non-DNA sequence-based manner in the context of adaptation.

In addition to WGBS for liver tissues, muscle tissues were also sequenced to serve as a control to distinguish between tissue-specific and species-specific epigenetic patterns - one can expect less methylome divergence in muscle tissues. Both liver and muscle tissues are rather homogenous and are mostly composed of a single cell type (primarily hepatocytes and myocytes, respectively). Other cell types might inadvertently be sequenced along with these two tissues (e.g. endothelial cells or cholangiocytes), and might be a source of confounding methylome variability (single cell WGBS could reduce this variability, however represents a much more expensive approach). I assume here that this contamination would be minor. In parallel to this work, the laboratories of Profs. R. Durbin and E. Miska have generated the genome assembly of another Lake Malawi cichlid, *A. calliptera* (available on NCBI), which included transcriptomic data of many different tissues. This offered a way to control that all the samples of this study were correctly labelled and assigned to the right tissue tissues (as a consequence, only one sample had to be removed from analysis [RL, liver], as it was wrongly labelled/dissected as liver, when it was most likely spleen [see Figs. A.1 and A.2]).

Five Malawi cichlid species selected for WGBS

Five cichlid species of Lake Malawi were selected for WGBS and RNAseq. Selection criteria were based on covering diverse species presenting unique and specialised diet-related

[1] RNA*later* is a storing solution consisting of a very high concentration of salts stabilising DNA/RNA molecules and inhibiting enzymatic reactions [154]

eco-morphological adaptations. The five species span four of seven ecological clades of cichlids of Lake Malawi described so far [102]. Furthermore, for all of them, whole-genome sequencing data have already been generated at variable sequencing depth [102].

The five different species of Lake Malawi studied in this chapter are the following:

Rhamphochromis longiceps

R. longiceps, RL, is a pelagic, carnivorous species, found in many habitats of the lake and often solitary. Its long elongated shape, featuring a prominent toothed mouth and adapted to fast swimming in pursuit of preys, makes it a predatory piscivore fish [155, 156].

Diplotaxodon limnothrissa

D. limnothrissa, DL, another carnivorous species, is know to populate deep parts of the lake (up to approximately 250m down, close to the limits of oxygenation). This species, part of the ecological clade of deep benthic species, is particularly adapted to low oxygen levels and dimly-lit environments – recent whole genome analysis revealed adaptation in the visual and oxygen-transport systems [102], with specific morphological adaptations (e.g. larger eye size) [157]. Its diet is mostly composed of zooplankton [155].

Aulonocara stuartgranti

Also known as the Grant's peacock, *A. stuartgranti* is found in the rocky and sandy shores of Lake Malawi. It has highly developed lateral line system, in particular under the jaws, enabling it to feed on small invertebrates buried under the lake ground [155, 158].

Petrotilapia genalutea

P. genalutea, PG, is a rock-dwelling species, one of the most widely distributed cichlids endemic to Lake Malawi. It feeds on algae found on rocks using its comb-like teeth on its large lips. PG belongs to the species-rich Mbuna clade.

Maylandia zebra

M. zebra, MZ, shares many similarities with PG, in that they are both shallow rock-dwellers, exhibiting similar morphological adaptations (protuberant toothed lips), enabling them to sift algae (more precisely, aufwuchs) off their rocky substrate. MZ is also an opportunist eater, and feeds on small invertebrates and zooplankton. MZ was the first cichlid species of Lake

Malawi to have its whole genome fully sequenced in 2014 [53]. MZ belongs to the Mbuna clade.

All the five species are haplochromine cichlids endemic of Lake Malawi and show haplochromine-specific features, such as considerable parental care (maternal mouthbrooders) [143, 139, 140].

2.3 Genetic polymorphism

Firstly, genetic diversity was assessed for each species studied in this thesis.

Overall, genome-wide pairwise sequence divergence amongst all the species of this study is low (Fig. 2.2a,b). Heterozygosity values (sequence divergence within one species population) is around 0.1% for all species (i.e., approximatively 1 single nucleotide polymorphism every 1kbp). On average, the sequence divergence is almost doubled between individuals of different species. In particular, ecologically more related species are genetically less divergent: herbivorous and dermesal species (MZ and PG) are grouped separately from the carnivore species (RL and RL). Interestingly, the demersal, omnivore (principally invertebrates) species (AS) is clustering away from the herbivore and carnivore species (Fig. 2.2b), probably due to its unique trophic and ecological adaptations (in particular its extensive development of the lateral line system). Also, the overall sequence divergence between offspring and parental specimens (parent-offspring trio sequencing) of AS is 0.09%, one of the lowest observed in vertebrates [102]. The outgroup species from Lake Victoria exhibits the highest levels of genome divergence, highlighting Lake-specific genetic polymorphism.

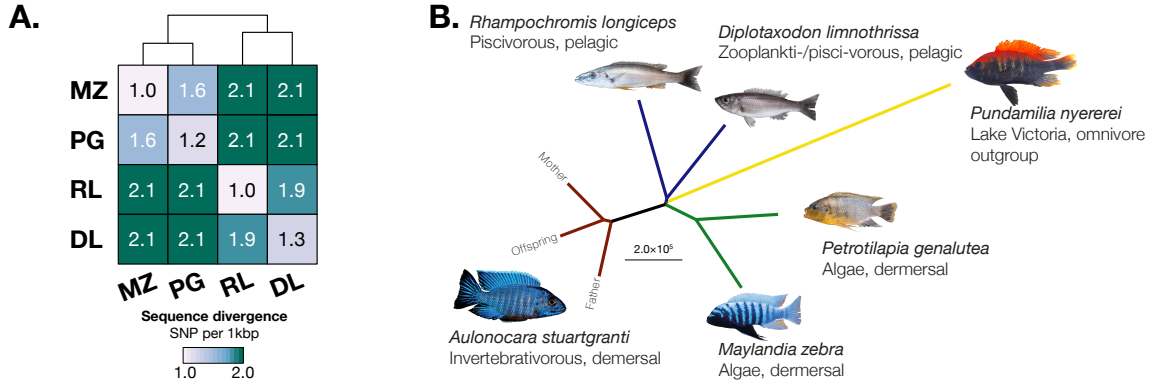


Fig. 2.2 Pairwise DNA sequence divergence within and between cichlid species studied in this thesis. Sequence variability in the fish of this study. **A.** Heatmap generated from sequence divergence (average SNP per 1kbp [callable site only], genome-wide). Within species: heterozygosity, nucleotide diversity; between species, sequence divergence. **B.** Maximum likelihood phylogenetic tree based on pairwise genome-wide sequence divergence. Lake Victoria cichlid, *Pundamilia nyererei*, used as an outgroup species. All values estimated with $\geq$16 individuals per species. Data generated and analysed by Hannes Svardal, published in Ref. [102]

Cichlids of Lake Malawi have been reported to exhibit one of the lowest sequence divergence between the numerous species, with low mutation rates as well – suggesting that other molecular events, such as hybridisation and ancient introgression might be fuelling speciation at higher rate than slow-rate mutations solely [102, 150, 159]. Hence, evolutionary

relationships between Lake Malawi cannot be inferred based on phylogenetic tree only, as some portions of the genome might not even be shared within individuals of one species [102]. It is therefore of great interest to investigate other molecular aspects contributing to speciation, and in particular the epigenetic mechanisms that might facilitate speciation. In order to quantify DNAme variation among Lake Malawi cichlids, only conserved DNA sequences between all the fish studied will be analysed.

2.4 Conservation of epigenetic genes

Proteins involved in DNA cytosine methylation pathways, such as TETs and DNMTs (Fig. 2.3) show a high level of sequence conservation across vertebrates (see section 1.2).

In the cichlid mbuna, the methyltransferase primarily involved in 5mC maintenance, *Dnmt1*, shares 82.2% and 73.3% sequence homology with its homologues in zebrafish and mouse, respectively (Fig. 2.3). Two isoforms have been identified in the genome of *M. zebra* (UMD2a), both for the same gene (*dnmt1*) in the same locus. They are therefore considered as one. The degree of sequence divergence between *Dnmt1* and other DNMT genes is considerable, as *Dnmt1* contains many more domains and is therefore longer in terms of amino acid sequence.

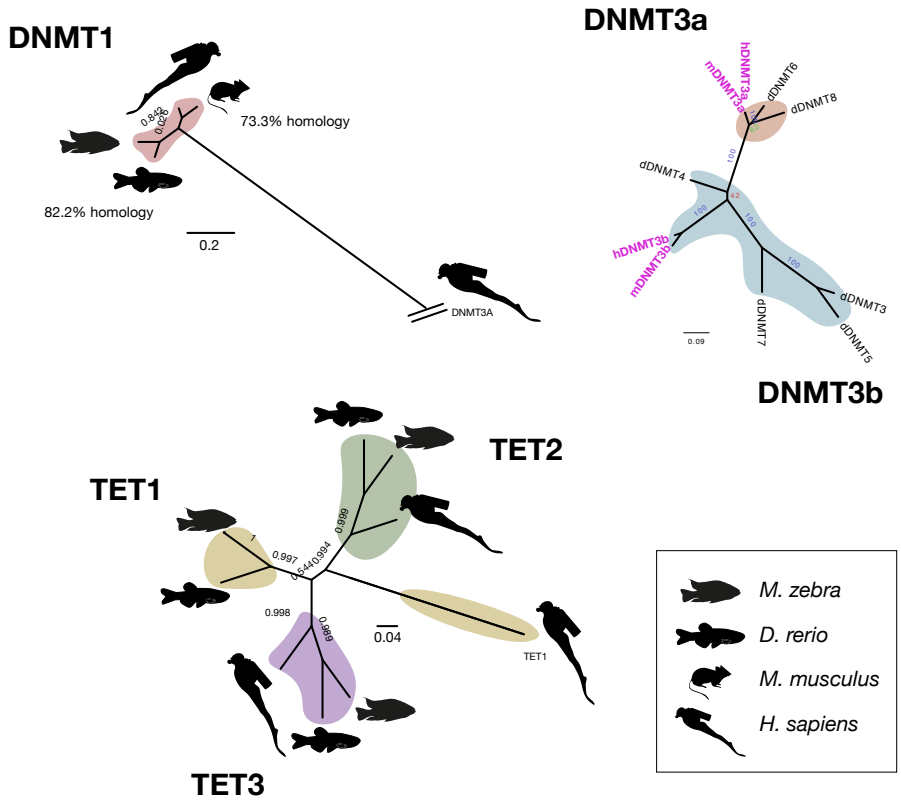


Fig. 2.3 DNMT and TET proteins are conserved in teleost fishes. Phylogenetic trees for the methyltransferase enzymes DNMTs and the methylcytosine dioxygenases TETs based on amino acid sequence similarities in teleost fish (*M. zebra*, mz and *D. rerio*, dr) and mammals (*M. musculus*, ms and *H. sapiens*, hs). Bootstraps are shown along the branches.

In case of other *dnmt* involved in *de novo* methylation (DNMT3), as introduced earlier, there is evidence of many duplication events, leading to the emergence of numerous paralogues in teleost fish (including zebrafish, salmon) [61], and is probably the case in mbuna

too, where numerous isoforms of *dnmts* are found. Six paralogous genes for DNMT3a have been reported in zebrafish, with different sequence homology (2 *dnmt* genes would be more related to mammalian DNMT3b, while 4 *dnmt* genes seem more related to DNMT3a; Fig. 2.3). It is currently unknown whether Lake Malawi cichlids also possess such DNMT3 paralogues, as many isoforms are present in the reference genome. This motivates further and thorough characterisation in *M. zebra*, using the reference genome sequence as well as RNAseq data (reference-free assembly of transcriptome), in order to decipher the level of conservation

The enzymes responsible for the catalysis of DNA-methylation seem to be present as unique, conserved orthologues in mammals, zebrafish and mbuna (Fig. 2.3). Unlike DNMTs, all the three known TET enzymes exhibit a higher degree of sequence conservation altogether. The proteins *tet2* and *tet3* are more conserved between teleost fish and mammals than *tet1* is: human TET1 seems slightly less conserved with teleost Tet1.

As discussed in the introduction (section 1.2), the degree of conservation of DNMT and TET enzymes in vertebrates is high, probably highlighting the biological importance of these pathways. In general, each protein is present in one single copy in mammals and teleost fish with low sequence variation. Interestingly, only proteins involved primarily in *de novo* methylation (DNMT3 families) present a higher levels of variation, in particular with many paralogues in teleost fish (including zebrafish, mbuna, trout, salmon) and even rodent-specific (DNMT3C) or mammals-specific (DNMT3L) proteins. This has generated novel function (neofunctionalization) in particular for this family of enzymes in many vertebrates, such as functions related to mammals imprinting (cofactor DNMT3L)[88] or methylation of species-specific repeats (DNMT3C) [62]. There seems to be a considerable evolutionary plasticity, while keeping the ancestral *de novo* function of Dnmt3, new functions have emerged in fish and mammals. Interestingly, DNMT3 in mammals are thought to mediate repression of evolutionary young TE activity through the recruitment of piRNAs during germ cell differentiation [17]. Such a mechanism might exist in teleost fish, as both piRNAs and argonaute proteins have been reported to be expressed in male and female germ lines in zebrafish [100]. In cichlids, piRNA-related mechanisms appear to be under evolutionary pressure, with sign of positive selection, which could mean novel functions or molecular interactions [102]. A possible DNMT-piRNA interaction mediating cichlid-specific TE repression might be present and warrant further experiment.

2.5 Characterisation of liver and muscle methylomes

2.5.1 Genomic DNA extraction and NGS library preparation

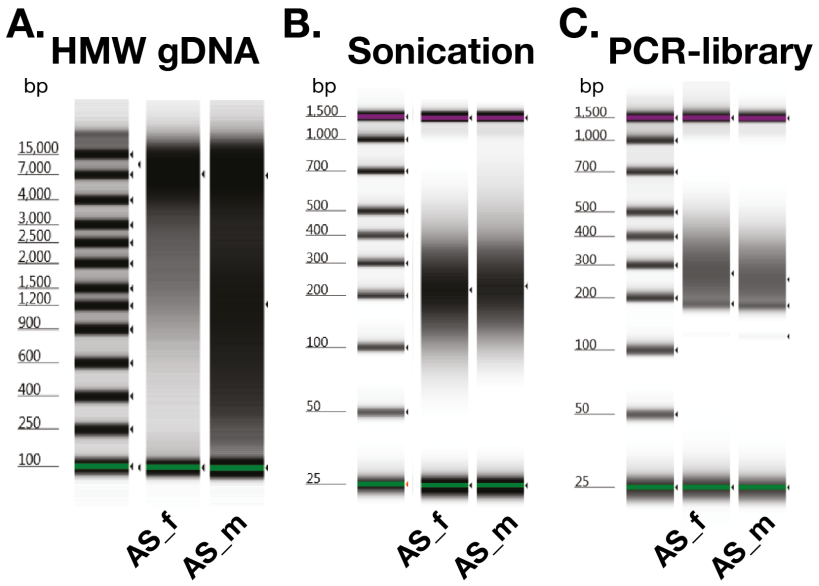


Fig. 2.4 WGBS NGS library preparation. Example of library preparation in the readiness for WGBS with two *A. stuartgranti* (AS) specimens (female and male, AS_f and AS_m, respectively). **A.** DNA profiles of HMW gDNA extracted from whole liver tissues using a silica column-based approach, preserved in RNA*later*. DNA electrophoresis performed using TapeStation. **B.** gDNA from step **a.** was sheared by sonication to fragment sizes of 300 - 400 bp. **C.** PCR amplification of NGS bisulfite-converted libraries.

High-molecular-weight genomic DNA (HMW gDNA) was successfully isolated from homogenised liver and muscle tissues, rapidly preserved in RNA*later* upon dissection in the field. Once extracted and purified, HMW gDNA samples were run on a electrophoresis gel to determine the integrity - one could expect size fragmentation for tissues collected in the field. Some degree of DNA degradation is indeed observed, probably due to natural fragmentation (mechanical or/and enzymatic) due to the limited access to chilled storage (Fig. 2.4a., smear on electrophoresis gel). gDNA samples were then successfully sheared by sonication to enrich for 300 - 400 bp-long fragments (Fig. 2.4b.). Note that sequencing reads are 150 bp-long, sequenced from both ends. It is therefore important to sequence gDNA fragment longer than 300bp to avoid any unnecessary sequencing overlap (i.e. same portion, in the middle, of the DNA insert would be sequenced with both reads twice, resulting in an overall loss of sequencing depth; Fig. 2.34a). After NGS library preparation and bisulfite conversion, fragments are then PCR amplified, enabling library indexing.

2.5.2 Alignment of bisulfite reads and CpG mapping

Alignment to Lake Malawi reference genome

First, sequenced reads of all tissues for all species were mapped to the same reference genome, namely *M. zebra* UMD2a [160]. This allows for direct genomic comparison and investigation of DNA methylation variability at conserved underlying DNA sequence between the five species studied in this thesis. The five species are genetically very similar (see Ref. [102] and section 2.3), with on average 1.5-2 SNP per kbp genome-wide. Thus, I expect high and comparable mapping efficiencies among all the samples. Methylation variability in non-conserved single cytosine nucleotides cannot be inferred and is ignored in this thesis.

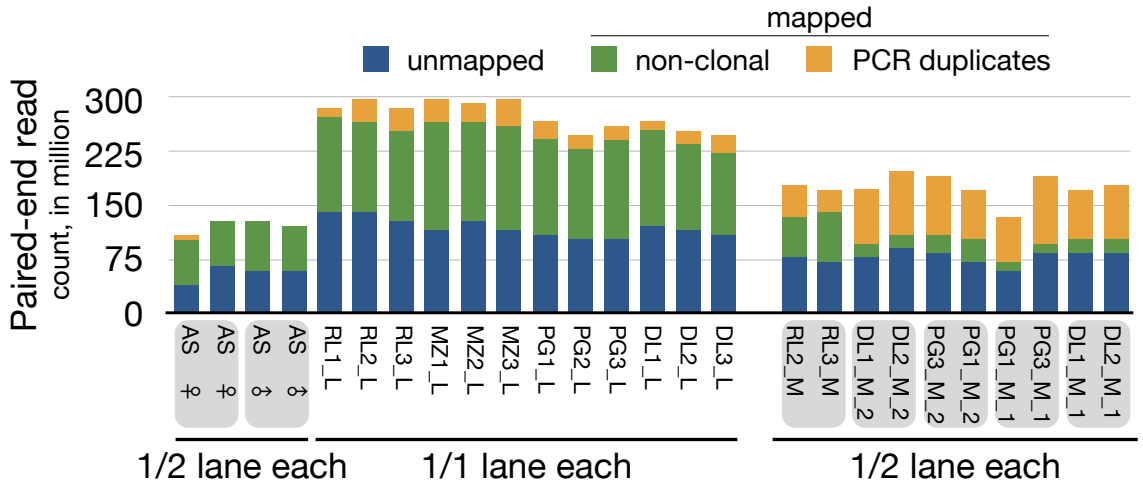


Fig. 2.5 Alignment of bisulfite sequencing reads to the same reference genome. Histogram representing the total number of 150 bp-long paired-end (PE) reads sequenced for each species, for both liver (L) and muscle tissues (M). Some samples were sequenced twice, sharing one HiSeq lane with another sample (2x1/2 lane each, merged later on; same samples boxed out in grey), while some other samples were sequenced once using an entire HiSeq sequencing lane (1/1). Unmapped reads in blue, mapped clonal reads (PCR duplicates) in yellow, mapped non-clonal reads in green. HiSeq4000 (PE 150 bp) for all samples, except for AS (HiSeq2500, PE 125 bp). Biological replicates: n=3 for all liver samples, except for RL and AS (n=2); n=2 for all muscle samples. Refer to section 2.2 for species abbreviations.

Consistently for all samples, more than half of all 150 bp-long paired-end reads were mapped uniquely (56.7%$\pm$2.8%, mean μ $\pm$ standard deviation σ; Fig. 2.5) - note the 3-letter based mapping of bisulfite results in reduced mapping efficiency (see Ref. [161] and section 2.10.3) . As DNA fragments are amplified during NGS library preparation, clonal reads (PCR duplicates) must be collapsed (the first of the clonal paired read mapped is kept) in order to avoid any erroneous inference of methylation state due to duplicated reads. These

clonal reads can represent a large fraction of the sequenced libraries and were filtered out for any downstream analyses (ranging from 2% to 87%, on average 34.5%±31.6% of all mapped reads were PCR duplicates; in yellow in Fig. 2.5). This resulted in a much lower number of useable reads (in green in Fig. 2.5). Muscle samples show particularly a high level of clonal reads. As all libraries went through the exact same protocol, the reasons explaining such high levels of PCR duplicates are unclear - this could be due to an increased level of fragmentation of HMW-gDNA in preserved muscle tissues, or a less efficient gDNA isolation of muscle gDNA (fibrous nature of muscle tissue compared to liver), both resulting in a much lower diversity of DNA fragments in the final libraries.

Bisulfite conversion rate

In parallel, the bisulfite conversion rate for each sample was assessed by mapping all sequenced reads to the spiked-in lambda phage genome (see section 2.10.3). Spiked-in lambda phage genome was entirely unmethylated. This means no C nucleotide should be sequenced if bisulfite conversion was efficient at 100%. Consistently for all samples, the bisulfite conversion rate was very high at 99.2%±0.5%.

Mapping and conservation of all CG dinucleotides in the genome

I then assessed the sequence specificity of methylated cytosines, as well as the sequence conservation of methylated cytosines.

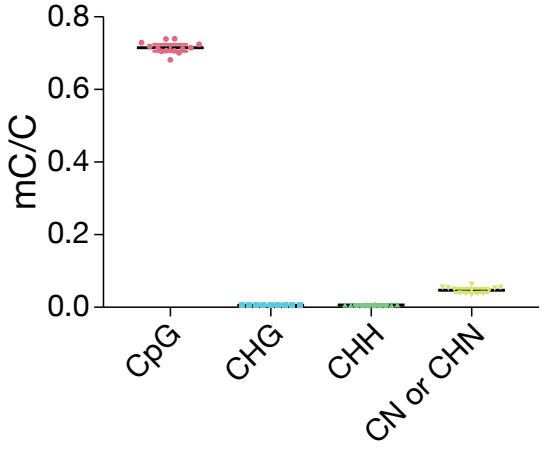


Fig. 2.6 Context-specific cytosine methylation in Lake Malawi cichlid genomes. Fraction of cytosine methylation in different genomic contexts, where H can be either A, T or C, and N can be any bases. Each dot represents one of the specimens studied, only liver WGBS data is shown. Mean ± IC95 shown. *n*=13, see Table 2.4.

Firstly, methylated cytosines were almost exclusively located in a CG dinucleotide context (Fig. 2.6). Cytosine methylation in other contexts is almost not detected, or at levels similar to bisulfite conversion rate (see section 2.5.2). This is a conserved feature of DNA methylation in somatic cells of most eukaryotes, except for plants, where DNA cytosine methylation can occur in any possible contexts [78, 18, 2].

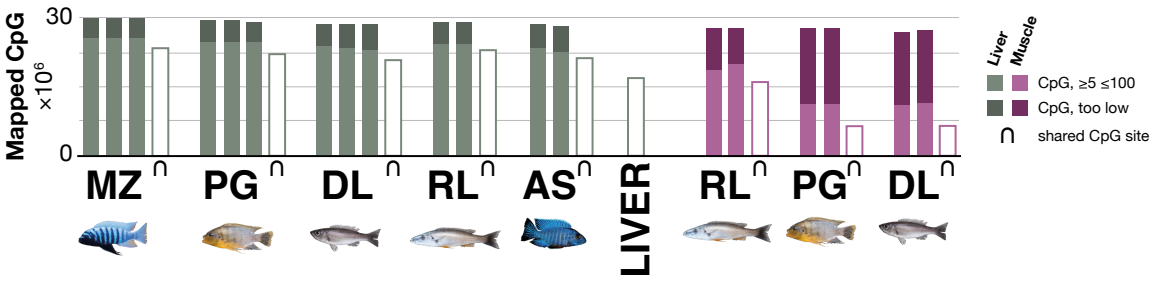


Fig. 2.7 Genome-wide mapping of CpG sites. Histogram of the total number of single CpG sites per individual. Only CpG sites with a coverage of 5-100 non-clonal paired-end reads are considered. Three biological replicates for liver tissues of MZ, PG and DL; two for liver of AS and RL and for all muscle tissues. "∩" represent the number of shared CpG sites per species (intersection). "Liver" represent all the mapped CpG sites, for which genomic coordinates are conserved between all liver samples of all specimens.

The conservation of DNA sequence at CpG sites, is very high, with close to 30 million CpG sites sequenced in all samples, reflecting the high degree of sequence conservation (Fig. 2.7). Any drop in CpG count for a given sample could be due to species-specific SNPs, low coverage and/or technical sequencing error. In order to ensure an accurate methylation status call, CpG sites with a read coverage of fewer than 5 or more than 100 non-clonal reads were filtered out. Sequencing coverage at these CpGs resulted >10 unique sequencing reads per CpG in liver tissues, and >5 in muscle tissues (Fig. 2.9b.). Of note, DMR calling using the software DSS utilises all CpG sites, regardless of read coverage, as this parameter is taken into account to predict CpG methylation status (see detailed methodology below). The high level of clonal reads (PCR duplicates) in muscle tissues eventually resulted in lower sequencing depths and much fewer CpG sites mapped.

2.5.3 Genome-wide characterisation of methylome patterns

Overall

To investigate the variation in methylome between cichlid species of Lake Malawi, only DNA methylation levels at conserved genomic cytosines (i.e. conserved underlying DNA sequences) and located in the specific CG dinucleotide context were analysed.

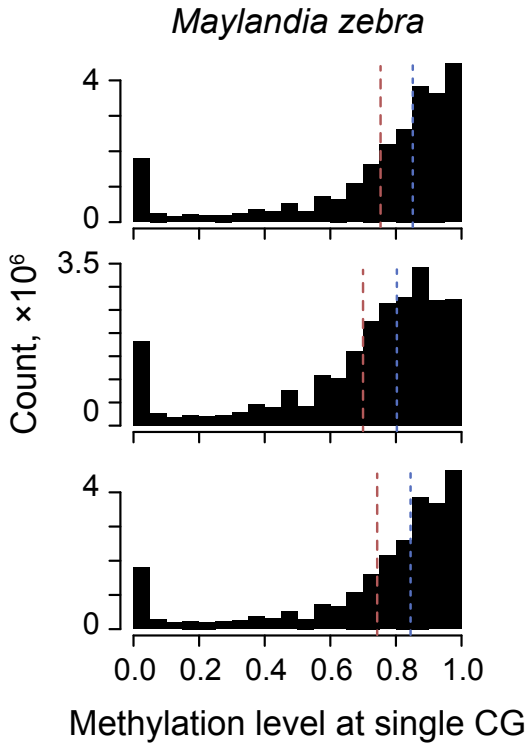


Fig. 2.8 Methylome at single CpG resolution. Frequency of methylation level (mCpG/CpG) at each CpG site genome-wide for three *M. zebra* male specimens. Only CpG sites with a coverage of 5-100 non-clonal reads were plotted and analysed further. 25.6 $\pm$0.14 million CpG sites are plotted. Red and blue dotted lines for mean and median values, respectively.

Overall, genome-wide DNA methylation levels are consistently high and similar in all tissues of all species, with low overall DNAme variation. Methylome at a single base resolution reveals a strong bimodal distribution of methylation levels (Fig. 2.8). In theory, in diploid cells methylation at CpG sites is expected to be binary: either 0% or 100%, and also 50% in case of allele-specific methylation patterns. Intermediate levels could be explained by the fact that bulk DNA sequencing was performed, capturing population- and cell-specific DNA methylation variability. Failure in bisulfite conversion could also cause non-binary methylation inference.

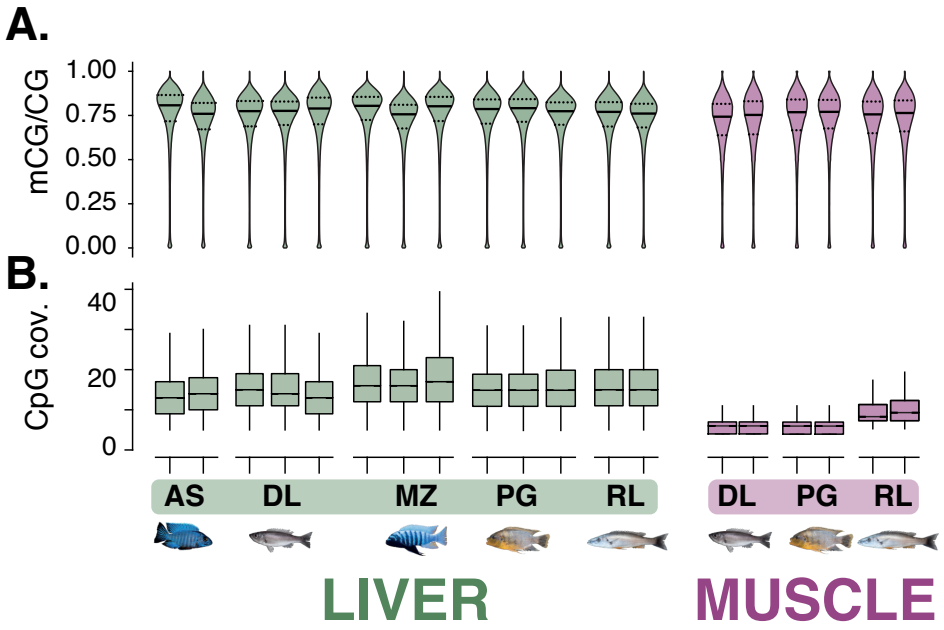


Fig. 2.9 Overall methylome levels in Lake Malawi cichlids. A. Violin plots representing genome-wide DNA methylation levels. Average of mCG/CG over non-sliding 1kbp-windows. Black line represents median values; dotted lines, 1st and 3rd quartiles. **B.** Boxplots representing the sequencing coverage at single CG dinucleotide (CpG) sites in all samples (CpG sites with coverage <5 or >100 non-clonal sequencing reads were filtered out).

Single CG dinucleotide (CpG) sites were then binned into 1kbp-long non-sliding windows - the average mCG/CG over such windows was then calculated and used for downstream analyses (except for identifying differentially methylated regions whereby a single CpG resolution approach was used). Muscle samples show methylation levels of 72.9%±1.4% (median, 76.0%±1.0%), and liver tissues, of 71.2%±1.1% (median, 78.2%±1.7%), all mean/median values ± st.deviation (Fig. 2.9).

Gene bodies and promoter regions

In cichlids of Lake Malawi, promoter regions consistently show lower methylation levels compared to genome-wide averages in all tissues of all species. There is a strong bimodal distribution of DNAme levels in regions around TSS (Fig. 2.10a.), with most of them being lowly methylated (<10%) and a small fraction of them showing high methylation levels, comparable to the genome-wide average (75-80%). Lower cytosine methylation at promoter seems to be a conserved feature of vertebrates and plants, but mostly lost in invertebrates (insects, nematodes, tunicates) and most fungi [162, 40, 163].

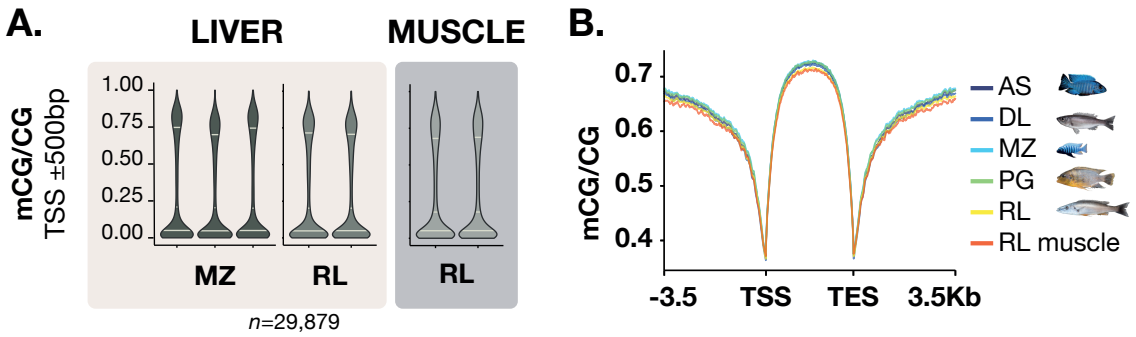


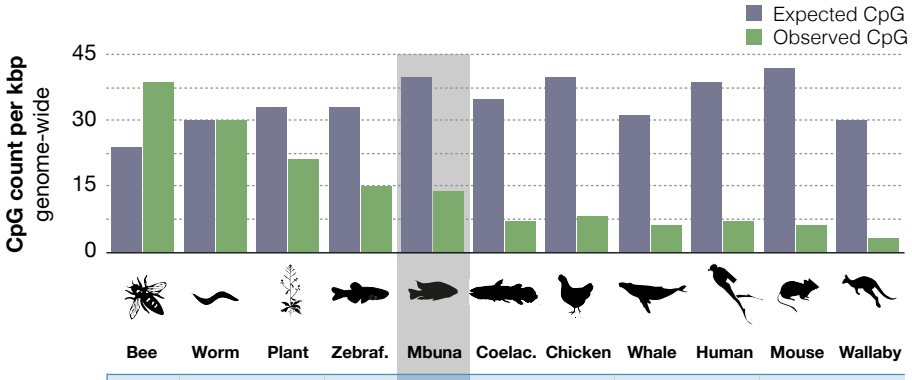
Fig. 2.10 DNA methylation levels at promoter regions. A. Violin plots representing DNA methylation levels at promoter regions (TSS±500bp) of all genes in liver tissues MZ (*n*=3) and RL (*n*=2) and muscle tissues of RL (*n*=2). Non-sliding 1kbp-bins, average of mCG/CG. White lines represents 1st, 2nd (median) and 3rd quantiles. **B.** DNA methylation levels at TSS, TES and gene body of all genes for all samples (*n*=3 except for RL liver and muscle tissues, *n*=2). Only CpG with coverage between 5-100 non-clonal sequencing reads analysed.

Interestingly, while promoters are largely hypomethylated in both tissues, gene bodies appear to show high levels of CG methylation, (Fig. 2.10b). Most plants, some invertebrates (many nematodes have lost gene body CG methylation) and most vertebrates show CG methylation in gene bodies [162, 40].

CGI

Mammalian genomes are highly CpG depleted, in that the frequency of CG dinucleotides genome-wide is extremely low compared to what would be expected by chance (low O/E ratio) [44, 164]. In contrast, fish genomes exhibit low CG depletion, with higher O/E ratios (ref. [71] and Fig. 2.11). It is thought to be mostly due to the increased mutability of methylated C, leading to global impoverishment for CG dinucleotides over extended evolutionary time (see Introduction, section 1.2.1). The roundworm *Caenorhabditis elegans*, lacking DNA methylation, shows the expected frequency of CpG for example. Many plants show cytosine methylation and consequently can be CpG-depleted (exemplified with *Arabidopsis thaliana*, whose genome is also composed CG-rich regions [165].

CpG islands (CGIs) are genomic regions showing significantly higher CpG frequency with a high C+G content. In mammalian genomes at least, CGIs tend to be lowly methylated regions overlapping promoters. In Lake Malawi cichlids, predicted CGIs exhibit a much higher CpG frequency than expected genome-wide (Fig. 2.12a). This ratio is also particularly high compared to mammals and plants (Fig. 2.12b). CGI in cichlids are 288bp-long on average for a total genomic coverage of 19.03Mb (Fig. 2.12c) – which is roughly the same



	Bee	Worm	Plant	Zebraf.	Mbuna	Coelac.	Chicken	Whale	Human	Mouse	Wallaby
E	24	30	33	33	40	35	40	31	39	42	30
O	39	30	21	15	14	7	8	6	7	6	3
O/E	1.63	1	0.64	0.45	0.35	0.2	0.2	0.19	0.18	0.14	0.1

Fig. 2.11 CG dinucleotide depletion in eukaryotic genomes. Most genomes exhibit a CG dinucleotide depletion, with lowest levels in mammals. Observed/Expected (O/E) ratio formula, see section 2.5.3. Teleost fishes exhibit a reduced CG depletion in comparison to amniotic species. See methods for detailed description of the organisms used (Coelac. and Zebraf. stand for Coelacanth and zebrafish, respectively). Silhouettes downloaded from http://www.phylopic.org/.

amount of CGIs as in humans (22.7Mb) for a genome three times smaller in cichlids; cichlids are therefore predicted to have many CGIs.

In stark contrast, only 31.1% of all annotated TSS regions in cichlids contain CGIs, which is half of what is observed in humans, and only 15% of all CGIs lie within promoter regions (Fig. 2.13a.). Hence, CGI regions are much less present in promoters in cichlids than in other organisms, in particular mammals, where CGIs compose more than two-thirds of all promoters (Fig. 2.13b. and Ref. [166, 44]). In contrast, promoter regions in cichlids are slightly more enriched in CGIs than other non-amniotic vertebrates, such as zebrafish or Xenopus (+10-15%) [71]. It therefore would be interesting to fully characterise CGIs that do not lie in TSS regions (coined 'orphan' CGIs) as they might exert unique biological regulatory functions, such as acting as ectopic promoters [44]. In mammals, 50% of all CGIs are predicted to be orphans (intergenic or intragenic regions); in cichlids, orphan CGIs are predicted to make up 77-85% of all CGIs (Fig. 2.13b). Moreover, methylation levels at CGI are strongly bimodal (either not methylated or highly methylated; Fig. 2.13c). Yet, CGIs located in promoter regions are practically not methylated at all (4.0%$\pm$0.3%, mean mC/C in % $\pm$ st.deviation; (Fig. 2.13d). Additionally, methylation at CGIs is not correlated with CpG density at CGI (Fig. 2.14a,b). Lack of cytosine methylation at CGI region of promoter regions is a conserved feature of vertebrate genomes, in contrast to CpG-poor promoters that are hypermethylated in general in humans [166, 167] and is often associated with a

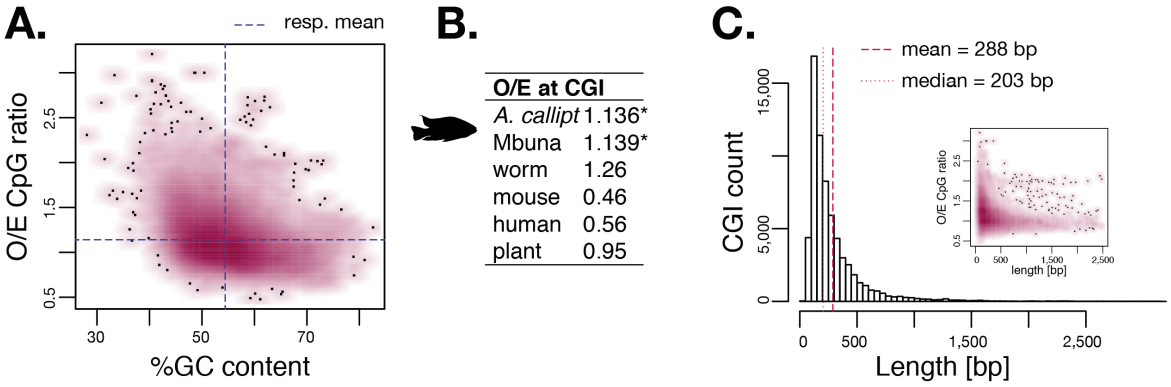


Fig. 2.12 CpG islands in Lake Malawi cichlids. CpG islands (CGI) in *M .zebra* genome were predicted using makeCGI (see methods). **A.** Plot of percentage CG content against O/E ratio at predicted CGIs. CGI, *n*=66,106. Dotted lines are respective mean values. **B.** Comparison of O/E ratios in CGIs of different organisms. Mbuna (*M. zebra*, UMD2a) and *A. calliptera* (fAstCal1.2) are Lake Malawi cichlid species. See methods for details on genomes used. **C.** Histogram of length distribution of all CGIs, given in bp. Mean and median values are shown with red dashed and red dotted lines, respect. Total genomic coverage of CGIs, 19.03Mb.

permissive chromatin state, promoting transcriptional activity [78, 44, 167]. Differential methylation state at CGIs has been reported to exert important regulatory functions in vertebrates [44, 167]. Also, many DNA-binding proteins possess either a CXXC or methyl-CpG-binding (MDB) domains, enabling them to specifically interact with unmethylated or methylated CG dinucleotides [78, 80].

The observed expected ratio is calculated as follows [164]:

$$\frac{O}{E} = \frac{\frac{CpG}{N}}{\frac{C}{N} \times \frac{G}{N}}$$

where N, CpG, C, G are the counts in bp in a given locus of total bases, CG dinucleotides, cytosines and guanines, respectively. CGIs were predicted using the unique genomic features pertaining to one species, as opposed to be predicted based on fixed, arbitrary threshold values, such as CG content or CpG frequency. Refer to section 2.10.7 for detailed methodology.

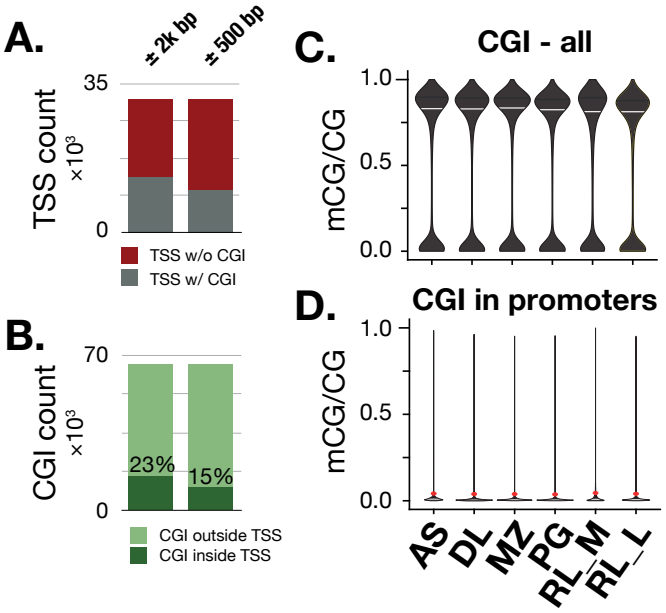


Fig. 2.13 CpG islands in promoter regions are associated with low DNA methylation levels. **A** Histograms representing the number of predicted CGIs overlapping promoter regions (either TSS±500bp or TSS±2kbp), and **B.** the number of promoters containing CGI in *M.zebra* genome. **C.** Violin plots summarising the distribution of DNAme at all CGIs (**E.**) and CGIs in promoters only (**F.**) in the five Malawi cichlids (liver methylome apart from RL muscle: RL_M). Average mCG/CG per species, $n \geq 2$. White lines represent median values. **D.** CGIs at promoter regions (n=8,356). White lines represent median values, red dots, mean values.

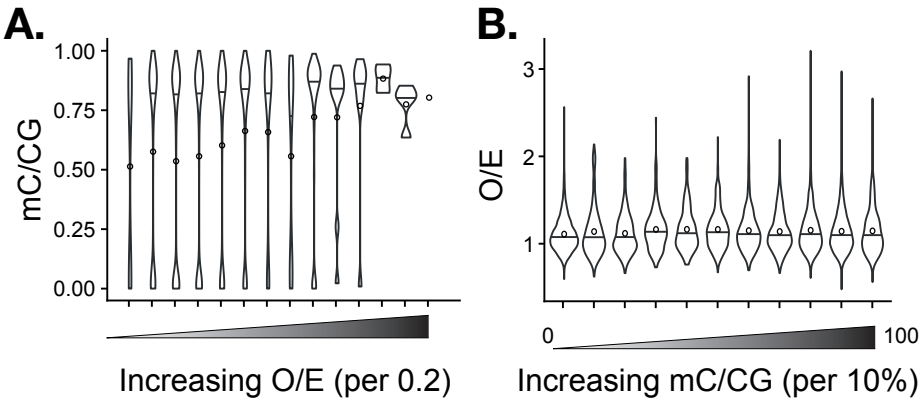


Fig. 2.14 DNA methylation at CGI is independent of CG dinucleotide density. **A.** Violin plots showing the DNAme levels at increasing O/E ratio at predicted CGIs (±0.2). **B.** Violin plots showing O/E ratios at CGI for increasing levels of DNAme (±20%). mCG/CG for *A. stuartgranti* liver (n=2). Black lines in violin plots are median values; round dots are mean values.

Transposable elements and repeats

Transposable and repetitive elements in the genome of *M. zebra* (UMD2a) were predicted and annotated using RepeatModeller and RepeatMasker (http://repeatmasker.org/). Overall, more than a third (35.4%) of the genome of *M. zebra* is predicted to be composed of transposable elements (TE) and repeats (Fig. 2.15a). This is >10% lower than in mouse and human (45% and 52.6%, respect.) and almost 20% less than in zebrafish genome [2]. The riverine cichlid lineage, Nile tilapia (*O. niloticus*), whose reference genome (high quality chromosomal assembly) has recently been published, shares a similar TE genomic load (37.6%, for a genome of similar size) [168]. Other African cichlids are expected to share similar TE genomic load however due to lower assembly quality, TE prediction might still be incomplete [53]. In East African cichlids, TE insertion have been reported to have mostly occurred near genes, which might impact transcriptional activity through *cis*-regulatory functions [110, 53, 92]. Genome assemblies of higher quality (use of PacBio long sequencing reads coupled with Illumina short and reliable reads, for example) will greatly improve TE prediction – this is the case for mouse, human, mbuna and tilapia genomes.

Table 2.1 Transposable element landscape in Lake Malawi cichlid *M. zebra*

	Count	Length [bp]	% total sequence	% TE seq
DNA transposons	414,229	125,219,920	13.10%	37.00%
Unclassified	409,831	98,453,305	10.30%	29.10%
LINEs[1]	180,029	73,523,563	7.70%	21.70%
LTR elements[2]	49,557	20,012,886	2.10%	5.90%
Simple repeats	299,920	13,388,563	1.40%	4.00%
SINEs[3]	26,063	3,461,624	0.40%	1.00%
Satellites	6,951	2,132,456	0.20%	0.60%
Low complexity	41,455	1,998,635	0.20%	0.60%
Small RNAs[4]	2,749	655,528	0.10%	0.20%
Total TE	**1,430,784**	**338,846,480**	**35.40%**	
Total nonTE		**618,574,682**	**64.60%**	

[1] Long interspersed nuclear element
[2] Long terminal repeats
[3] Short interspersed nuclear element
[4] ribosomal RNA, e.g.

Among all the TE and repeat families, DNA transposons, unknown repeats and simple repeats are the most frequent (Table 2.1) and together make up ∼70% of the repeat genome of MZ (Fig. 2.15b).

[2]Data retrieved for hg38, mm10 and danRer7 from RepeatMasker, http://repeatmasker.org/species/

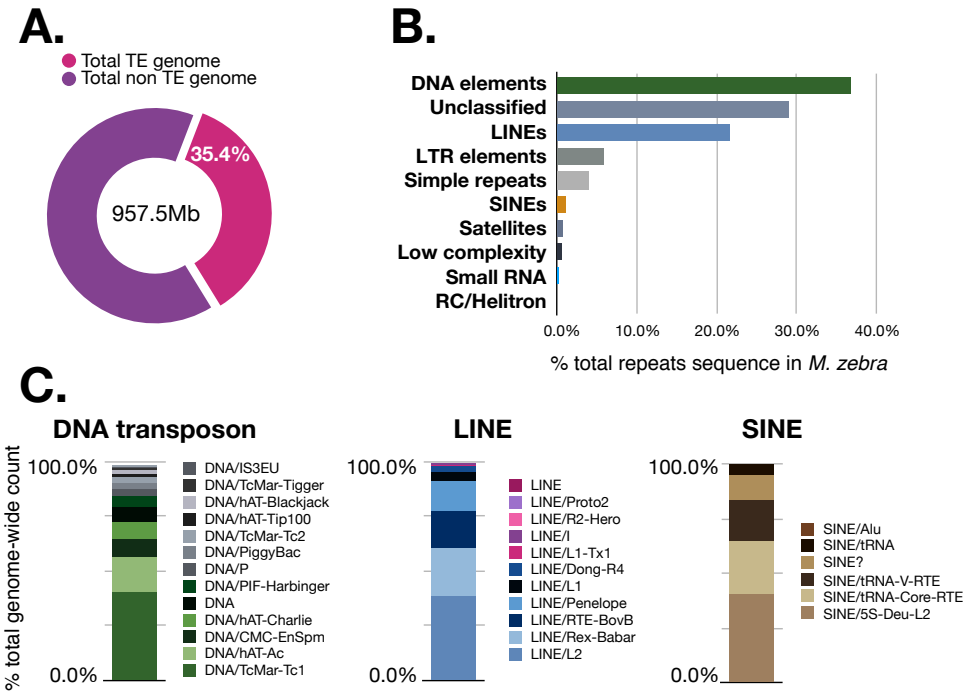


Fig. 2.15 Genomic landscape of transposable elements and repeats in Lake Malawi cichlid *M. zebra*. **A.** Proportion of TE (transposable elements and repeats) sequences predicted in *M. zebra* genome. **B.** Percentage of total TE sequences for each TE category. **C.** Genomic frequency of each TE family and subfamily (genome-wide count, in percentage).

In contrast to mammalian genomes, retrotransposons (LINE, SINE, LTR) are under-represented in cichlids and account for only 22%, 6% and 1% of the repeat genome, respectively. The predominance of DNA transposons has been observed in other non-amniotic vertebrates (e.g. zebrafish), while LTR and LINE/SINE are a genomic feature of mammalian repeat genomes. Interestingly, both LINE and DNA transposons appear to have expanded rapidly and recently (as shown by the increased genomic coverage of certain elements sharing similar sequence, Fig. 2.16). Some of these TE elements could still be active, with associated genomic transposition, explaining their recent expansion in the genome. In particular, a population of DNA transposons, Tc1/*mariner*, and of LINE L2 have recently expanded in copy number and now compose a larger portion of the genome. These might still be contain active catalytic domain enabling genomic transposition.

On the other hand, in human and mouse, the TE family LINE L1 is predicted to have recently expanded and now composes 17.5% and 19.9% (respectively) of their genomes. DNA transposons in human only make up less than 4% of the human genome, and do not show any recent activity in terms of transposition.

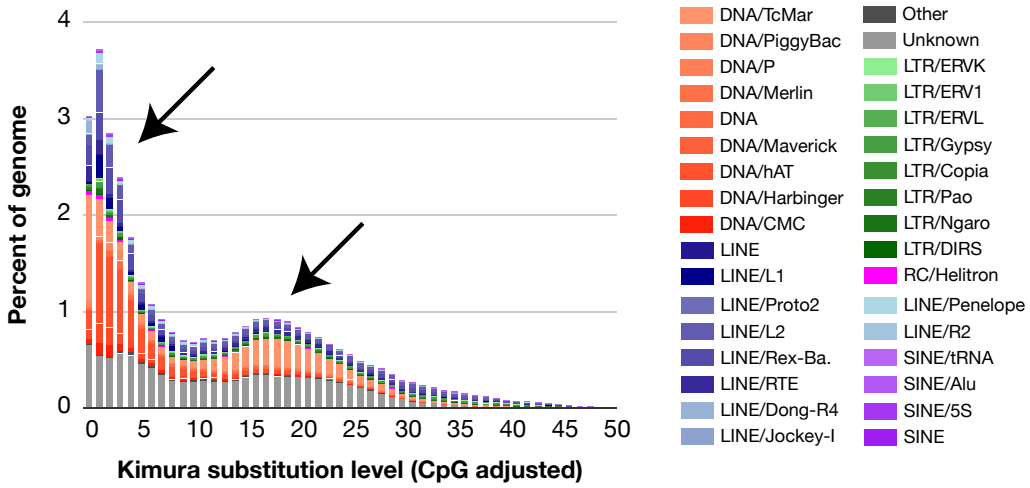


Fig. 2.16 Evolutionary recent expansion of certain families of DNA transposons and LINE elements in Lake Malawi cichlids. Landscape of TEs in *M. zebra* genome. TEs are classified into known TE categories and also according to sequence divergence (Kimura substitution levels, analogous to TE age). Two distinct waves of burst of TE activity are indicated by black arrows.

Interestingly, two waves of TE activity is predicted to have happened, and some populations of TEs, in particular DNA/TcMar and more importantly LINE/L2 and LINE/Rex-Barbar might still be particularly active in the genome of Lake Malawi cichlids (Fig. 2.16). This most recent wave of TE activity have been responsible for the expansion of LINE elements in cichlid genomes. A similar burst of activity could be traced back to older evolutionary times – DNA transposons were at that time the primary cause of repeat expansion. This recent and probably still active burst of activity, associated with specific expansion of some LINE families, seems to be unique to cichlids (to a lesser extent, to the riverine lineage tilapia as well). In zebrafish, there is no prediction of TE burst of activity. Moreover, LINE elements compose less than 2% of the genome, in stark contrast with cichlid genomes. LINE (L2, in particular) might turn out to be relevant in the recent adaptive speciation of cichlids.

Finally, a large fraction of repeats found in the genome could not be reliably classified into any known TE families (10.3% or 29.1% of total genome or total TE sequence, respectively). This is considerable and TE prediction in other genomes might shed light on new TE families.

Given their potential deleterious actions on the host genome due to transposition events, the activity of repetitive elements, i.e. transposition, is under tight control, suppressed by different cellular mechanisms. DNA methylation is thought to be at the front of defence in repressing their activity - often along with piRNA-mediated mechanisms, in particular

in germ cells. Young and active TEs are thought to be associated with high levels of DNA methylation in many organisms, resulting in their transcriptional silencing. More evolutionary ancient transposons might exhibit variable levels of DNAme, as they might be controlled by other defence mechanisms, such as KRAB zinc-finger proteins in mammals only (see Introduction), or might not be active due to mutational load acquired over evolutionary time.

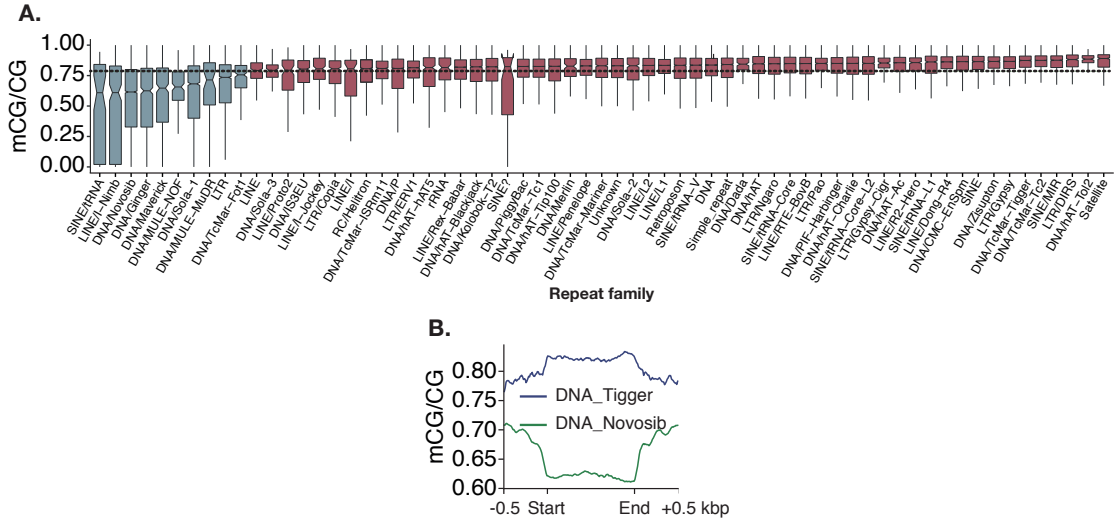


Fig. 2.17 Transposable and repetitive elements are differentially methylated in Lake Malawi cichlids. A. Boxplots representing the average DNA methylation levels (mCG/CG) for all the transposable elements and repeats families of Lake Malawi cichlids. TE families showing lower (blue) or higher (red) methylation levels (median) compared to genome-wide levels (dotted line). **B.** DNAme levels in two selected families of DNA transposon. Liver methylome for MZ shown only (average mCG/CG across given TE regions, biological triplicates).

In Lake Malawi cichlid species of this thesis, transposons are differentially methylated (Fig. 2.17a.). Most of TE families exhibit high levels of DNA methylation, some much higher than the average genome-wide level. This might reflect the high state of repression that is required in somatic cells. Some families, including DNA transposon families, such as Novosib (Fig. 2.17b.), Ginger, Maverick, and some LINE and LTR, show a much broader variation in DNA methylation, with some individual TE exhibiting very low levels of DNAme. This could be explained either by the fact that these lowly methylated TEs have accumulated deleterious mutations over evolutionary time, leading to their self-repression, explaining why DNA methylation at these TE might not be maintained. Alternatively, lowly methylated TE could have been co-opted for cellular functions [38], such as ectopic promoters or enhancers and are therefore lowly methylated to enable biological functions.

Overall, repeats and transposons are heavily targeted by CG methylation in mbuna. This could be one of the most conserved functions of DNA methylation, from plants, invertebrates, fungi to mammals: most organisms with an active DNA methylation machinery all show high levels of DNA methylation in repeats, in particular CG methylation, but also CHG methylation in plants [41, 40, 163]. This might reflect an important and ancestral mechanism of genome defence [34].

2.6 Methylome variation in cichlids of Lake Malawi

2.6.1 Overall

To investigate the variation in methylomes between cichlid species of Lake Malawi, I analysed only the variation in DNA methylation at conserved genomic CpG sites (i.e. conserved underlying DNA sequence between the five species).

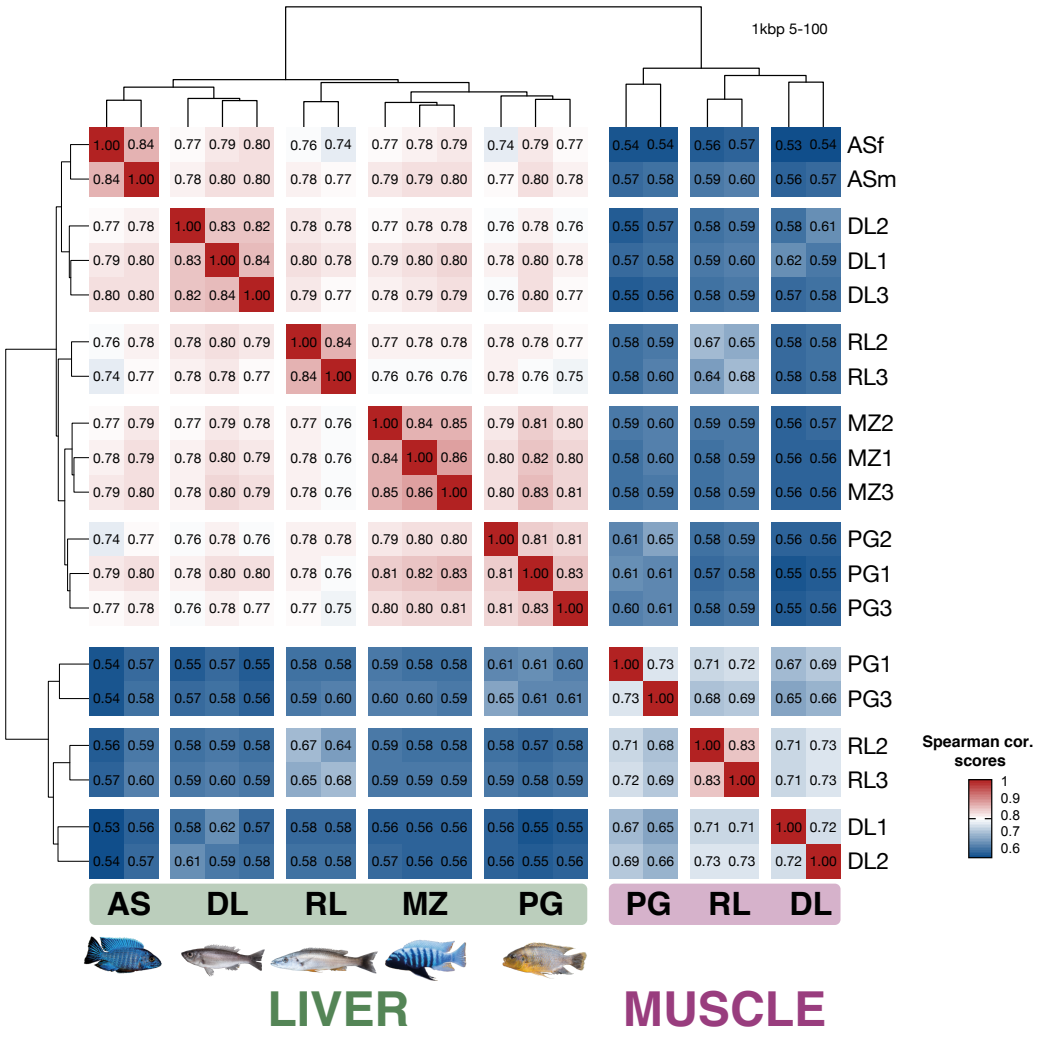


Fig. 2.18 Cluster analysis of methylome variation among tissues and species in Lake Malawi cichlids. A. Heatmap (with unsupervised clustering) representing Spearman correlation scores between genome-wide, single-CpG resolution methylomes of liver and muscle of Malawi cichlids. Average mCpG/CpG in non-sliding 1kbp-windows (coverage between 5-100 non-clonal seq reads; three biological replicates for all samples, except for RL, biological duplicates).

Overall, the variation in methylomes was first highly tissue-specific, with a lower inter-tissue correlation of 56-61%. Liver tissues, independently of the diet or habitat of the species, share more DNAme patterns with any other liver tissues than with any muscle (intra-liver methylome correlation, 74-84%). The same observation holds for muscle tissues, although to a lesser extent (intra-muscle correlation of methylome, 65-83%). Hence, inter-species liver methylomes are more similar than inter-species muscle methylomes. Remarkably, the variation in liver and muscle DNAme is also species-specific: individuals of the same species share the highest levels of DNAme variation, in particular for liver tissues (intra-species intra-tissues methylome correlation, 81-85%). In brief, DNAme variation in cichlids is tissue- and species-specific. Solely based on DNAme patterns in liver and muscle tissues at conserved underlying DNA sequences, phylogeny can be constructed accurately (Fig. 2.18).

In addition to the species and tissue specificities of DNAme patterns genome-wide, species sharing similar ecological habitats and some morphological features, such diet-related adaptation, tend to exhibit more similar patterns of DNAme genome-wide: the two rock-dwelling herbivores (MZ, PG) cluster together, while zoo-planktivores form another group, along with the piscivore RL. Principal component analysis of DNAme patterns at conserved CpG sites between the five species was able to decompose the species-specific variation seen in liver methylome, as well as the variation shared between species with similar diet (Fig. 2.19).

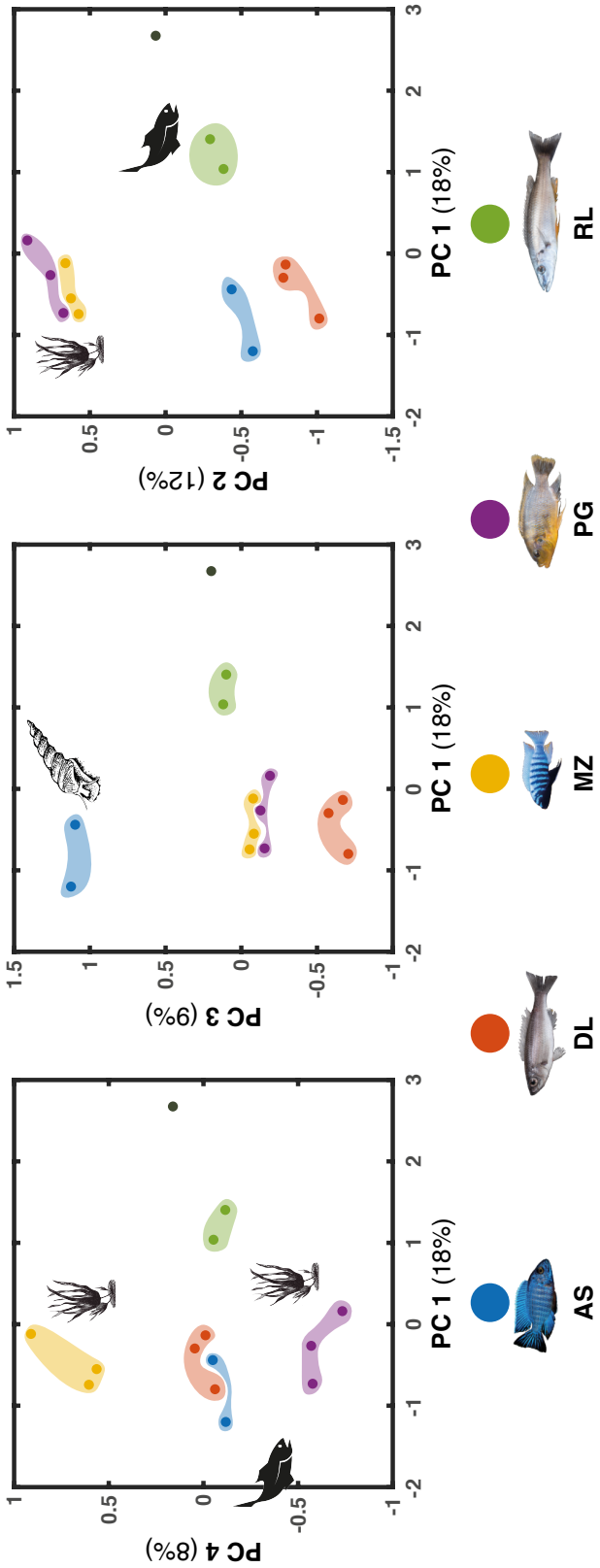


Fig. 2.19 Liver methylome variation in Lake Malawi cichlids. Principal component analysis of methylation variation at conserved CG dinucleotide sequences between all liver samples (coverage, 10-100 non-clonal reads). Intra-species methylome variation is in general smaller than inter-species variation. Liver DNAme variation could cluster individuals by species, and by habitat/diet for some PCs. PC2 could be explained by diet (carnivore vs herbivore), while PC1 would cluster the pelagic, piscivore RL away. Interestingly, distinct liver methylation patterns is observed for each species. One RL sample was dissected/mislabelled as liver, when it was probably spleen (grey dot).

To conclude, natural populations of cichlids dwelling in similar habitats and feeding on similar diet tend to show higher degrees of liver methylome similarity. Hence, liver methylome could reflect some level of adaptation, in particular with regards to diet adaptation. This warrants further characterisation of liver methylome with regards to trophic adaptation.

In this PhD work, the liver methylome will be characterised in greater details, identifying genomic localisation of DNAme variation. Such DNAme divergence will be compared to cichlids of other East African Lakes to investigate any epigenetic variability and convergence in genomic localisation (Chapter 3). Furthermore, upon environmental perturbation, such as alteration of the diet, dynamics in liver methylome will be assessed in order to quantify the plasticity in liver methylome variation in response to external stimuli in the context of adaptation and phenotypic diversity (Chapter 4).

2.6.2 Differentially methylated regions

Overview

Most of the methylation variation between two species is usually grouped in large genomic regions (from 50bp to several kbp; Ref. [43]). Such differentially methylated regions (DMR) are composed of a variable number of neighbouring CpG sites (≥5 sites), all exhibiting similar methylation levels (see Fig. 2.25 for an example of DMR).

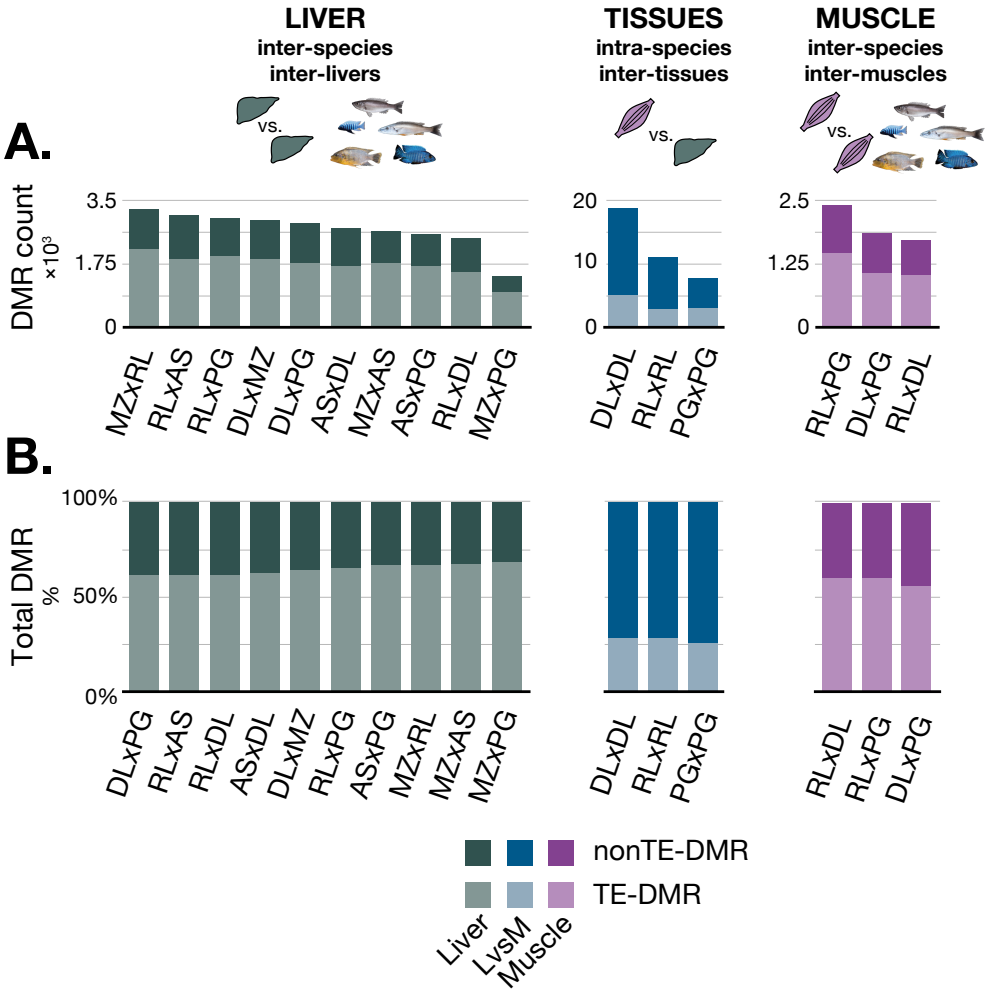


Fig. 2.20 Characterisation of DMRs in liver and muscle of Lake Malawi cichlids. A. Barplots representing the total number of DMRs found between each pairwise comparison of species. The proportion of DMRs for each comparison located in transposable elements or repetitive regions (TE-DMRs) is shown with a lighter colour for each barplot. Inter-species DMRs between liver (green, left), intra-species DMRs between liver and muscle tissues (blue,middle) and inter-species DMRs between muscle tissues (right, purple). **B.** Same as in **a.**, but values given as percentage over total DMR count, and bar plots sorted by increasing TE-DMR proportion.

To this aim, the software DSS [43] was used to predict DMRs among Lake Malawi cichlids, based on read coverage at each CpG sites, variability between biological replicates and methylation status of neighbouring CpG sites. DMRs showing a significant DNAme difference over genomic regions of at least 25% and with at least 5 CpG sites were analysed further.

I characterised and compared DMRs found between same tissues of different species (inter-species and intra-tissues, that is liver vs liver , muscle vs muscle between the five species) and between tissues within each species (intra-species, inter-tissues, that is liver vs muscle) (Fig. 2.20), in a pairwise fashion. Note that only three of the five species had both muscle and liver methylomes sequenced (Table 2.4). The highest number of DMRs between any liver methylomes is found between the algae-eating rock-dwelling *M. zebra* and the pelagic piscivore *R. longiceps* (*n*=3,264). These two species are also the most phenotypically divergent in terms of diet and ecological habitats. More closely related species based on diet have fewer DMRs: 2,480 and 1,435, between the liver methylomes of two pelagic and piscivore species (DL vs RL) and the two rock-dwelling algae-eaters (MZ vs PG), respectively. It would be important in the future to test this hypothesis further by sampling and investigating liver methylome difference in ecologically differentiated species within clades (e.g. herbivorous vs. carnivorous mbuna or herbivorous vs. carnivorous shallow benthic).

Likewise, DMRs between muscle tissues are also the most numerous when the two most phenotypically divergent species are compared, although overall there are fewer DMRs found between muscle tissues. Note, that this could also be due to an overall reduced sequencing coverage for muscle samples (Fig. 2.7).

Finally, DMRs between two tissue types (muscle vs liver) were far more numerous with a maximum number of close to 11,000 DMRs. This could reflect methylome patterns unique to cell type, essential to cellular identity (Fig. 2.20b).

In addition, close to two-thirds (64.4%$\pm$2.5%) of the inter-species methylome variation between liver tissues is located in transposable elements and repeat sequences (Fig. 2.20a,b, green histograms on the left). This is also true when comparing DMRs between muscles, with 59.0%$\pm$2.3% of them localising in transposon and repeat elements (Fig. 2.20a,b, purple histograms on the right). Interestingly, the proportion of DMRs overlapping TEs is not correlated with the ecological and phenotypic distance of the species compared.

In stark contrast, DMRs that are found between different tissues within the same species are far less associated with transposable elements and repetitive sequences, with only 26.9%$\pm$1.7% overlapping repeat elements.

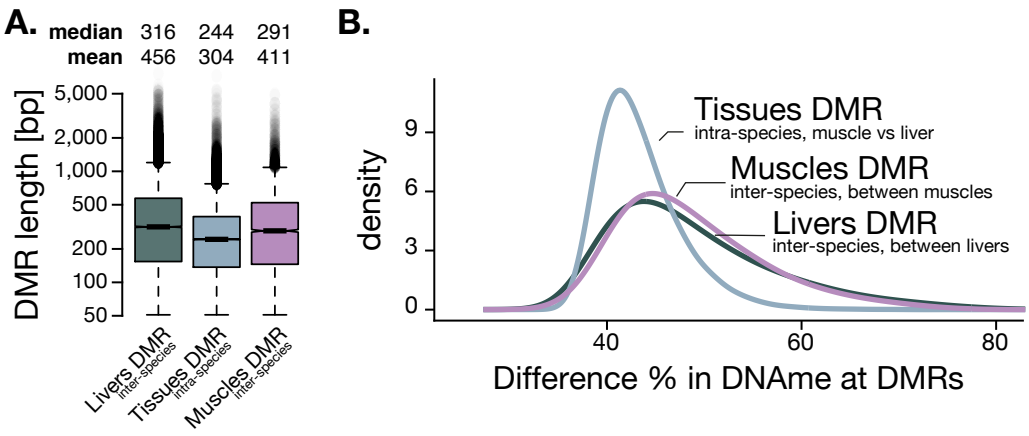


Fig. 2.21 Properties of DMRs in liver and muscle tissues of Lake Malawi cichlids. A. Boxplots of the length (in bp) of all DMR found between livers, between muscles and between tissues. Median and mean values of DMR length are given above boxplots. Logarithmic scale for y-axis. **B.** Distribution (kernel density estimation) of DNAme difference (converted to absolute % values) between DMR found in livers, muscles and tissues.

Overall, DMRs between tissues and between species are slightly different in their properties: inter-species intra-tissues DMRs (muscle vs muscles and liver vs liver) were slightly longer than the intra-species inter-tissue DMRs (411-456 bp long for inter-species DMR and 304 bp-long for intra-species tissues DMRs; Fig. 2.21a). In addition, although most DMRs in any comparison exhibit on average between 40-45% DNAme difference over DMRs, inter-species intra-tissue DMRs show a much broader DNAme difference with large differences (>50%), while most of the differences in methylation between tissues peaks at around 45% (Fig. 2.21b). Further work would include full characterisation of DMR found according to their types (TE vs nonTE, for example), which might reveal unique features.

In conclusion, inter-species variation in methylomes appears to be mostly associated with transposable elements, in particular for inter-species liver methylomes, meanwhile TEs are much less represented in the inter-tissue variance of methylomes. Nevertheless, intra-species DMRs between two tissues, although slightly shorter in size, are far more numerous than inter-species DMRs, which could arise from major epigenetic differences underlying two very distinct cellular types that are hepatocytes of the liver and myocytes of skeletal muscle.

Liver-specific DMRs

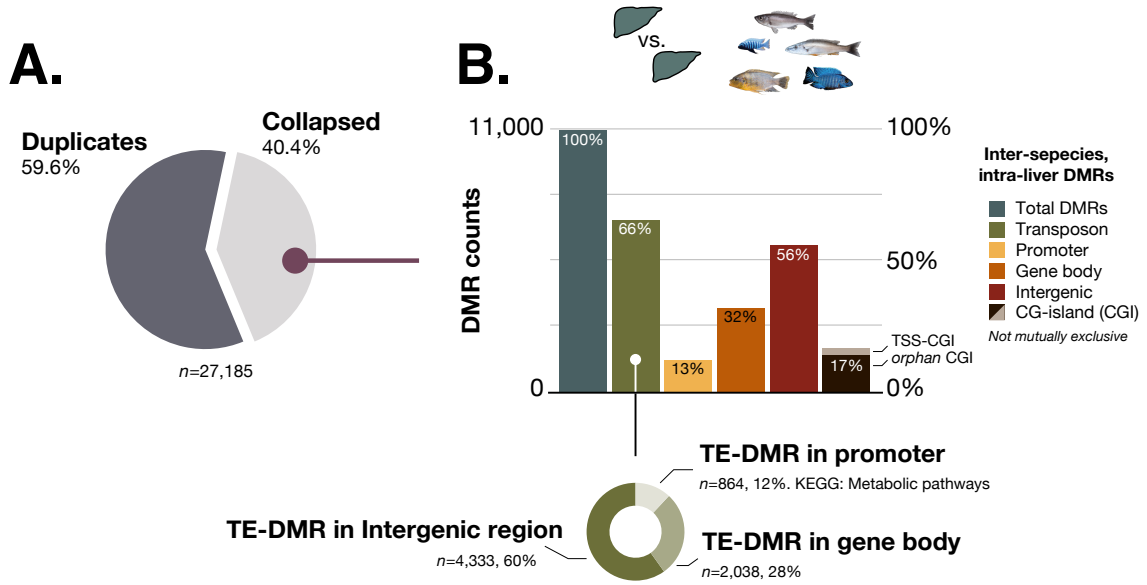


Fig. 2.22 DMRs in liver of Lake Malawi cichlids. A. Pie chart representing the number of DMRs found in more than 1 pairwise comparison ("duplicates"), and total and collapsed number of DMRs found among all species ("collapsed"). Total count of DMRs found between livers in all the 10 pairwise comparisons, *n*=27,185. **B.** Genomic localisation of all DMRs found in liver of the 5 Malawi cichlid species (collapsed). DMRs were found to overlap transposable and repeat elements (TE-DMR), promoter regions (TSS±500bp), gene bodies, intergenic regions (outside intron, exons and promoters) and CpG-rich regions (CpG-islands, CGI; overlapping with TSS or not [orphan CGIs]).

To further characterise liver methylome variation between all the five species at once (rather than pairwise), all liver DMR were merged/collapsed. Approximatively 60% of all DMRs are found in more than one pairwise comparison (Fig. 2.22a). This redundancy in the genomic localisation of DNAme variation is high between livers of the five species: some loci tend to show variation consistently in many species or to be unique to one species only (Fig. 2.22a).

Strikingly, of all the liver DMRs, two-thirds overlap transposable elements or repeat regions in the genome (TE-DMRs; Fig. 2.22b.), which represents a small enrichment over chance (1.13-fold, Fig. 2.23). Most of these TE-DMRs (60%) are located in intergenic regions, outside promoter regions and gene bodies. Intergenic DMRs, whether in TEs or not, make up 56% of DNAme variation in liver. Altogether, there is a considerable fraction of DNAme variation outside promoter regions and gene bodies, which may include important *cis*-regulatory elements, such as enhancers and ectopic promoters, with possible functions

Table 2.2 Enrichment analysis of liver DMRs genomic localisation.

	Expected[1]		Observed [2]	O/E
	mean	sd		
TSS	484.1	18.4	1,375	**2.84**
TE	6,389.2	34	7,235	**1.13**
GB	5,595.6	52.9	3,481	**0.62**
Intergenic	4,908.3	53.1	6,132	**1.25**
CGI	526.8	30.4	1,849	**3.51**

[1] Expected number, or expected by chance (random shuffling, genome-wide, 10 iterations, same DMR coordinates; see methods)
[2] See 2.22

in transcription regulation (Fig. 2.22b.). Overall, liver methylome variation is enriched for intergenic regions (1.25-fold enrichment over chance; Table 2.2).

Loci in the vicinity of TSS regions might exert important regulatory functions. Although they account for only 13% of the variation in liver methylome (Fig. 2.22), this represents 2.8-fold enrichment for liver methylome variation in promoter regions (Fig. 2.23 and Table 2.2). Promoter regions seem to be enriched for DNAme between species, surprisingly for the same tissues. It would be interesting to fully characterise DMRs associated with TE sequences and located in the vicinity of promoter regions overlapping TEs, as some TE insertions upstream of genes have been reported to bear regulatory functions, underlying novel expression patterns [20, 38].

Moreover, only 2.1% of all DMRs are located in TSS overlapping CpG-dense regions (CGI). CGIs show the highest enrichment for liver methylome variation, with a 3.5-fold enrichment. These regions are particularly targeted by DNAme variation in Lake Malawi cichlids, especially CGIs outside promoter regions (orphan CGIs), whose functions remain unclear [44].

Almost one-third of the variation is located in gene bodies, which comprises exons, introns and intronic enhancers (Fig. 2.22b.). This includes DMRs associated with TE sequences (28% total), but also many intragenic elements that are independent of repeats, and could include exons, introns and enhancers amongst others. Interestingly, gene bodies are significantly less targeted by liver methylome variation (almost half of what would be expected by chance, see Table 2.2). On the other hand, intergenic regions account for 56% of liver methylome variation (1.25-fold enrichment), which includes many TE-related loci, enhancers, suppressors and ectopic promoters.

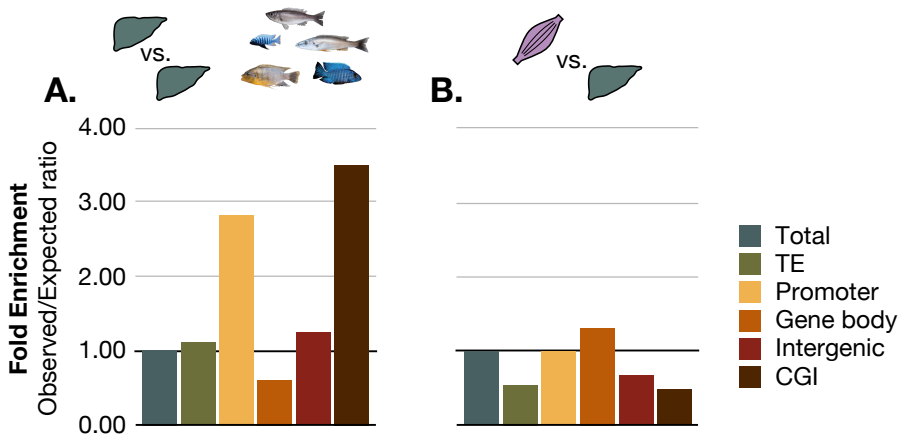


Fig. 2.23 DNAme variation is enriched for different genomic elements. Histograms showing the O/E ratios of the number of liver (**A.**) and tissues (**B.**) DMRs localised in transposable elements and repeats (TE), promoter regions (TSS), gene bodies (GB), intergenic regions or CpG islands (CGI) and the expected numbers by chance (random shuffling). Refer to Table 2.2 for detailed description of liver DMRs. Expected values are generated by randomly shuffling DMR coordinates genome-wide, 10 iterations. Total number of DMRs: between livers, *n*=10,988; between tissues, *n*=26,459.

In summary, most of the inter-species liver methylome variation is located in transposons and intergenic regions. There is a considerable enrichment for promoters and CGI regions, even though they account for less than 17% of the variation each, while gene bodies appear to be more preserved from this variation. In the genome of non-amniotic vertebrates, promoter are depleted in CpG-islands, in stark contrast with mammalian genomes. In cichlids, more than 70% of CpG islands are located outside TSS (namely, orphan CGIs). Interestingly, CGIs exhibit a high degree of DNAme variation between cichlid species, with a close to 4-fold enrichment. Such CGIs might exert relevant biological functions in DNAme variation and might be linked to specific cellular machinery.

Tissue-specific methylome

In contrast with inter-species liver methylome variation, there is a significant depletion of inter-tissues DMR in transposable elements, intergenic regions and CpG-islands (0.53x, 0.66x, 0.49x the expected values by chance; Fig. 2.23a.). Instead, gene bodies are slightly enriched for DMRs. Surprisingly, promoters are not the main source of DNAme variation. Rather, gene bodies are enriched for intra-species inter-tissue DNAme and might be important for cellular identity.

Species-specific/multi-tissue methylome

Intra-species methylome variation is more similar than inter-species variation (Fig. 2.18). I then investigated whether some variation could be unique to one species, regardless of the tissue analysed.

To this aim, intra-species, inter-tissues DMRs (i.e., DNAme variation between liver and muscle tissues within one species) and inter-species, intra-tissue DMRs were compared (Fig. 2.24a.). Interestingly, ca.30% of all DMR found in the three comparisons were common to the species, i.e. these DMRs were present in both tissues and are instead species-specific rather than tissue-specific (also referred to as species-specific, multi-tissue DMRs). The largest fraction of DMRs is specific to liver tissues (41-49%), while DMRs unique to muscle tissues compose 20-26% of all DMRs found (Fig. 2.24b). This means that liver tissues seems to show more DNAme differences between species than between muscle tissues: ca.60% of all DMRs found between muscle tissues of the three species were species-specific [found in liver also], while only 35-49% of all DMRs found between livers of different species were species-specific (Fig. 2.24c).

Approximately a third of the variation found in liver and muscle methylome between species may be tissue-independent. These DMRs are present in both tissues, and might also be in the whole organism. Sequencing other tissues could be done to confirm such species-specific, multi-tissue DMRs. This species-specific, tissue-independent variation may have been particularly relevant at earlier developmental stages and could have been passed on across cell divisions from one sister cell to the next during differentiation without erasure. This represent some sort of embryonic epigenetic memories [169].

Gene ontology enrichment analysis on genes associated with species-specific DMRs provides some evidence for these embryonic epigenetic patterns[3]: key developmental genes are particularly targeted by methylation variation (significant enrichment for genes related to hindbrain development).

Not all genes associated with multi-tissue DMRs are related to development: others appear to be relevant for different cellular functions. These could be unique to one species, and relevant not only to both liver and muscle tissues, but to other tissues and at other developmental times as well. For example, one species-specific multi-tissue DMR shows hypomethylation state of the gene predicted to be derived from a piggyBac transposable

[3]examples of candidates: beta-carotene oxygenase 1, *bco1* involved in retinal metabolism; BRF1 RNA polymerase III transcription initiation factor subunit b, *brf1b*; neuregulin 1, *nrg1*, involved in organogenesis; retinoic acid receptor alpha, *raraa* involved in the visual system; exosome component 8, *exosc8*; frizzled class receptor 3a, *fzd3a*, involved in brain development; *prickle1b*, mediating facial branchiomotor neuron migration

element (piggyBac transposable element-derived protein 4-like, *pgdb4* like, specifically in one species *R. longiceps*. piggyBac elements belong to an ancient family of DNA transposons [170], and are known to have been domesticated in many eukaryotic genomes, using, for example, the transposase activity [92] and might play a role in genome reorganisation (through transposase activity; see section 1.4.5). Additionally, it is used as a gene vector of choice in insects. This specific hypomethylated state in RL only is not fully understood, and could maybe be associated with transcriptional activity of this cryptic piggyBac-derived protein and nearby genes (Figs. 2.24d and 2.25).

In conclusion, a considerable portion of the variation in methylome is shared between different tissues in one species and might represent species-specific, multi-tissue patterns, contributing to distinct phenotypes. These species-specific DNAme patterns, representing ca.30% of the total DNAme variation, might also reflect embryonic epigenetic memories (possibly vestigial DMRs). These marks may be epigenetic relics following embryogenesis when they may have had a functional roles. Many essential functions associated with DNAme take place early on during development. Differential methylation at key developmental *cis* regulatory loci may underlie important phenotypic differences, in particular in the context of species radiation. This motivates further characterisation, in particular of the methylomes and transcriptomes of embryos of different species and possibly inter-species cross at different developmental times. Such experiments might highlight the impact of these embryonic DMRs on the expression of key developmental genes, possibly underlying phenotypic change.

Current work focuses on characterising any functional link between liver-specific DMRs unique to one species and gene expression in liver, with a particular emphasis on TE-DMRs localised in promoters. Thus far, overall correlation has been drawn (see below).

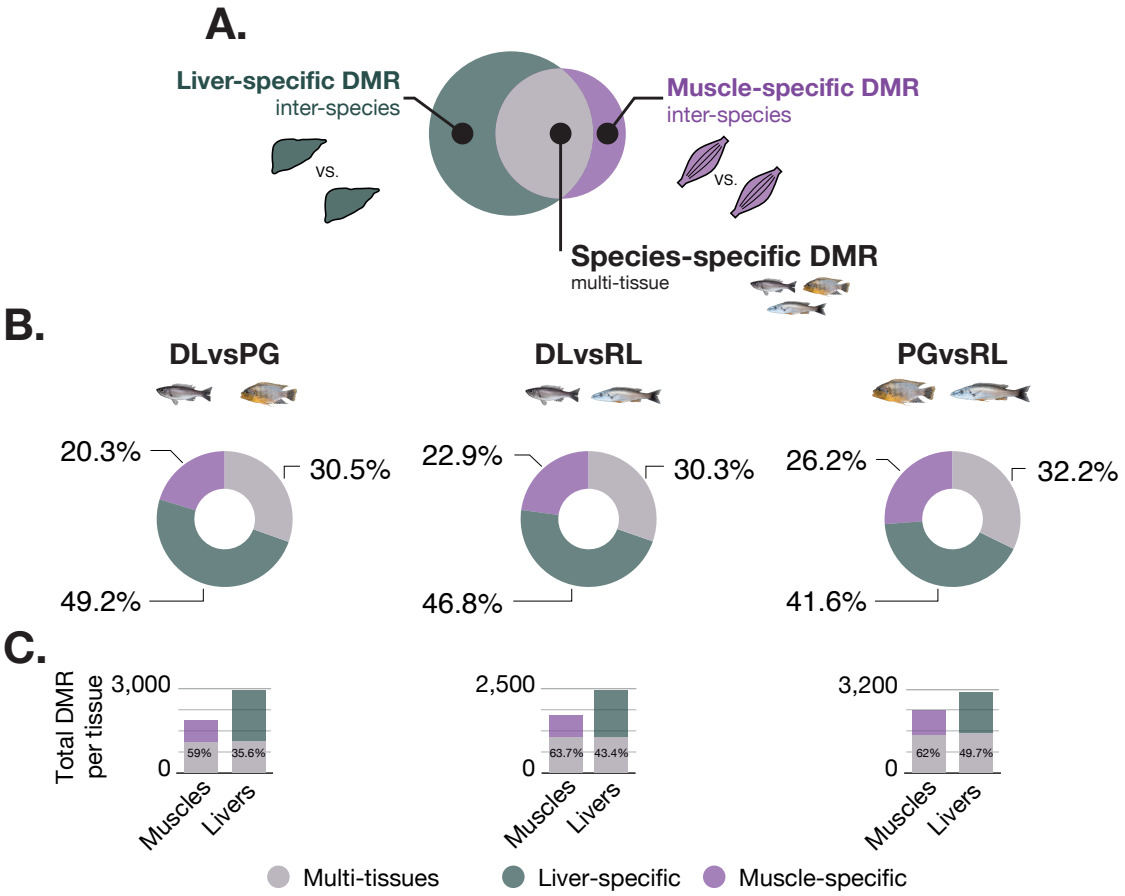


Fig. 2.24 Tissue- and species-specific DNAme variation. Methylome variability is tissue-specific and also species-specific (multi-tissue at least). **A.** Pairwise comparison of DMRs between livers and muscles in three species. All the DMRs found between livers (left, green) and muscle (right, purple) of two species (i.e., intra-tissue, inter-species DNAme variation). Some variation is shared between the two tissues (multi-tissue) and might be species-specific rather than tissue-unspecific (overlap in grey). **B.** Detailed description using pie charts of pairwise comparison of DMRs found between liver and muscle tissues of three different species. DMRs can be species-specific (green), muscle-specific (purple) or common to both tissues (species-specific/multi-tissue, grey). **C.** Histograms summarising the number of DMRs in muscle (left) and liver (right) that is common in both tissues (grey bar) for each comparison (inter-species within one tissue).

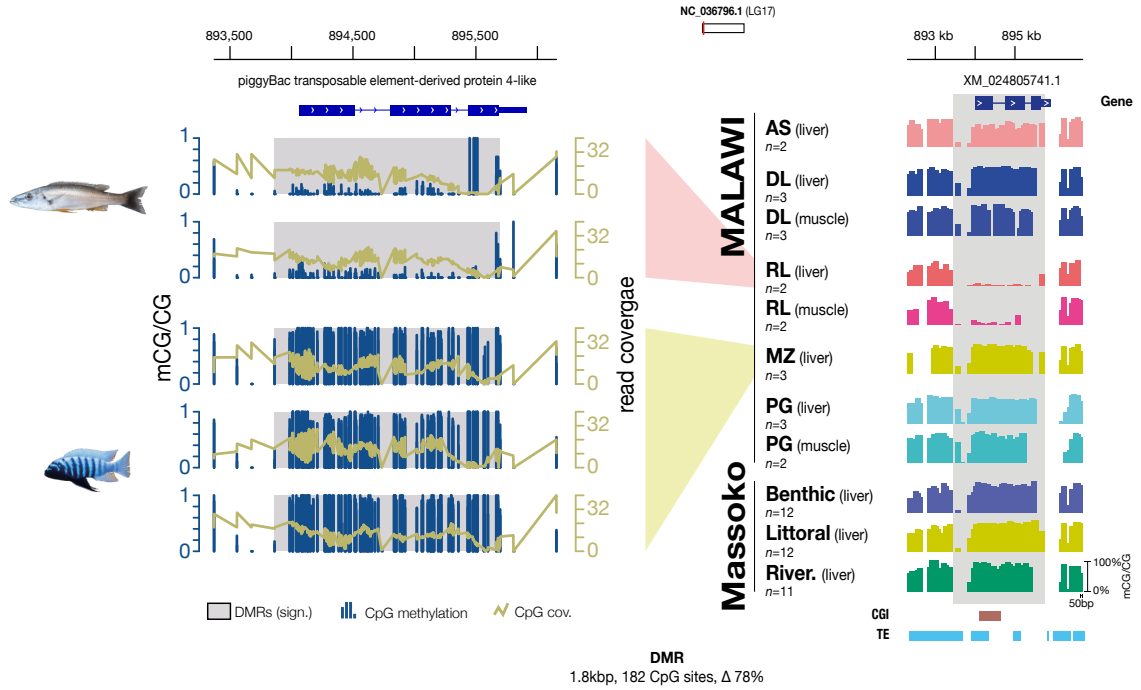


Fig. 2.25 Detailed genome-browser view of one species-specific multi-tissue DMR. Here, an example of the 1.8kbp-long tissue-independent (or multi-tissue) DMR specific to *R. longiceps* (RL) only (from Fig. 2.24d). Only *M. zebra* and RL (n=3 and n=2, respectively), both liver tissues, are shown (left). Individual CpGs comprising the DMR (in total 182 CpG sites) are shown in blue, non-clonal read coverage at each CpG site is shown in yellow. The DMR is delimited in grey. Hypomethylated (78% of mC difference on average over this region between MZ and RL livers) in RL liver and muscle (not shown) compared to all other fish sequenced, independently of tissues. This DMR spans the promoter region and gene body of the DNA transposon, piggyBac transposable element-derived protein 4-like. (Right) The methylation levels at this particular DMR is shown for all the five Malawi fish are shown. Figure created with DSS [43].

2.7 Transcriptome of Lake Malawi cichlids

The role of DNAme in regulating gene expression, in particular at gene promoters, has been extensively studied. Change of DNAme levels has been reported to be associated with differential transcriptional activity of key developmental genes, enabling cell-type differentiation and normal development [18, 17]. Early life establishment of stable and heritable DNAme patterns may thus be correlated with transcriptional changes in somatic cells as well, thus underlying phenotypic variation. Many DNAme readers, such as methyl-sensitive TFs, would be greatly affected by change in DNAme. Furthermore, it is important to note that change in gene expression might be mediated by other mechanisms, in parallel with DNAme, pertaining in particular to TF machinery present in one specific cell type.

To test this hypothesis in cichlids, I then sought to investigate the correlation between liver methylome and gene expression patterns. I generated and analysed the muscle and liver transcriptomes (RNAseq) of three of the five cichlid species, in order to understand whether any DMRs would be correlated with change in expression of genes, thus reflecting diet adaptation. This next section will first characterise gene expression landscapes in each species, to then investigate the correlation between gene expression and DNAme variation. Refer to Table 2.4 for a detailed description of sampling design and sample size.

2.7.1 Species-specific transcriptome variability

Overall, patterns of gene expression are first tissue-specific, and then species-specific (Fig. 2.26). Correlations of gene expression between any liver samples or any muscle tissues are higher (76-83%) than between two different tissues, even if of the same species (69-75%). Lastly, liver or muscle tissues of species thriving in similar diets/habitats tend to be less divergent transcriptionally (Fig. 2.26).

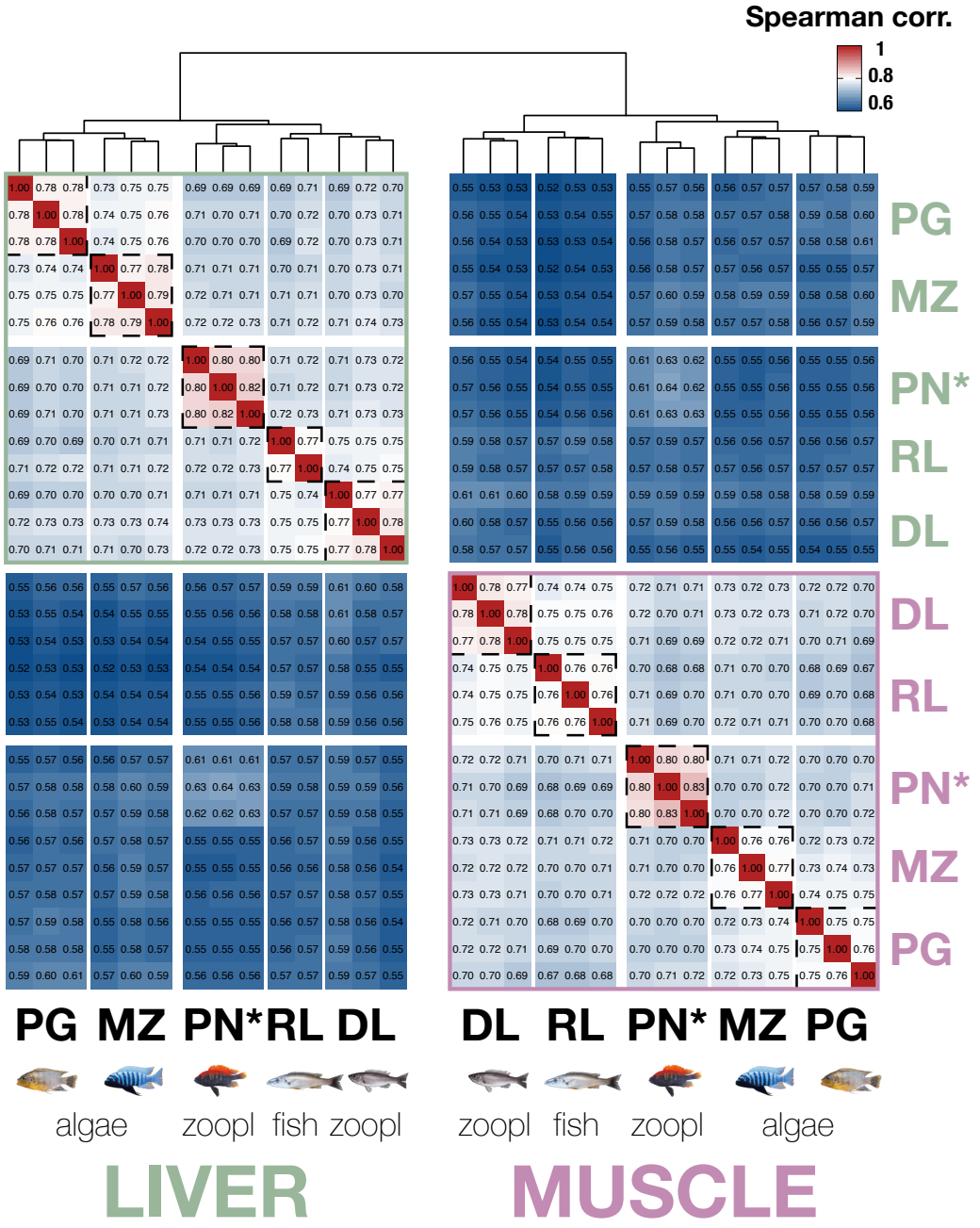


Fig. 2.26 Transcriptome landscape in Lake Malawi cichlids. Unsupervised hierarchical clustering based on transcriptome divergence in both liver and muscle tissues of Lake Malawi cichlids. One cichlid species of Lake Victoria, *P. nyererei* PN, is used for comparison. Spearman correlation matrix was generated using expression profile (ln(tpm+1)) for all transcripts of the reference genome for Lake Malawi UMD2a.

2.7.2 Differentially expressed genes

In total, there are 3,447 significantly differentially expressed genes (DEGs) between liver tissues of four Lake Malawi cichlids (many of which are differentially expressed in more than 1 comparison, referred to as "shared", as opposed to unique DEGs that are found only in one specific comparison). As expected, there are more DEGs between phenotypically distant species, with the highest number of DEGs (n=2,176) between the rock-dwelling, algae-eater *P. genalutea* (PG) and the pelagic *D. limnothrissa* (DL). Unsupervised clustering of DEG expression cluster individuals by species followed diet/habitat/morphology (Fig. 2.28a). More closely related species, sharing more similar diet or habitat, have much fewer DEG, with the lowest number of DEG found between the livers of the two carnivorous, pelagic species (DL vs *R. longiceps* RL, n=195). The two herbivore, rock-dwellers (*M. zebra* [MZ] vs PG) still exhibit a considerable amount of DEGs (n=933). Interestingly, few DEGs are unique to one comparison: the same genes are differentially transcribed in livers of many species (grey part of the histograms; Fig. 2.27a).

As expected, most of the differences in liver transcriptomes in Lake Malawi fish are related to metabolic pathways overall: many cytochrome P450 genes are differentially expressed, as well as genes involved in xenobiotic, fatty acid and steroid metabolic processes (Fig. 2.27b). This could reflect adaptation to different diets and to distinct sources of fatty acids, such as the ones possibly explained by a animal- vs plant-based diets (Fig. 2.27b,d). Genes that are upregulated in herbivore species pertain to processes related to fatty acid metabolism (especially cytochrome P450; Fig. 2.28b). In carnivorous species, common upregulated genes are also related to metabolic processes, but more precisely to amino acid metabolism (Fig. 2.28d).

Furthermore, PG exhibits a somewhat unique liver transcriptome, with many genes only differentially expressed consistently in the three biological replicates of this species (28.2% of all DEG in liver). Interestingly, DEGs in PG are particularly enriched for processes related to ribosome biogenesis, rRNA metabolic pathways and necroptosis (Fig.2.28a,c). This unique pattern of transcription in liver PG is difficult to interpret: it could reflect cellular response to some external and local environmental stressors or toxins, unique to these three individuals (all caught in the same lake area), at the time of collection. It could also be actual biological differences in hepatic transcriptome. Different epigenetic factors, such as *tet3*, *nsun2* and *dnmt3bb.3* show higher transcriptional activity – similar dynamics of key epigenetic genes (in particular *tet* enzymes) have be identified as the hallmark of cellular regeneration in damaged liver in mice [70].

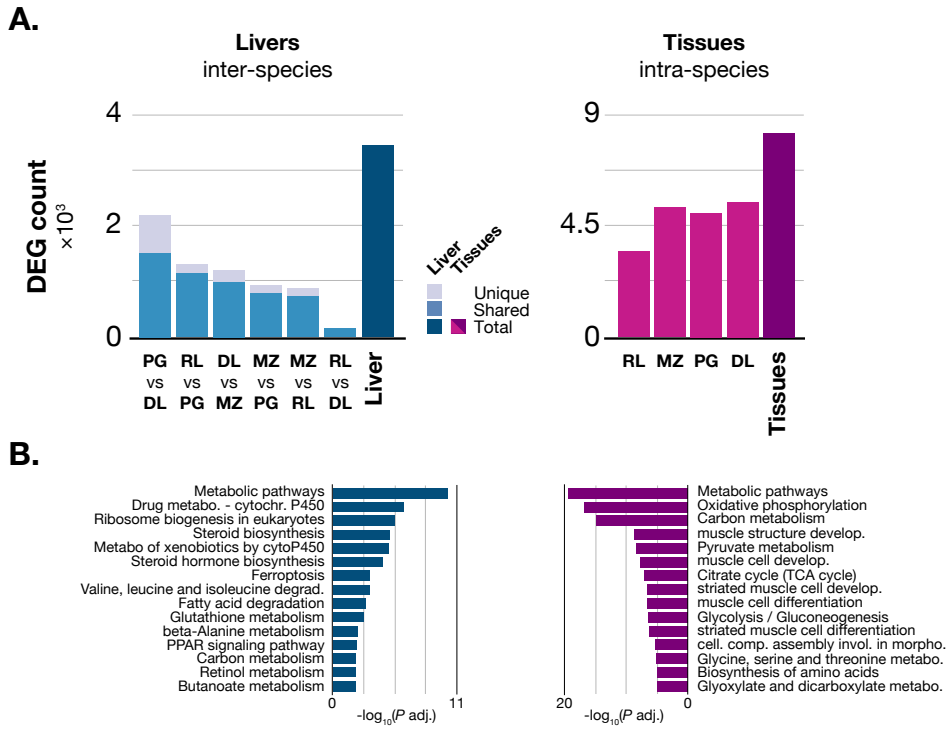


Fig. 2.27 Differentially expressed genes in Lake Malawi cichlids. A. Number of differentially expressed genes (DEG) in four Malawi cichlid species. **Left**, DEsG between livers, inter-species pairwise comparison (DEGs unique to one comparison in light blue; DEGs shared with at least one other comparison in darker blue; total number of DEGs found in at least one comparison, *n*=3,437). **Right**, DEGs between tissues (liver and muscle) in each species, intra-species pairwise comparison (total number of DEGs in at least one comparison, *n*=8,258). All DEGs, FDR<0.01. DL, *D. limnothrissa*(*n*=3); RL, *R. longiceps* (*n*=2); MZ, *M.zebra* (*n*=3); PG, *P. genalutea* (*n*=3). **B.** GO enrichment analysis for DEGs (top 15). Left, intra-species DEGs in liver; right, intra-species between tissues (muscle vs liver).

On the other end, differences in transcriptomes between tissues (liver and muscle) of any of the four species studied are approximatively twice as numerous: 4,811±886 genes are differentially expressed between liver and muscle tissues, with the lowest number of inter-tissue difference found between liver and muscle transcriptome of RL (*n*=316). In total, there are 8,258 DEGs, which means that a large fraction of the same genes are differentially expressed consistently in many species, as expected. Genes pertaining to one specific cellular type are differentially expressed, such as the ones involved in muscle structure development, or in the citric acid (TCA) cycle and pyruvate metabolism.

Furthermore, it would be interesting to draw a comparison with genes that are differentially expressed in the muscle tissues between two species, in order to understand and put in context the difference in transcriptomes between livers of two species. I hypothesise

that inter-species variation in gene expression in muscles may be more homogeneous and therefore smaller than in livers, as liver functions might be reflect adaptation to the different sources of food present in Lake Malawi, and therefore under selective pressure.

In conclusion, in terms of number of genes differentially expressed, livers of two different species (in particular ecologically-related species) have five times fewer DEG than between tissues of one species. As expected, muscle tissues exhibit patterns of expression very distinct to liver tissues (54-58% overall correlation between liver and muscle transcriptomes). However inter-species variation in liver transcription exhibit considerable change, with many genes unique to some species, and these may be important to ensure fast and plastic liver adaptation to different diet. *Cis*-regulatory elements and differential methylation levels in the genome could participate in the fine-tuning and alteration of transcriptional network in liver.

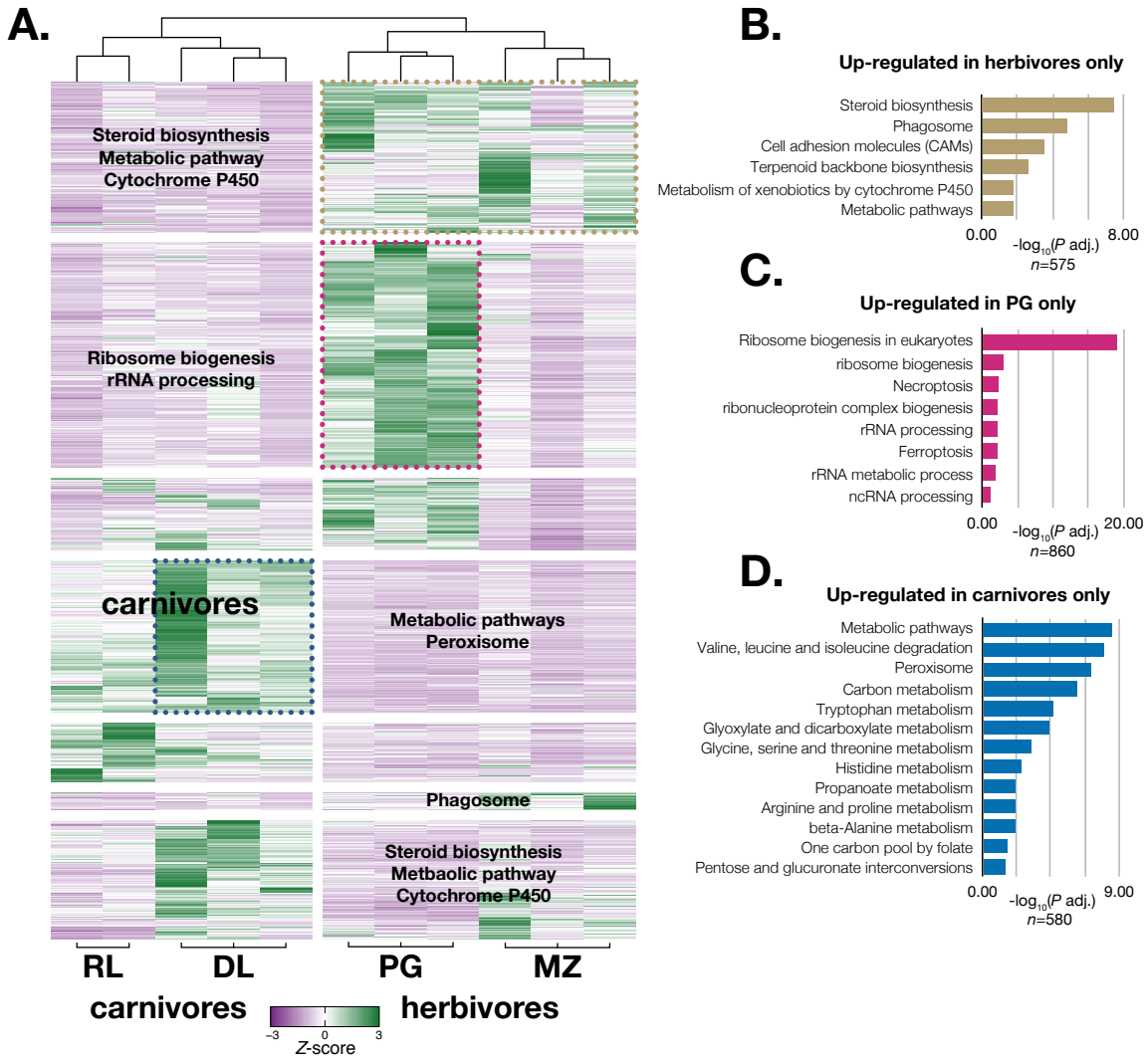


Fig. 2.28 Gene expression profiles of the livers of Lake Malawi cichlids. A. Heatmap and unsupervised clustering of the expression pattern for all differentially expressed genes (DEGs) in liver of Lake Malawi cichlids (averaged TPM per gene per individual, *z*-score). Only genes with TPM $\geq$5 in any individual are shown. *n*=3,044. FDR<0.01. *n*=3 for all (male specimens). DL, *D. limnothrissa*; RL, *R. longiceps*; MZ, *M. zebra*; PG, *P. genalutea*. GO enrichment analysis for upregulated genes in the herbivores MZ and PG (**B.**), in PG only (**C.**) and carnivores DL and RL only (**D.**). Only significant terms (adj.pval<0.05) are shown.

2.8 Transcriptome and Methylome interplay

I then sought to understand the correlation between transcriptional activity and DNAme variability at different genomic loci. In particular, whether differential levels of DNAme at *cis*-regulatory regions, such as promoter regions, could be correlated with variable levels of gene expression. It has been shown that lowly methylated promoter are more permissive for the transcriptional machinery to anneal to TSS regions, sometimes via a crosstalk with histone marks [78, 44, 25]. DNAme at gene bodies has also been shown to bear some biological functions, such as regulation of splicing events, and is usually correlated with higher transcriptional activity [78, 86].

To address this aim, I correlated all the DMRs that could be linked to genes (due to their location in gene promoters, gene bodies and if linked to TEs in the direct vicinity [1kbp] of DMRs) with differentially expressed genes in liver tissues. Of note, all sequencing data (WGBS and RNAseq) were generated using the exact same homogenised tissues for all the species (see method 2.33).

Table 2.3 Correlations between gene expression and DNAme at promoters and gene bodies.

		M. zebra	*R. longiceps*
TSS	**liver**	-0.36	-0.38
	muscle	N/A	-0.42
Gene body	**liver**	0.02	0.12
	muscle	N/A	0.2

[1] Values represent Spearman's correlation scores, p values <0.002 for all pairwise correlations
Refer to Fig. 2.29 for graphs.

In fully differentiated tissues of Lake Malawi cichlids (liver and muscle at least), the correlation between methylation at promoter regions and gene expression is negative, although weak (Spearman correlation in liver and muscle, -36% to -38% for liver tissues of MZ and RL, respectively, and -42% for RL muscle, p values <0.001; see Figs. 2.29 and 2.30, as well as Table 2.3). In other words, higher levels of DNAme at promoter regions are correlated with lower gene expression on average, and vice versa. Interestingly, in muscle tissue, this negative correlation is slightly stronger than in livers (Fig. 2.29). Both in muscle and liver tissues, promoter regions tend to be highly methylated (median around 72% methylation), while showing drastically lower methylation levels for even lowly expressed genes (ca.15% methylation at TSS), with the lowest levels for the second highest category of gene expression

(5% on average ; Fig. 2.29a and Fig2.30). The distribution of DNA methylation levels at TSS is bimodal, with the majority of TSS exhibiting 0-5% methylation levels (and more permissive transcription), and a smaller fraction of promoter regions exhibiting 70-85% DNAme levels and low transcriptional activity (Fig. 2.29c). Even though the negative correlation is highly significant, there is a considerable variation in DNA methylation levels in each category, especially for genes showing low and intermediate levels of expression.

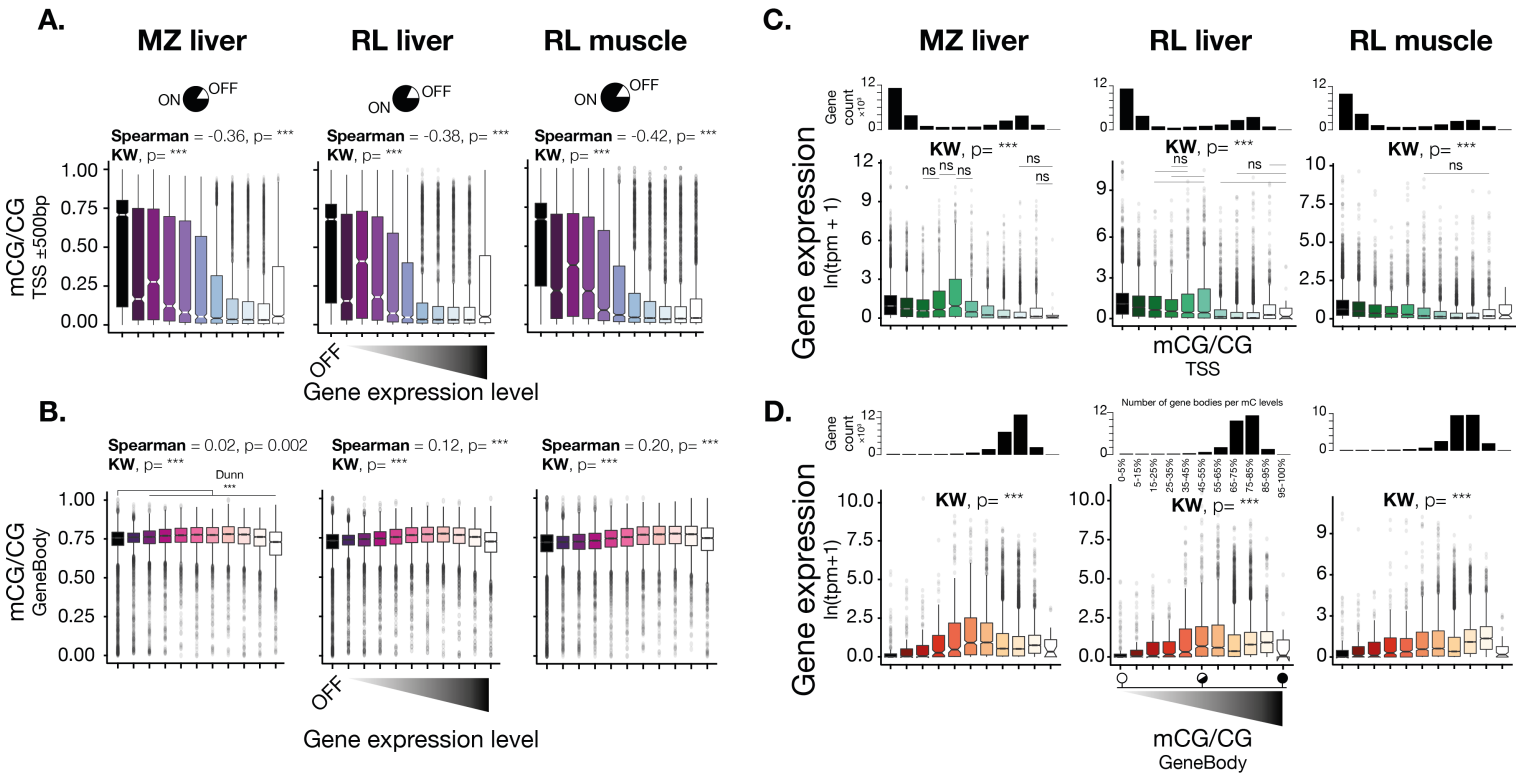


Fig. 2.29 Functional correlation between methylomes and transcriptomes. A,B Boxplots representing methylation levels at TSS and gene body according to gene expression levels in liver and muscle tissues. y-axes represent methylation levels at TSS (500±TSS, **a.**) and gene body (**b.**) against gene expression levels. Genes were grouped in categories (with same number of genes, *n*=2,103), including one category for unexpressed/silent genes ("OFF") and 10 categories of increasing transcriptome values. Pie chart represent the number of silent/unexpressed ("OFF") and of expressed genes ("ON"). Gene expression levels are averaged TPM values per gene for each species (*Maylandia zebra* [MZ] liver and Rhamphochromis longiceps [RL] liver, *n*=3 each; RL muscle, *n*=2). Methylation values are averaged mCG/CG ratios per loci per species (MZ liver and MZ muscle, *n*=3 each; RL liver, *n*=2). Scores for Spearman correlation tests, as well as scores for Kruskal-Wallis and Dunn's tests are indicated above boxplots. **C, D.** Boxplots representing gene expression levels at different methylation level categories at TSS (**c.**) and gene body (**d.**). TSS (**b.**) and gene body (**d.**) are grouped into categories based on their methylation levels (10 categories, from 0-100%). x-axes represent averaged TPM values per species for each category. Histograms showing the number of genes per methylation category are given above boxplots.

In the case of gene bodies, the variation in methylation levels is smaller, with most gene bodies consistently exhibiting very high levels of methylation (75-80%; Fig. 2.29b.). Very few gene bodies show methylation levels lower than 55% on average (Fig. 2.29d.). Of note, in addition to exons and introns, gene bodies can include some intronic enhancers and alternative splicing sites. Higher gene expression levels are correlated with higher methylation levels at gene bodies. However, the most highly expressed genes show similar DNAme levels to the ones observed for lowly expressed genes (Fig. 2.30). The correlation between gene expression and DNAme at gene bodies is positive, yet very weak: almost null for liver tissues of the rock-dwelling algae-eater *M. zebra* (Spearman, rho=0.02, $p<0.002$), but slightly higher for liver tissues of *R. longiceps* (Spearman, rho=0.2, p<0.001; see Table 2.3). Like for promoter regions, this positive correlation is slightly higher in muscle tissues (20%, p<0.001). Higher CG methylation levels at gene bodies of genes showing intermediate levels of transcription activity are a well conserved feature, from some plants, to invertebrates and vertebrates [35]. Interestingly, the correlation between gene body methylation and transcriptional activity does not extent to the most highly expressed genes in cichlid, which are not the ones presenting the highest levels of gene body methylation. Paradoxically, gene body methylation levels of the most transcriptionally active genes resemble these of silent genes. Similar observations have been made in many invertebrates (insects, tunicates, cnidarians), vertebrates and plants (angiosperms and tracheophytes) [35].

In conclusion, actively transcribed genes in fully differentiated tissues (muscle and liver, at least) is correlated with increased methylation levels in gene bodies, and with very low methylation levels in promoter regions. This contradictory link between DNAme and transcription reflects the complexity and the wealth of cellular functions and the many diverse interactions that DNAme can exert. A realm of different readers can interact with DNAme, and their activity might vary depending on the spatial distribution of DNAme as well as on the levels of modification (e.g. transcription factors that are CpG methylation sensitive) [78, 80]. In mammals, similar correlations have been observed, and this could reflect evolutionary conservation of the DNAme-related mechanisms of transcriptional regulation in many vertebrates [17]. It is important to note that the activity of some TFs might change DNAme levels at promoter regions for example, or that SNPs in one particular TF-binding genomic locus might also influence local DNAme levels [78, 80] – all the aforementioned scenarios could be identified in our datasets and are of particular interest, as they may underlie important adaptive changes in phenotypes.

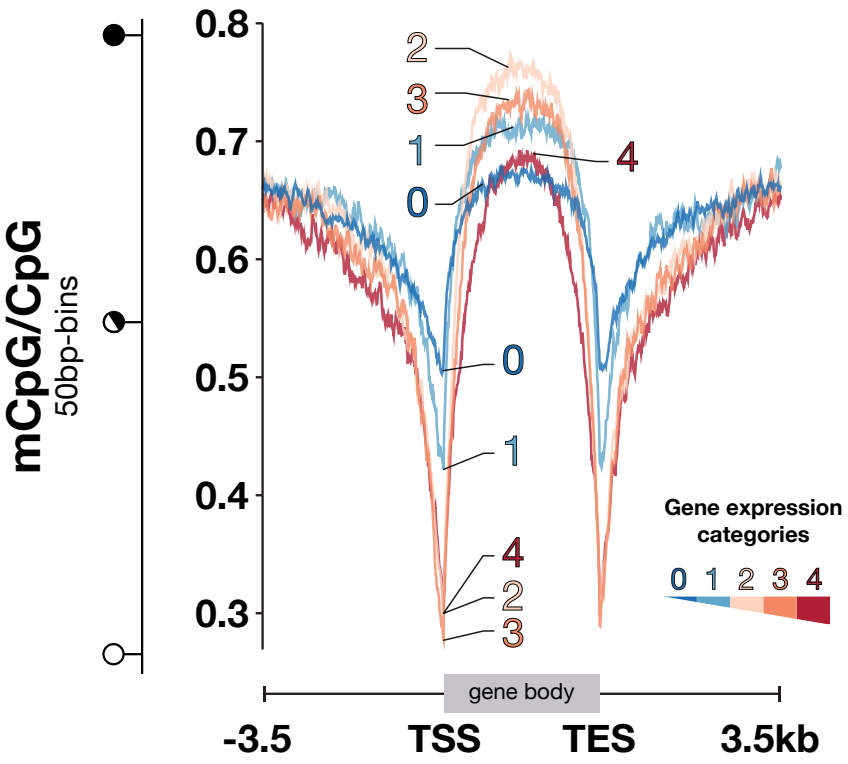


Fig. 2.30 DNAme levels at promoters and gene bodies is associated with differential transcription activity. Methylome profiles around the transcription start and end sites (TSS and TES, respectively) and in gene bodies (x-axis) for each gene expression category. y-axis represents averaged DNAme levels in *M. zebra* males, liver (n=3, averaged mCG/CG in 50bbp-long windows, genome-wide). Averaged expression values (mean TPM values per gene) in liver of *M. zebra* males (n=3). Genes were grouped into 11 categories by transcription (deciles plus one category for silent ("OFF", 0) genes), from low (1), intermediate (2), high (3) to highest (4) expression values (five of the deciles are shown for clarity).

2.8.1 Impact of inter-species liver methylome variation on transcriptome

I then sought to investigate the link between high DNA methylation variability and its impact on gene expression levels. Of note, all DMRs, both tissue- and species-specific, are used for this analysis. This allows for a broader correlation to be made, including DMRs relevant in fully differentiated liver, that could have been established very early on during development and could therefore be present in muscle tissues as well ('embryonic relics'). In particular, only 63% of all the DMRs (n=6,974) could be associated with a gene (i.e. DMRs located either in promoter region and/or exon/intro of known genes or lying in intergenic transposon

elements in the vicinity of genes; see Fig. 2.25b). Many DMRs are intergenic and therefore cannot be linked to any genes.

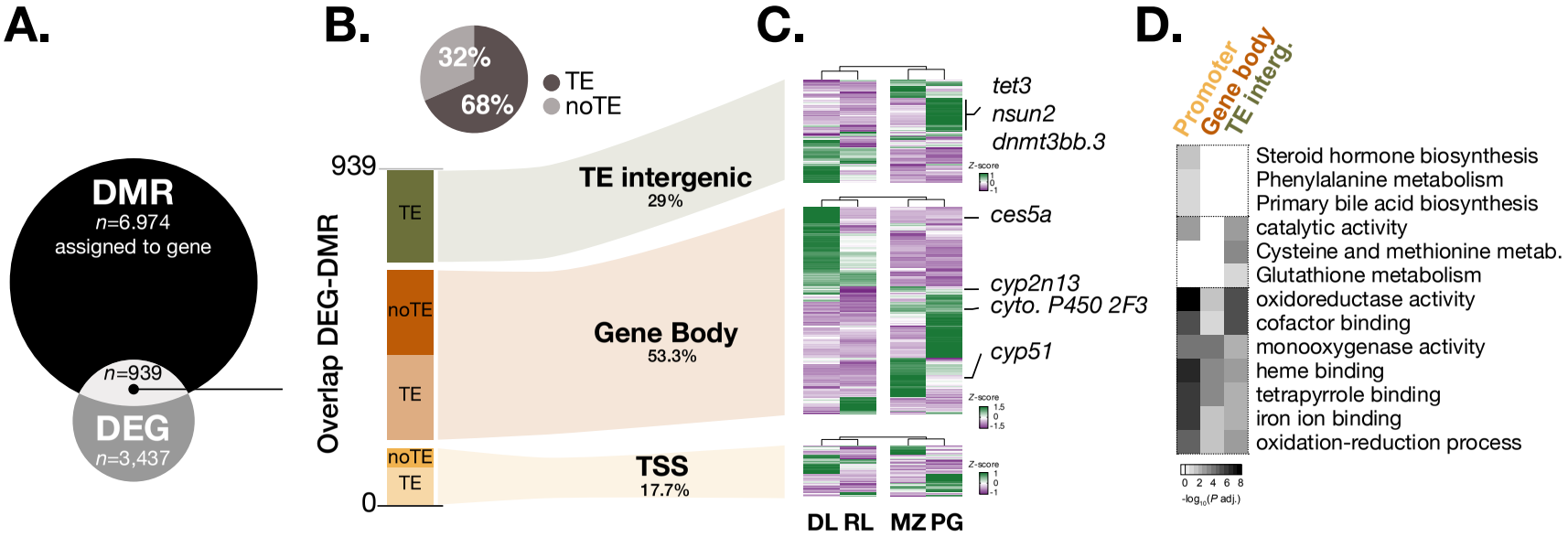


Fig. 2.31 Differentially expressed genes showing DNAme variation in livers are related to metabolic and developmental processes. A. Number of genes (overlap, *n*=939) that are both differentially expressed in liver (DEGs) and also differentially methylated in liver (DMRs). Here, the total number of liver inter-species DEGs found is shown (see Fig. 2.27 a., dark blue bar). The total number of liver inter-species DMRs (*n*=6,974) that could be spatially allocated to one gene. **B.** Genomic localisation of DMRs associated with DEG in liver. DMRs were localised in nearby intergenic transposons ("TE intergenic", from 500bp to 4kbp from gene bodies), in promoter regions (TSS) and gene bodies. Pie chart summarising the overall proportion of DMRs associated with TE sequences. **C.** Heatmap of gene expression profile in liver for DEG with DMRs at nearby TE (*n*=242), gene body (*n*=386) and TSS (*n*=161). Averaged TPM values per gene per species, *Z*-score (scaled). **D.** GO enrichment analysis for DEG that are associated with DMRs. Heatmap of -$\log_{10}$(adj.pval.) for each GO term (not significant in white).

Of 6,974 DMRs that could be associated to one gene (i.e. DMRs located either in promoter region and/or exon/intro of known genes or lying in intergenic transposon elements in the vicinity of genes), 939 (15%) were correlated with significant transcriptional changes at the respective genes. Close to 3,500 genes are differentially expressed in any liver of Lake Malawi cichlids (Fig. 2.27a). 27% of them exhibit significant degree of liver methylome variation (Fig. 2.31a.). Of important note, such an overlap could be considered somewhat small if one assumed that any change in methylation would directly affect transcriptional activity of one particular gene in a fully differentiated cells. However, liver methylome not only harbours the epigenetic patterns essential to the cellular identity of a fully differentiated hepatic cells, but also shares methylation patterns with muscle tissues (Fig. 2.24), hinting at species-specific methylome, independently of the tissues analysed. This suggests that some patterns might instead reflect embryonic methylation memories [169], that might have been established very early on during development, with important embryonic functions (for example, in fine-tuning gene expression patterns). Throughout cell divisions, such species-specific, multi-tissue methylation patterns might have been maintained, with no clear biological functions in fully differentiated cells, as lacking the machinery capable, probably methyl-sensitive DNA-binding factors, of interacting with it. Generating WGBS and RNASeq of different tissues, as well as of embryonic tissues could test the hypothesis that some DMRs are functionally relevant only during development, resulting in multi-tissue DMRs in fully differentiated tissues. Finally, in light with gene ontology enrichment analysis hinting at DMRs associated with key developmental genes (i.e. homeobox genes), I expect most of the methylome variation to participate in early life development. Nevertheless, it remains that a considerable amount of liver-specific methylome variation is associated with changes in gene expression and might be explained adaptive changes in response to distinct diet (Fig. 2.31). Further characterisation of DEG-DMR is being carried out at the moment to identify the best gene candidates. This could lead to a more targeted approach for example, aiming at validating the expression and methylation for specific candidates, using liver and muscle tissues from other Malawi cichlid species.

Furthermore, it is important to note that the transcription of certain genes might be orchestrated by other cellular factors, in concert sometimes with DNA methylation. These factors (such as methyl-insensitive TF) might be specifically expressed in liver tissues and may underlie changes in transcriptional network.

2.8.2 Localisation of inter-species DMRs associated with transcriptional changes

Intergenic TEs

Interestingly, more than two-thirds of the DNAme variation associated with significant changes in transcription (DMR-DEGs) is located in transposon or repeat sequences (TEs, Fig. 2.31b.): this includes DMRs overlapping TEs that are either in promoter regions (12.5% of all DMR-DEGs) or gene bodies (27.1%). This also includes DMRs overlapping intergenic TEs located nearby genes (29% of all DMR). This last category of TE-DMRs (arbitrary defined as located >500 bp and <4kbp upstream of TSS) could include extended promoters or nearby enhancers. Remarkably, these DMRs are found significantly in genes involved in oxido-reduction processes (fatty acid metabolism, such as many cytochrome P450 genes), but also, less expectedly, in genes involved in cysteine and methionine metabolism (Fig. 2.31d.). This includes specific enzymes with methyltransferase activity, such as the *de novo* DNA methyltransferase *dnmt3bb.3* [4] and the t-RNA methyltransferase *nsun2*. The enzyme *dnmt3b* is known to be involved in the *de novo* DNA methylation patterns, adding new methyl groups onto cytosines, as opposed to solely maintaining DNA methylation patterns upon cell division for example. Other key epigenetic factors exhibit DNAme variation in nearby TE sequences and are differentially expressed in Lake Malawi cichlids, such as the methylcytosine dioxygenase *tet3* (LOC101464948), catalysing the conversion of methylated cytosines into hydroxymethylated cytosines (5hmC); 5hmC is a stable modification with unclear biological functions [66], which could be an intermediate modification in the complete erasure of modified cytosines. DNA demethylation process could be required to reshape methylome in liver tissues, and perhaps to remodel chromatin state as well in response to external stimuli, to eventually change the transcription activity of certain genes. In particular, both *tet3* and *nsun2* show increased gene expression specifically in livers of *P. genalutea* (not in muscle tissues), although at low levels (Fig. 2.31d.). It has been shown that TET1-mediated DNA methylation would promote liver regeneration via cell reprogramming [70]. In addition, the activity of TET enzymes can be modulated through the interaction with metabolite- and nutrient-sensing proteins, thus promoting DNA demethylation in response to environmental changes [65, 67–69]. Of note, *P.genalutea* shows the highest levels of uniqueness in terms of transcriptome: a large fraction of all the DEGs found in livers of Lake Malawi cichlids are species-specific, and mostly upregulated in PG (Fig. 2.28a.).

[4] LOC101473965

However, when investigating TE-DMRs that lie in the vicinity of genes, it is tempting, based on their proximity, to allocate to them *cis*-regulatory roles, such as enhancers or ectopic promoters. It would be therefore useful in the future to experimentally characterise these elements (e.g., ChIP-seq, chromatin immunoprecipitation followed by DNA sequencing in order to identify binding sites of DNA-interacting proteins, such as TFs or histones), which will allow for a more reliable association to genes. This would be of particular relevance as most of DMRs are found in intergenic regions (ca. 56% all of DMRs found in liver; Fig. 2.22b).

Gene bodies

When not associated with TE sequences, most of the variation in liver methylome correlated with transcriptional changes is located in gene bodies (53.3%). DNAme variation in gene bodies could be involved in splicing event processes [86] (see section 1.4.2), as well as in transcriptional regulation via intronic enhancers. Like for DMRs in intergenic TEs and TSS, DEG with signifiant liver methylome variation in gene body are enriched for functions linked to oxidation-reduction processes. Of note, the pelagic species *D. limnothrissa* accounts for the largest (39%) fraction of DEGs associated with gene body DMRs. These genes are particularly involved in dicarboxylic acid metabolic processes. The gene carboxylasesterase 5A, *ces5a*, shows an almost 6-fold increase in expression and is associated with hypomethylation levels in the first intron of the gene in DL liver. This enzyme is involved in the metabolism of fatty acyl and cholesterol ester (figure not shown).

Promoters

Variation of DNAme in promoter regions (TSS-DMRs) makes up less than 18% of liver methylome variation associated with transcriptional change in gene expression (importantly, 70% of TSS-DMRs are spanning TEs; Fig. 2.31b). Yet there is enrichment for genes important for liver metabolic functions, mostly linked to fatty acid metabolism, such as primary bile acid production, amino acid production and steroid acid biosynthesis. This could be particularly relevant for adaptation to carnivorous and herbivorous diets, which would imply efficient assimilation and metabolism of different types of fatty acids (Fig. 2.31d.). Differentially expressed genes showing liver methylome variation in their promoter regions show patterns of transcription mostly unique to one species (for upregulated genes). For example, eight genes are strongly upregulated, most of them involved in oxidation-

reduction processes[5]. *bco1* is involved in vitamin A metabolism – beta-carotenes being converted into retinal, with liver being a major storage of beta carotene. Retinol is also an important molecule involved in the visual system (photoreceptor), and therefore may be related to visual adaptation to dimly-lit habitats of the lake (occupied by the cichlid *D. limnothrissa*). Many specific cytochrome P450 genes, involved in xenobiotic and lipid metabolism (signalling pathway, steroid biosynthesis, vitamin A pathway) among others, are differentially expressed in specific species.

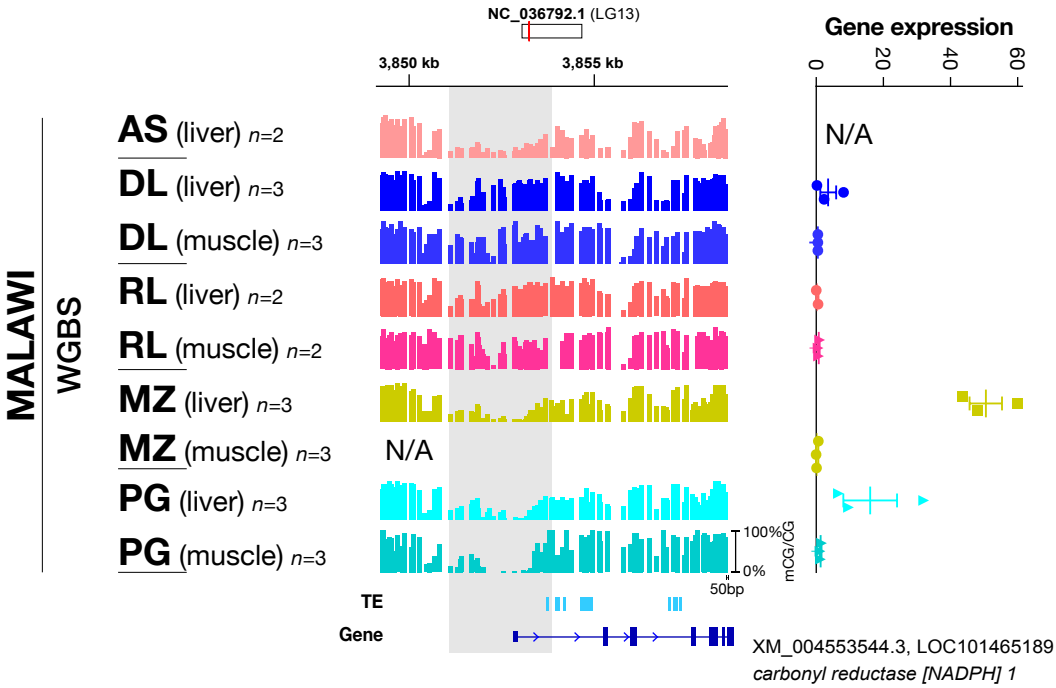


Fig. 2.32 Hypomethylated promoter of the carbonyl reductase *nadph1* is associated with increased expression levels in liver. Example of a differentially expressed gene, carbonyl reductase *nadph1*, in livers of Lake Malawi cichlids, showing significant variation in DNAme at promoter regions of 5 different species. **Right**: genome browser view of WGBS liver methylomes. Averaged mCG/CG levels in 50bp-long windows per species, genome-wide. **Left**: gene expression levels for both muscle and liver samples, in TPM values. N/A refers to tissues for which no sequencing data is (yet) available.

The gene carbonylreductase *nadph1* is significantly up-regulated in both herbivore species (MZ and PG) in liver tissues only (Fig. 2.32). Carnivore species (DL and RL) do not express this gene in liver or muscle tissues. Interestingly, the promoter region of this gene shows significant hypomethylated levels in algae-eating rock-dwellers (MZ and PG) in both muscle and liver tissues, while both pelagic carnivores show almost 100% methylation over this ca.

[5] *dio1*, iodothyronine deiodinase 1; LOC101476301, cytochrome P450 2K1; *cyp2n13*, cytochrome P450 2J6; *bco1*, beta,beta-carotene 15,15'-dioxygenase)

3.5 kbp-long promoter in both tissues. This is an example of a species-specific DMR, with a tissue-specific gene expression profile and methylation pattern probably established very early in development in specific species [169]. It is likely that the cellular factors/methyl-sensitive readers, interacting with that promoter are only active/present in liver and might be methylation-sensitive factors (such as transcription factors). The gene carbonylreductase *nadph1* is involved in the metabolism of fatty acid, in particular of arachidonic acid. The enzyme has been shown to be up-regulated in fish in response to environmental threats, in order to provide protection against oxidative stress [171] and might be important in fatty acid-mediated cellular signalling.

To conclude, most of the DNAme variation that is associated with changes in gene expression is located in gene bodies and intergenic repeat regions in the vicinity of genes. DNA variation associated with DEGs is significantly located in repeats and transposable elements. It would be of particular interest to characterise these repetitive elements showing DNA variation in order to evaluate their age, and activity in other Lake Malawi cichlids. Possible knock-out of the most promising TE-DMRs, in particular these in promoter or enhancer elements, would provide further evidence of the link between transcriptional changes associated with differential methylation.

2.9 Discussion and future work

In this chapter, I characterised the liver methylome landscape of five eco-morphological distinct wild-caught cichlid species of Lake Malawi. This study represents the first whole-genome bisulfite sequencing analysis investigating epigenetic-mediated trophic adaptation in a natural population of speciating vertebrates.

2.9.1 Cichlid methylomes share conserved features with other vertebrates

High CG methylation levels genome-wide

The first analysis of chapter 2 pertains to the characterisation of the methylome landscape in Lake Malawi cichlids. Overall, cichlids of Lake Malawi share many conserved features with other vertebrate genomes. Firstly, cytosine methylation is highly sequence specific in cichlids and other vertebrates, in that almost all methylated cytosines are located in CG dinucleotide sequences (Fig. 2.6). Methylation of DNA cytosines in other sequence contexts is rare in vertebrates (low levels at detected in neural and stem cells in certain vertebrates), while being more frequent in plants in particular at TE sequences [35, 78, 169]. Another characteristic of cichlid methylome shared with most vertebrates is that more than 70% of cytosines are methylated genome-wide in CG dinucleotide sequence context (Fig. 2.8). Overall, just like most vertebrates, the cichlid methylome is globally highly methylated in the two tissues analysed in this thesis, which is expected to be true for many somatic cells in cichlids (Fig. 2.9). Plants and invertebrates usually exhibit variable levels of cytosine methylation, which can sometimes also be restricted to gene bodies and/or repeats [41, 40]. In cichlids and vertebrates, such high methylation levels are thought to have come at the cost of a CG depleted genome (Fig. 2.11), due to the propensity of mC for spontaneous deamination [45, 46].

Low CG methylation levels at promoters and CGIs

While most of the genome in cichlids and vertebrates exhibit high levels of CG methylation, some specific genomic regions are unmethylated. In cichlids, as well as in most vertebrates [41], promoter regions are devoid of methylated cytosines (Fig. 2.10). Interestingly, while two-thirds of the promoters in mammals are associated with CpG-rich regions (CGIs), less than a fifth of cichlid promoters are CpG-rich sequences (Fig. 2.13b), although almost all CGI-containing promoters in cichlids are unmethylated (Fig. 2.13d), like in other amniotic

vertebrates [44]. Furthermore, I showed that methylation at promoter regions was negatively correlated with gene expression in both liver and muscle tissues (Spearman correlation rho=ca. -40%), while gene body methylation was positively but weakly correlated with transcriptional activity (Spearman correlation rho= 2-20%; see Figs. 2.29 and 2.30). Gene body methylation at actively transcribed genes is a widely conserved feature, from vertebrates, plants to invertebrates, while lack of methylation at TSS regions might mostly be a characteristic of vertebrate genomes only [41]. Transcriptional regulation, such as alternative splicing or transcription repression/initiation, is highly dependent on differential methylation levels in gene bodies, enhancers and promoters in vertebrates, and is thought to mostly happen through the recruitment of specific methyl-sensitive DNA-binding factors, such as transcription factors or repressors just to cite a few [78, 172, 86]. Such interactions, organised in a highly timely-manner, are essential for organismal development, embryogenesis and cell differentiation in most vertebrates and might be responsible for phenotypic plasticity. Loss-of-functions associated with most factors involved in the metabolism of cytosine methylation usually lead to embryonic lethality [63, 17, 44], highlighting the biological relevance of DNA methylation in vertebrates.

High CG methylation levels at TE sequences

The activity of transposable elements is under the tight control of the host. In eukaryotes, post-transcriptional silencing of TE transcripts is carried out by small non-coding RNAs (such as piRNAs and siRNAs, among others). Furthermore, DNAme-mediated mechanisms to silence transcription activity directly at TE genes are a conserved feature in eukaryotes, and can be mediated by sncRNAs as well as KRAB-ZFPs in higher vertebrates (excluding teleost fishes) [103]. In cichlids, most TE families are heavily methylated (Fig. 2.17), which suggests that a DNAme-based mechanism to silence TE activity also exists. Likewise, in zebrafish, TE sequences exhibit high levels of methylation. Moreover piRNAs are expressed in the germ line of both sexes and target genomic TE sequences, hinting at DNAme-mediated repression of transposon activity driven by sncRNAs [100]. Nile tilapia has been shown to express both piRNAs and PIWI proteins [99], known to interact with piRNAs to possibly mediate DNAme-based transcription silencing. Furthermore, 35.4% of the genome of the Lake Malawi cichlid *M. zebra* is composed of TE sequences, half of them being DNA transposons (Fig. 2.15a,b), which could be a conserved feature of teleost fish only, as the genomic TE landscape in most mammals is composed of retrotransposons [93].

In summary, the methylome of Lake Malawi cichlids shares many features seen in other vertebrates and is therefore expected to exert similar regulatory functions, in particular related to transcriptional control and repression of TE activity.

2.9.2 Species- and tissue-specific methylome variation in Lake Malawi cichlid fishes

I then sought to characterise the methylome variability between five different Lake Malawi cichlids, exhibiting distinct trophic adaptations. Strikingly, the liver methylomes in Lake Malawi cichlids exhibit important variability at conserved underlying DNA sequences and appear to be both tissue- and species-specific (Figs. 2.18 and 2.19), in that a phylogeny of each species and possibly based on the diet could be reconstructed based on methylome divergence only in the case of both tissues. Interestingly, promoter regions, CpG islands (mostly outside TSS regions with potential regulatory functions [173]; also known as *orphan* CGIs) and transposable elements are the genomic regions showing the highest levels of epigenetic diversity (Fig. 2.23), highlighting the importance of such regulatory regions in possibly mediating phenotypic plasticity in the context of adaptation to different diets in cichlids.

2.9.3 Vestigial DMRs in fully differentiated tissues

I described that some epigenetic variation may not only be liver- or muscle-specific but rather multi-tissue species-specific, in that both liver and muscle share specific DNAme variation. In fact, 35-50% of liver methylome variation is shared with muscle tissues, and inversely, 59-64% of inter-species methylome variation observed in muscles is also present in livers (Fig. 2.24), which indicates that liver methylome exhibits slightly more tissue-specific inter-species DMRs, in line with the putatively significant evolutionary pressure on liver adaptation. I postulate that such a multi-tissue variation is likely to be found in other tissues (e.g. heart, brain), and that it might represent embryonic memories or vestigial methylation patterns [169], in line with gene ontology enrichment highlighting key developmental genes [169]. I hypothesise that a large fraction of epigenetic variation might indeed be relevant/active in early stages of development, promoting long lasting, species-specific changes in phenotypes (as likely located in distal regulatory regions [Fig. 2.23] and Ref. [169]), while being probably dormant in adult tissues, in liver and muscle tissues (unless they exert biological functions relevant to multiple tissues [housekeeping roles, e.g.]). Upon cell division during development, such vestigial DMRs are faithfully passed on from one mother cells to the

daughter cells without epigenetic erasure in fully differentiated tissues. The gene visual system homeobox 1 (*vsx1*), participating in eye development, is a good example of putative vestigial multi-tissues DMRs, as its function is likely to be restricted to early stages of development (Fig. 3.8). Consequently, this warrants further comparative characterisation of the methylome and transcriptome at different developmental time points to understand the role of such vestigial methylation variation during embryogenesis. In addition, a complete identification of genomic regulatory regions (enhancers, promoters) would allow for a detailed characterisation of vestigial DMRs, highlighting their possible functions.

2.9.4 DNA methylation variation is correlated with differential transcriptional activity at key metabolic and developmental genes

Finally, I then characterised the interplay between inter-species methylome variation and transcriptional activity in liver tissues to test the hypothesis that any divergence in DNAme between species could underlie considerable phenotypic plasticity, in particular related to traits associated with trophic adaptation.

Strikingly, a considerable amount of differentially expressed genes (DEGs) could be associated with methylome variation (DMRs) in liver (Fig. 2.31a). In total, 27% of all inter-species liver DEGs were linked to liver DNAme variation. Importantly, this might represent an underestimation, as DMRs were spatially assigned to genes, in that no annotation of regulatory elements, such as enhancers or promoters, were available at the time of this study. Inversely, ca. 14% of all species-specific DMRs were associated with DEGs, which might probably be an underestimation as well, in that vestigial/multi-tissue DMRs were not excluded from the analysis and might putatively exert no functions in adult liver, out of the developmental context.

Furthermore, most genes associated with DMRs were differentially expressed in a highly species-specific manner, probably highlighting the unique liver-related phenotypic adaptation in these eco-morphologically distinct species. Even the more closely-related herbivorous, rock-dwelling species (*M. zebra* and *P. genalutea*) differ in their DEGs associated with DMRs (Fig. 2.31c, bottom). Interestingly, gene bodies accounted for the majority (53.3%) of all liver transcriptional changes associated with methylome variation. This hints at regulatory roles for gene body methylation in regulating and interacting with the transcription machinery (RNA pol II, for example) [86]. Promoters, important actors in controlling transcription in vertebrates, make up close to 18% of the transcriptional variation linked to species-specific DNA methylation.

Importantly, DNA methylation variation associated with significant changes in gene expression consistently pertains to similar biological processes regardless of its genomic localisation (Fig. 2.31d), suggesting that methylome variation at both promoters and gene bodies could act in concert in order to promote phenotypic plasticity. In particular, specific hepatic processes seem to show transcriptional changes associated with DNAme variation at promoters and gene bodies (Figs. 2.31d and 2.32). The biosynthesis of steroid hormones and of lineage-specific types of bile acids, important in lipid metabolism/breakdown and in energy homeostasis [174], appears to be targeted by DNA methylation variability at their transcriptional start sites.

Finally, another important observation is the predominant implication of TE sequences in producing species-specific liver methylome variation, hinting at a role in promoting phenotypic plasticity. Although TE elements are only slightly enriched overall for DNAme variation (Fig. 2.23), they account for more than two-thirds of the transcriptional changes associated with liver methylome variation, in particular at promoter regions (Fig. 2.31b). Due to the high propensity of TE sequences to contain regulatory elements, novel genomic insertion of TEs could be accompanied by changes in transcription in neighbouring genes [38]. Differential methylation states at such TE-derived ectopic promoters have been linked to be influenced by environmental perturbations, possibly resulting in altered gene expression activity – see the agouti mouse model and in toad flax for example, in section 1.4.5 and Ref. [92, 126]. It would be interesting to further characterise the implication of TE-DMRs associated with DEGs, which is the focus of current work.

Adaptive traits, such as the ones associated with liver functions as diverse as lipogenesis, steroid hormone production and xenobiotic metabolism, can greatly participate in the adaptation to different diets in Lake Malawi. Liver methylome variation was associated with transcriptional changes in such biological processes in a species-specific manner. Altogether, these results postulates an adaptive role of DNAme variation in facilitating plasticity of liver-related phenotypes. It is tempting to speculate that functional methylome variation could be fixed in different species of Lake Malawi cichlids and could participate in the short- and long-term phenotypic divergence by shaping transcriptional landscape. Such epigenetic variation could be selectable as affecting phenotypic plasticity of adaptive traits.

Although these results highlight significant liver species-specific methylome variation associated with possible traits related to diet adaptation, it is crucial to investigate further whether such variation could be fixed in a population and could be transmitted to subsequent generations in an environment-independent manner. In other words, how much of the methylome landscape is shaped by the local environment as opposed to being environment-

independent? I speculate that a large fraction of methylome variation is directly influenced by environmental factors, underlying selectable phenotypic plasticity. However, I expect some fixed and heritable DNAme variation associated with adaptive traits. For this latter, the implication of genetic polymorphism in generating epigenetic variation can not be ruled out at the moment and will be investigated in chapter 4. Although this thesis analysed liver methylome variation at conserved underlying DNA sequences, genetic polymorphism in *trans* may also modify the methylome landscape [78]).

2.10 Detailed methodology

Overview

All Malawi cichlid fish were wild caught. Upon collection (during the field trip, 2015-2016), tissues were placed in RNA*later*, and were then stored at -20°C. The protocol explained hereafter was optimised for tissues stored in RNA*later*. The main difference with fresh tissues would be the additional purification step (to clean samples of any salt carry-over), not necessary for fresh tissue, and the slight degradation of HMW gDNA observed for wild caught fish; see Fig.2.4a.

Table 2.4 Sampling size - Methylome of Lake Malawi cichlids

Species	WGBS [1]		RNAseq [2]	
	liver	muscle	liver	muscle
Aulonocara stuartgranti	2*	0	0	0
Maylandia zebra	3	0	3	3
Rhamphochromis longiceps	2	2	3	3
Petrotilapia genalutea	3	2	3	3
Diplotaxodon limnothrissa	3	2	3	3

* one female and one male.
[1] WGBS, whole genome bisulfite sequencing
[2] RNAseq, total RNA (transcriptome) sequencing

2.10.1 Protein sequence homology

TET and DNMT proteins were found in *Maylandia zebra* with BLASTp using canonical amino acid sequences of either *Homo sapiens* or *Danio rerio* for each protein analysed (% seq identity, query coverage and genomic location of isoforms were the major criteria of selection). Phylogenetic trees based on sequence homologies were then generated by first aligning all the sequences (MUSCLE, 3.7), curating the alignments (Gblocks 0.91b) and finally building and predicting the sequence phylogeny (PhyML 3.0) – all done using http://phylogeny.lirmm.fr/phylo_cgi/index.cgi [175]. Trees were visualised using FigTree (v1.4.4).

The following NCBI Reference Sequences were used:

TET3: XP_014263906.2 (mz), NP_001314875.1 (dr), O43151 (hs)

TET2: XP_004549491.1 (mz), XP_005159960.1 (dr), Q6N021 (hs)

TET1: XP_004551793.2 (mz), XP_021328376.1 (dr), Q8NFU7 (hs)

mz, *Maylandia zebra*; dr, *Danio rerio*; hs, *Homo sapiens*.

2.10.2 Isolation of genomic DNA and NGS Library preparation

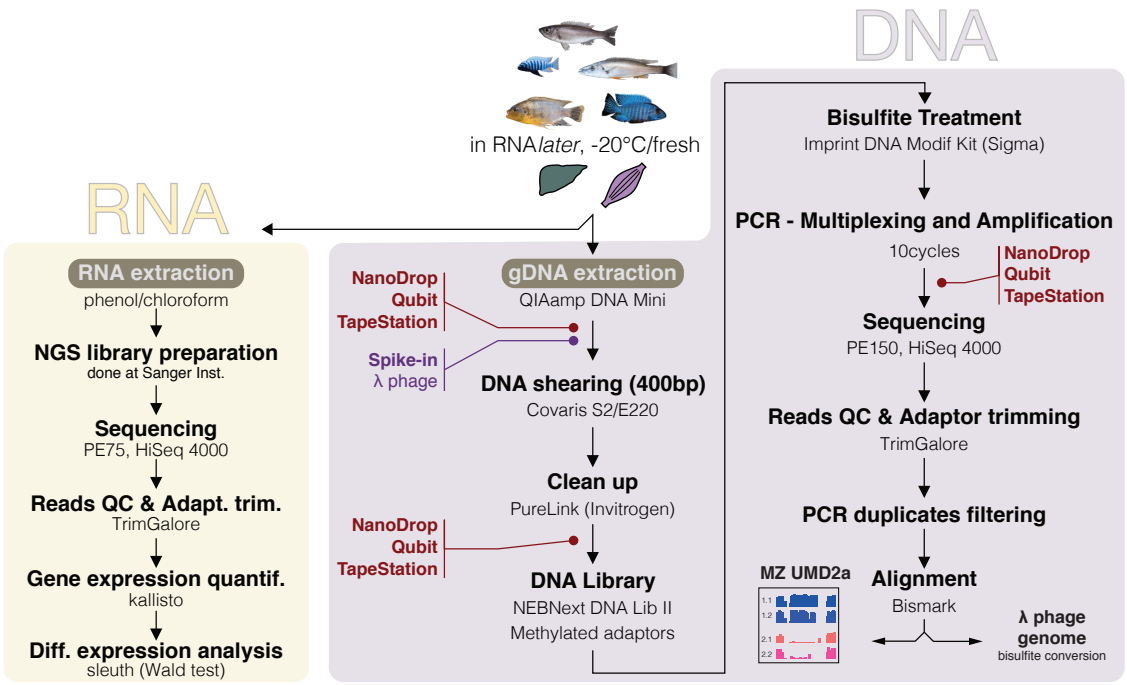


Fig. 2.33 Overview of the methods used to generate WGBS and RNAseq data. Summary of the main methods used to extract high-molecular-weight genomic DNA and RNA from both liver and muscle tissues and to prep NGS libraries for both WGBS and RNAseq. Bisulfite conversion of genomic fragments is performed after DNA library prep. Once sequencing is done, adaptor sequences in read sequences are trimmed out and low quality reads are filtered out. Then, adaptor-trimmed reads are mapped to both Lake Malawi reference genome *(M.zebra*, UMD2a) and to lambda phage genome to assess bisulfite conversion rate. RNAseq reads were mapped to the same reference genome. Gene expression was quantified to finally perform differential gene expression analysis.

Figure 2.33 summarises the main method chosen to produce WGBS libraries. In detail, high-molecular-weight genomic DNA (HMW-gDNA) was extracted from homogenised tissues (<25mg) using a silica-column based approach QIAamp DNA Mini Kit (Qiagen 51304) according the manufacturer's instructions. HMW-gDNA was then fragmented using sonication to the target size of 300 - 400bp (Covaris, S2 and E220; Fig. 2.34). Fragments were then purified (mainly to remove carried-over salts of RNA*later*) with PureLink PCR Purification kit (ThermoFisher). Before any downstream experiments, quality and quantity of gDNA fragments were both assessed using NanoDrop, Qubit and Tapestation (Agilent). To estimate bisulfite conversion efficiency, the unmethylated genome (48.5kbp) of the lambda

phage cl857 Sam7, isolated from infected GM119, a Dam$^-$ Dcm$^-$ *E. coli* strain (lacking the methyl transferase activity for both cytosine and adenine nucleotides; Promega) was spiked in (0.5% (w/w)) before sonication.

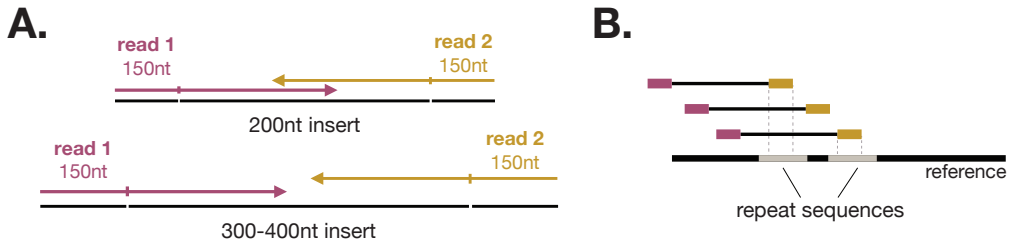


Fig. 2.34 Overview of paired-end sequencing reads. A. Paired end sequencing enables both ends (reads 1 and 2) of the DNA insert to be sequenced. An insert size of 300-400nt long is ideal to avoid any sequencing overlap (bottom). **B.** As the distance between each paired reads is known, higher mapping efficiency is to be expected, in particular at repetitive regions, such as transposons. Figures modified from Illumina website.

Typically, 200ng of sonicated fragments were prepped in the readiness of NGS with NEBNext Ultra II DNA Library Prep (New England BioLabs, E7645S). Briefly, blunt fragments are first end-repaired and A-tailed using T4 DNA polymerase and Klenow Fragment. They are then ligated on both flanks with Illumina methylated adaptors (NEB, E7535S). Adapted fragments were then purified with Agencourt AMPure Beads at a 0.8x ratio (Beckman Coulter, Inc). Libraries (ca.50ng) were then treated with sodium bisulfite (see section 2.10.3 and Fig. 2.35) according to protocol (Imprint DNA Modification Kit; Sigma, MOD50). Converted libraries were then indexed (NEBNext Multiplex Oligos for Illumina, NEB E7335S) and amplified by PCR (10 cycles) with KAPA HiFi HS Uracil$^+$ RM (KAPA Biosystems), followed by a purification step with 0.7x volume of Agencourt AMPure Beads. Size and purity of libraries were determined using Tapestation and quantified using Qubit (Agilent). Indexed libraries were sequenced on HiSeq 4000 (High Output mode, v.4 SBS chemistry, at CRUK, Cambridge Institute, UK) to generate paired-end 150 bp-long reads. Of note, both male and female specimens of *A. stuartgranti* only were sequenced on HiSeq 2500 to generate paired-end 125bp-long reads.

2.10.3 Bisulfite conversion - detailed overview

In order to distinguish between modified cytosines and unmodified (5-methylcytosine, mC, and 5-hydroxymethylcytosine, 5hmC), DNA is treated with sodium bisulfite. Upon bisulfite treatment, only unmodified cytosines will be converted into uracil through a first sulfonation

reaction, followed by a deamination process and a final desulfonation step after DNA denaturation by heat into single-stranded fragments (Fig. 2.35a). Modified cytosines are immune from this conversion (Fig. 2.35b). As a result after PCR amplification, unmodified cytosines will be read as thymine upon sequencing, while modified cytosines, as cytosines (Fig. 2.35c.). It is therefore essential to monitor the efficiency of the conversion, as any failure in converting any unmodified C into U would result in an erroneous methylation determination. To this aim, a spike-in was used for all WGBS samples as conversion control (see section 2.10.2): 0.5% w/w of genomic DNA of lambda phage produced in a mutant strain of *E.coli*, lacking any methyltransferase activity (dam^- and dcm^-) was spiked in before the sonication step. As only unmodified C is present in the phage genome, any failed conversion of unmodified C will result in a false positive mC call. The conversion efficiency is therefore calculated based on how many false-positive mC were sequenced in the lambda phage genome. The mapping software bismarck [161] is able to infer methylation status at each CpG site in the genome by comparing mapped converted libraries with reference genome [6].

2.10.4 Quality of WGBS sequencing reads

`TrimGalore --paired --fastqc --illumina` (v0.5.0, Babraham Inst.; github.com/FelixKrueger/TrimGalore) was used to determine the quality of sequenced read pairs and to remove Illumina adaptor sequences and low quality reads (Phred quality score <20).

2.10.5 Alignment and visualisation of mCG sites

All adaptor-trimmed paired reads were aligned to the reference genomes of *M. zebra* cichlid (assembly version, `UMD2a`) and to the lambda genome (to determine bisulfite non-conversion rate) using `Bismark` (v0.20.0 [161]), a mapping software making use of `Bowtie2`. The alignment parameters were as follows: 1 mismatch allowed with a maximum insert size for valid paired-end alignments of 500bp (`bismark -p5 -N 1 -X 500`). Clonal mapped reads were removed using `deduplicate_bismark -p`.

[6] Both the reference genome and sequenced reads will be converted *in silico* twice (one C-to-T conversion and one G-to-A conversion). Converted reads are mapped then to the respective converted reference genome, and the unique best alignment is then determined from the two mapping processes (these are directional libraries, in that complementary strands will not be sequenced). The methylation state at each CpG is determined by comparison with the reference original genome (reviewed in Fig.1 in [161])

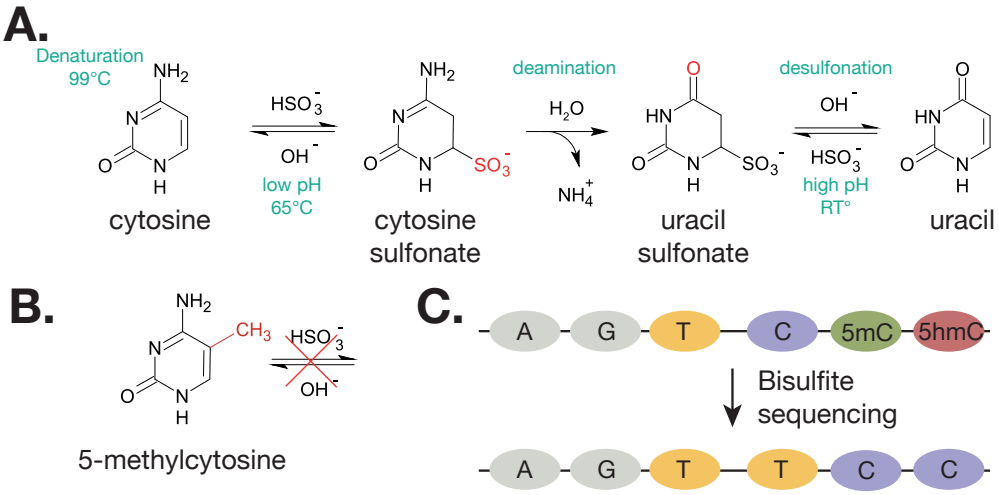


Fig. 2.35 Bisulfite conversion of unmodified DNA cytosine. Principal chemical reaction of WGBS enabling distinction between unmodified and methylated cytosines upon sequencing. **A.** Unmodified cytosine is converted to uracil through a first sulfonation reaction, followed by a deamination process and a final desulfonation step after DNA denaturation by heat into single-stranded fragments. **B.** Modified cytosines are immune from this conversion. **C.** Upon whole genome bisulfite sequencing, unmodified cytosines are read as thymines, while modified cytosines, as cytosines. **a.** and **b.**, adapted from NEB website. **c.** adapted from Ref. [176].

Methylation at CpG sites were called using `bismark_methylation_extractor -p --multicore 9 --comprehensive --no_overlap --merge_non_CpG` and `methylpy`.

Downstream analysis, including DMR calling and gene expression correlation, were done with `R` package `DSS` (v2.32.0 [177]), `R` (v3.6.0) using custom scripts. `Samtools` (v1.9) and `bedtools` (v2.27.1) were used to generate, handle mapped reads and generate averaged methylation levels across non-sliding windows of various sizes genome-wide. `R` packages `ggplot2` (v3.1.1) and `pheatmap` (v1.0.12) were used to visualise methylome data and to produce non-biased clustering based on methylome similarities. `R` in-house commands were used to produce principal component analysis (`prcom`, centred and scaled), Spearman's correlation (`cor`), Euclidean distances (`dist`). Statistical tests, including Kruskal-Wallis H test and Dunn's multiple comparisons test were performed with `R` using the package `FSA` (v0.8.24). Of note, for all methylome analyses apart from DSS-mediated DMR calling, only CpG sites with >4 and $\leq$ 100 unique mapped reads were used for analysis.

The genome browser IGV [178] (v2.5.2) was used to visualise DNA methylation genome-wide (in 50bp-long windows).

2.10.6 DMR calling

Differentially methylated regions (DMRs) were called using the `R` package DSS (Dispersion Shrinkage for Sequencing data with single replicates)) [177]. The authors have developed a statistical method to predict DMRs that takes into account spatial correlation of CpG methylation, read depth and biological variation. In brief, DSS includes a binomial distribution test capturing the random sampling process of the WGBS dataset, a beta distribution modelling the biological variation among replicates, and a smooth function accounting for the spatial correlation among nearby CpG sites (estimated through an empirical Bayes (EB) procedure). Finally, significantly differentially methylated single CpG is calculated by performing Wald test. Finally, regions with $\geq$ 5 CpG sites showing significant difference in methylation levels ($\Delta \geq$ 25%, pvalue < 0.05) and if within 50bp distance of each other were grouped into DMRs.

Enrichment analysis for genomic elements of DMRs was done using `bedtools shuffle` with 10 iterations.

2.10.7 Genomic annotations

The reference genome *M. zebra* (UMD2a) was used to generate all annotations.

Custom annotation files were generated and defined as follows:

Promoter regions

Promoter regions were defined as TSS$\pm$500 bp.

Gene bodies

Gene bodies included both exons and introns and other intronic regions, and excluded the first 1kbp regions downstream of TSS to avoid any overlap with promoter regions.

Transposable and repetitive elements

Repetitive regions were modelled using RepeatModeler (v1.0.11), masked and classified using RepeatMasker (v4.0.9.p2).

CpG islands

CpG-islands (CGIs) were predicted with makeCGI [164] - a software using a Hidden-Markov model approach to define CGI in an unbiased way in any species (without pre-defined parameters).

The following genomes were used to compare genomic CG contents: Honey bee (*A. melifera*, Amel_4.5), worm (*C. elegans*, WBcel235), Plant (*A. thaliana*, TAIR10), Dario/zebrafish (*D. rerio*, GRCz10), Mbuna (*M. zebra*, UMD1), Coelac (*L. chalumnae*, LatCha.1), Gallus/chicken (*G. gallus*, Gall_5), Grey whale (*E. robustus*, v1), human (*H. sapiens*, GRCh38.p10), mouse (*M. musculus*, GRCm38.p5), tammar wallaby (*N. eugenii*, Meug1.1).

2.11 Sequence divergence

Pairwise sequence divergence matrix was generated by Hannes Svardal, using whole-genome sequencing data published in Ref. [102]. $\geq$16 individuals (male) per species. Tree based on SNP difference was build using FigTree (v1.4.4; github.com/rambaut/figtree).

2.12 Transcriptome analysis – RNAseq

As part of a collaborative project, RNA isolation was performed by Dr. M. Du and NGS library preparation and sequencing were done at the Sanger Institute by the sequencing facility.

In brief, for each species, three biological replicates of liver and muscle tissues were used to sequence total RNA (see Fig. 2.33). Refer to Table 2.4 for a detailed description of sample size for both WGBS and RNAseq. Of note, the same specimens were used for both RNAseq and WGBS.

In brief, total RNA from RNA*later* preserved and homogenised liver and muscle tissues were isolated using the phenol/ chloroform approach. RNA samples were then treated with DNAase. ribosomal RNA (rRNA) fragments were also removed using RiboZero, according to manufacturer's instructions. Quality and quantity of rRNA-deprived RNA extracts were determined using NanoDrop and Bioanalyser. cDNA NGS libraries were directional and sequenced on HiSeq 2500/4000, paired-end 75bp and were performed by the sequencing facility of the Sanger Institute.

2.12.1 Alignment of RNAseq reads and gene counting

Quality of sequenced read pairs was determined, adaptor sequences and low quality reads removed with `TrimGalore --paired --fastqc --illumina`.

Reads were then aligned to the *M. zebra* genome assembly (UMD2a) and expression of each transcript was quantified (TPM) using `kallisto quant --bias -b 100 -t 1` (v0.43.1) [179]. Note that for all downstream analyses, gene expression, rather than transcript, was used (mean TPM of all transcripts/isoforms per gene).

2.12.2 Differential gene expression analysis

Differential gene expression (DE) analysis was performed on gene expression matrix using `sleuth` (v0.30.0 [180]; `sleuth_wt`, Wald test) with FDR <0.01. Gene expression of TPM>5 in any individuals of any of the 5 species was required for DEG.

2.12.3 Visualisation of DEG and transcriptomic data

Principal component analysis (centred and scaled) and Spearman correlation matrix were produced with build-in `R` programme `prcomp` and `cor` on gene expression matrix (averaged TPM values for biological replicates for each species). Graphs and unsupervised clustering and heatmaps were produced with `R` packages `ggplot2` (v3.1.1) and `pheatmap` (v1.0.12). Heatmaps of gene expression show scaled TPM values (TPM $\geq$ 5 in at least one replicate per comparison).

To study the correlation between methylome and gene expression, genes were binned into 11 categories based on their expression levels (increasing gene expression levels, from cat 1 to 10; cat "OFF" groups silent/not expressed genes, i.e. TPM=0 in all replicates for a particular species. RL liver (n=2): 10 'ON' categories , n=2,129; 1 'silent' category, n=5,331. MZ liver (*n*=3): 10 'ON' categories , *n*=2,199; 1 'silent' category, *n*=4,704. RL muscle (*n*=2): 10 'ON' categories , *n*=2,101; 1 'silent' category, *n*=4,622). TSS (500bp$\pm$ TSS) and gene bodies were also binned into 10 categories according to methylation states (0-100% methylation, by 10% incremental; RL liver (*n*=2), 11 categories, *n* ranging from 34 to 11,202. MZ liver (*n*=3), 11 categories, *n* ranging from 28 to 11,192. RL muscle (*n*=2), 11 categories, *n* ranging from 60 to 9,946). Categories were generated using the R package, `tidyverse`. TPM values and methylation levels were averaged for each tissue and per species using all biological replicates.

Chapter 3

The methylome of *A. calliptera* sp. Massoko - early stages of speciation

3.1 Background

A recent study by Malinsky *et al.* investigated the genomic basis underlying the ongoing speciation of two distinct ecomorphs of *A. calliptera* in Lake Massoko [181]. Surprisingly, the authors found no fixed SNPs between the two ecomorphs, instead they described large genomic regions showing a high degree of genetic divergence, the so-called islands of speciation (see section 3.1.2), suggesting that the speciation process is highly polygenic. In addition to this remarkably low genetic variability, restricted to some genomic regions on a few chromosomes only, striking phenotypic differences were highlighted, such as different morphological traits, distinct male breeding colours, partial assortative mating and adaptation to different sources of food.

3.1.1 Geography of Lake Massoko

Lake Massoko (or Lake Kisiba) is a volcanic crater (maar-type) lake in the Rungwe highlands in the Southern end of Tanzania, approximatively 20 km north from the Northern tip of Lake Malawi [182, 183] (Fig. 3.1). Lake Massoko has a diameter of 700m and a maximum depth of 36m [183]. The lake bottom can be divided into three zones: the littoral (5-10m deep) is composed of some reeds and grass, then the crater wall zone, which is steep and rocky, and finally the mud plain, composed of a flat and thick layer of silt. Recent studies could reliably date the formation of this crater lake at 50 thousand years ago [182].

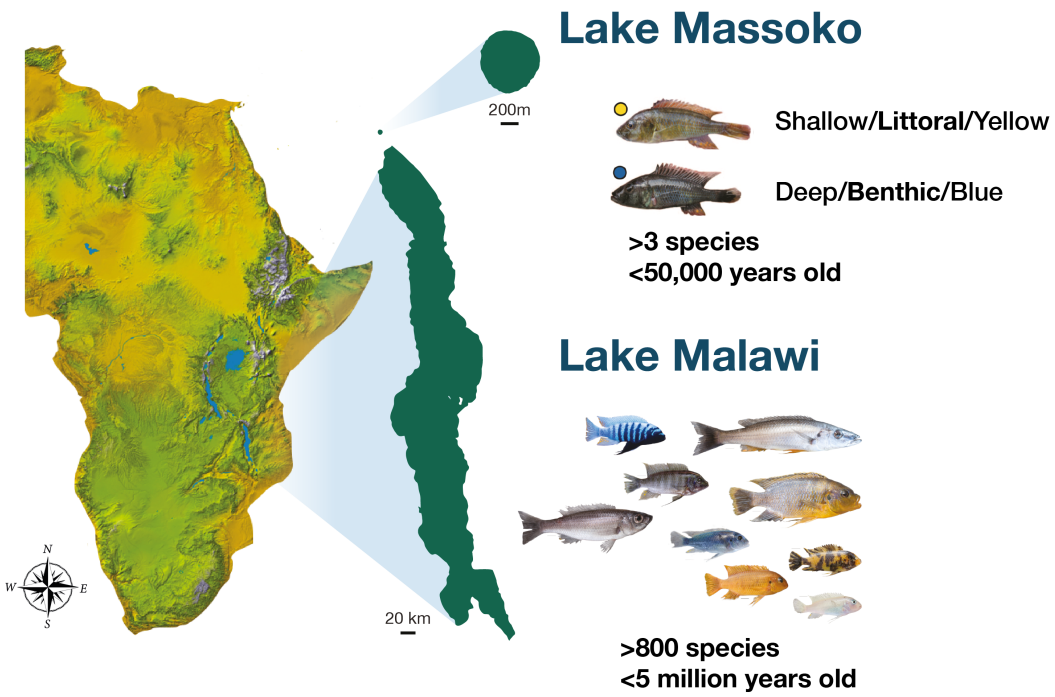


Fig. 3.1 Map and cichlids of Lakes Massoko and Malawi. Cichlids of Lakes Malawi and Massoko are studied in this thesis to understand variation and convergence in genomic loci of DNA methylation variation in liver tissues. More than 800 different cichlid species have been described in Lake Malawi (estimated formation: 5 million years ago), and only 3 cichlid species have been found in the crater lake Massoko (crater lake formation: 50 thousand years ago). Two subgroups of speciating *A. calliptera* individuals are found in lake Massoko: the benthic, blue ecomorph populating deeper parts (20-25m) of the lake, while another ecomorph, yellow breeding colours for males, populate shallower habitats (0-10m), closer to the shore.

3.1.2 Fishes of Lake Massoko

Lake Massoko is home to four fish species (three of which are cichlids) [183]. The most numerous species is the polymorphic *Astatotilapia*, followed by a population of *Oreochromis*. Two other species, although much less frequent, have been observed, namely *Coptodon rendalli* and the catfish *Clarias gariepinus*, both of which are endemic to Lake Malawi and its catchment and might have been introduced in Lake Massoko [183].

Interestingly, different ecomorphs of *A. calliptera* are found throughout the lake (Fig. 3.1). The first population dwells in the shallower parts of the lake (0-10m) and is characterised by males showing bright yellow breeding colours, resembling the riverine and probably ancestral *A. calliptera* specimens. It is therefore referred to as the littoral or yellow ecomorph. The second population is found deeper in the lake (15-20m) and is called the benthic or blue ecomorph, after the male breeding colour. No observation of yellow morphs has ever been

made in the benthic parts of the lake, however the benthic specimens have been reported to inter-breed with the littoral population, in particular in the intermediate zone (between littoral and benthic regions) [183]. In addition, there is evidence for moderate but significant assortative mating (in particular for littoral females) between the two ecomorph groups [181], which is a key element in the speciation process [184, 143].

Morphological differences of *Astatotilapia* sp. Massoko

The haplochromine cichlids of Lake Massoko show striking morphological differences. In particular, the benthic specimens have on average (i) longer head and jaw, (ii) lower body mass, (iii) narrower pharyngeal teeth and (iv) lighter lower pharyngeal jaws (see Fig. 1.8c). In addition, differences in stable isotope ratios indicated (v) a greater influence of planktonic-based food in the diet of the benthic population. Altogether, this suggests clear ecomorph separation and adaptation to different ecological environments in *Astatotilapia calliptera* sp. Massoko.

Genomic islands of speciation in *Astatotilapia* sp. Massoko

The two *A. calliptera* ecomorph populations of Lake Massoko have been reported to be very closely related genetically. Strikingly, whole-genome sequencing of the subpopulation highlighted no fixed SNPs clearly segregating the different ecomorph groups (more individuals are being sequenced and analysed at the moment). Rather, Malinsky, Challis and colleagues identified 55 genomic regions showing high genetic polymorphism, called HDRs for highly diverged regions (both high F_{ST} and D_{xy}[1], in brief high allele frequency differences and high nucleotide diversity between populations) without fixed SNPs between the two populations. Half of these HDRs are localised on only five chromosomes [181], thus forming *archipelagoes* of speciation. Interestingly, these regions of high divergence were enriched in genes associated with morphogenesis, sensory systems and steroid hormone signalling, in line with morphological divergences and visual adaptations to dimly lit parts of the lakes [181].

Moreover, whole-genome sequencing analyses confirmed the early stages of divergence of Lake Massoko *A. calliptera* ecomorphs. These populations could have originally populated the lake ca.10,000 years ago after the separation from the Mbaka *A. calliptera* population

[1] F_{ST} represents a measure of allele frequency differences between populations, that is the population-specific change in allele frequency, while D_{xy} represents an absolute measure of population difference, capturing the number of nucleotide difference among populations [185]

(the river Mbaka is located nearby Lake Massoko). Then, the benthic groups is estimated to have formed a distinct groups within the past 500-1,000 years [181].

Overall, the Lake Massoko system offers a unique opportunity to investigate the methylome variation associated with this on-going sympatric speciation in natural populations of *A. calliptera* of Lake Massoko, and whether any convergence in genomic loci showing high DNAme variation exists between Lakes Malawi and Massoko cichlid systems (Fig. 3.1).

To address these aims, I generated genome-wide bisulfite sequencing data from multiple specimens from benthic, littoral and proximate riverine populations. In line with the adaptation to different diets, I hypothesise that the liver methylome would affect liver function and thus be related to diet.

3.2 Whole Liver RRBS Methylomes of 35 *A. calliptera* specimens of Lake Massoko

To investigate the variability of liver methylomes in Lake Massoko *A. calliptera* ecomorphs and to draw comparisons with the Lake Malawi cichlid radiation, $\geq$ 11 males of each ecomorph (Riverine, Littoral, Benthic) were collected. Two bisulfite sequencing (BSeq) approaches were employed:

- **RRBS**, or reduced representation bisulfite sequencing;
- **WGBS**, or whole-genome bisulfite sequencing.

First, the targeted approach, RRBS, was used [186]. This method uses the restriction enzyme, MspI, that cleaves the phosphodiester bond of double-stranded DNA at 5′C.CGG 3′. This results in an enrichment for sequencing of GC-rich genomic regions (where most changes in DNA methylation is expected, at least in mammals). While allowing many more samples to be sequenced at a reasonable cost, it differs from the WGBS approach in that only ca. 25% of all CpG sites genome-wide will be sequenced. Subsequently, WGBS data for two breeding male individuals for each of the three groups was performed to obtain a genome-wide resolution of the Massoko methylome. Refer to table 3.1 for a detailed description of sample size for each BSeq technique.

Table 3.1 Sampling size - Methylomes of *A. calliptera* sp. Massoko

	BSeq type[1]	
Ecomorph	RRBS	WGBS
A. calliptera Mbaka (Riverine)	11	2
A. calliptera Littoral	12	2
A. calliptera Benthic	12	2

Number of biological replicates of liver tissue of breeding males, collected in the wild
[1] Bseq stands for bisulfite sequencing technique

3.2.1 Methylome of *A. calliptera* males of Lake Massoko

In order to characterise the variation in liver methylome between the three ecomorphs, RRBS single-end 50bp-long reads were all mapped to the same reference genome of Lake Malawi cichlid (*M. zebra*; UMD2a). On average, 11.07 $\pm$ 3.4 million SE50 reads ($\mu \pm \sigma$) were generated (Fig. 3.2a). Alignment scores were high (i.e., compared with WGBS data), with on average 83.84 $\pm$ 1.6% of reads mapping uniquely (Fig. 3.2b). Unmapped reads might arise from overly-short read sequences and/or too many alignments (repetitive nature of the read sequence). In addition, mapping efficiency is decreased because bisulfite converted read mapping relies on three letters instead of four as the nucleotide C is used to infer methylation status.

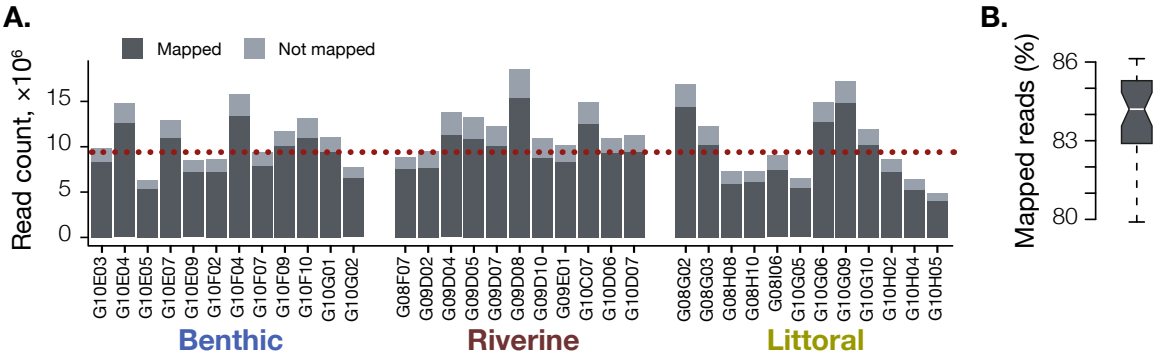


Fig. 3.2 Mapping of RRBS reads - Methylome of liver tissues for the two different *A. calliptera* ecomorph groups of Lake Massoko and riverine *A. calliptera*. Single-end 50bp-long reads of liver methylome were mapped to *M. zebra* reference genome. **A.** Count of uniquely mapped and unmapped reads for all *A. calliptera* specimens of each ecomorph group. The red dotted line represents the mean number of mapped reads for all RRBS sequencing data. **B.** Boxplot showing the percentage of mapped reads for all RRBS datasets combined. Fish IDs are provided below graphs.

The RRBS approach enriches for CpG-dense regions of the genome. The total count of CpG sites in the Massoko RRBS dataset (i.e. count of all CpG sites that have been sequenced in at least one of the 35 samples, in any population and with a minimum coverage of 4 sequencing reads) is 3.5 million (Fig. 3.3a, "All CpG"). Only 35% (1.2 million) of the CpG sequenced is common among all three populations (i.e. CpG common in at least one individual in each population, Fig. 3.3a, "Ecomorph"). Of note, only 225k CpG sites (7.1%) are common to all the 35 individuals (Fig. 3.3a, "Individual") - these CpGs, shared among all samples, are required to evaluate methylome patterns in the whole population (see below). This drop in the number of CpG common to all individuals could be due to technical drawbacks (low methylation coverage, variable sequencing depth at different loci, fragments absent in NGS library) or due to actual SNPs between individuals.

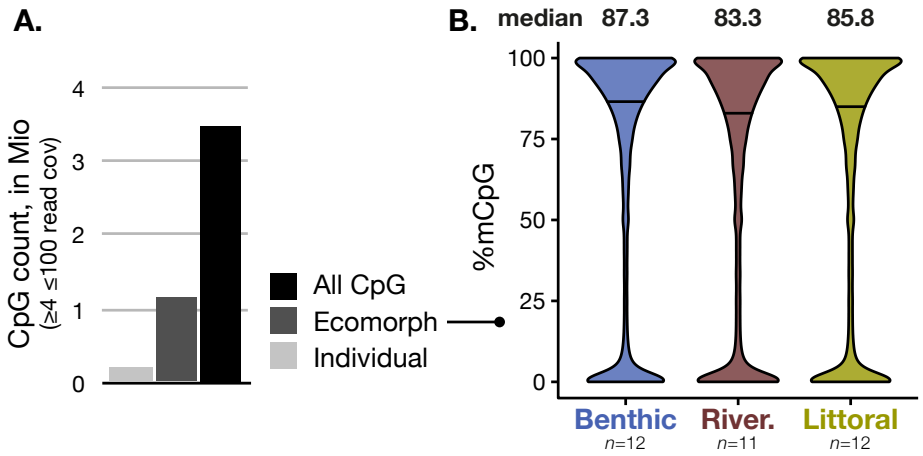


Fig. 3.3 Total CpG count and methylation levels in RRBS datasets. A. Histogram showing the total number of CpG sequenced in the 35 individuals of Massoko dataset. "All CpG": all the CpG ever sequenced in any individual of any population (black); "Ecomorph": CpG common in at least one individual in each population (darker grey, common between general groups); "Individual": CpG sequenced in all the 35 individuals (light grey, common in all). Sequencing coverage at any CpG sites of at least 4 reads. **B.** Violin plots representing methylation levels (%mCG/CG) in each population, at CpGs shared between ecomorph groups (darker grey in **a.** – 1.15 million CpG sites). Median values are given above violin plots.

Overall, methylation levels at CpGs in the three populations are consistently high, showing a clear bimodal distribution: most CpG sites show a high degree of methylation (around 100%) with a smaller fraction of CpG showing almost no methylation at all. On average, the deep population shows a slightly higher methylation level overall compared to the Riverine group exhibiting the lowest level of the three populations (median values of 87.3%, 85.8% and 83.3% overall DNA methylation levels, for Benthic, Littoral and Riverine populations

respectively; Fig. 3.3b), while the littoral population shows a more intermediate DNAme level overall.

3.2.2 Methylome patterns unique to each *A. calliptera* sp. Massoko population

Overall differences

In order to investigate whether different Massoko morphs of *A. calliptera* would show distinct patterns of liver methylation, and whether this could be correlated with different trophic adaptations, Spearman correlation scores were generated and Euclidean distances were plotted. Principal component analysis of methylation variation at conserved CpG sites was also performed.

Methylome variation between populations is almost always greater than variation within-population: specimens mostly cluster by populations based on Spearman correlation scores of liver methylome (Fig. 3.4a). The benthic population forms a more distinct group, with liver methylome patterns being more unique than these shared between the riverine and littoral populations. Of note, some specimens, identified as showing the benthic blue phenotype, cluster with shallow individuals in terms of liver methylome patterns (Fig. 3.4a). Specimens caught in the wild were sampled and identified as belonging to the benthic or littoral population according to the blue vs yellow male breeding colours and the depth of the catchment. George Turner has reported that no littoral specimens have ever been observed in the benthic areas of the lake [183]. Moreover, upon collection, Malinksy, Challis and colleagues could not rule out the fact that some littoral specimens could have been by-caught inadvertently while sampling benthic individuals [181]. Interestingly, all the specimens exhibiting methylome patterns specific to the benthic population also show diet specific to the deeper parts of the lake (Fig. 3.5). This suggests that some benthic could have been found in shallower parts of the lake (i.e. the intermediate zone), however no littoral specimens were found in the benthic habitats. This is in line with the idea that benthic populations could sometimes interbreed with littoral fish [181, 183].

Similarly, the first principal component (6.6% of variance) might mostly be explained by the unique patterns of liver methylome in the benthic population compared to both the littoral and riverine groups (clustering together). The second principal component, explaining 4.2% of the methylome variance, is required to distinguish the riverine from the littoral populations (Fig. 3.4b). This might be reflected by the slightly higher methylation levels observed overall in the benthic specimens, compared to the two other ecomorph populations (Fig. 3.3b). Of

note, Malinsky and colleagues [181] have shown that the benthic and littoral individuals exhibit a significant change in stable isotope ratios (more depleted in C^{13} in benthic fish), which indicates significant dietary changes to more offshore-planktonic food sources for the benthic populations. Stable isotope ratios similarly can group the two populations apart (Fig. 3.5).

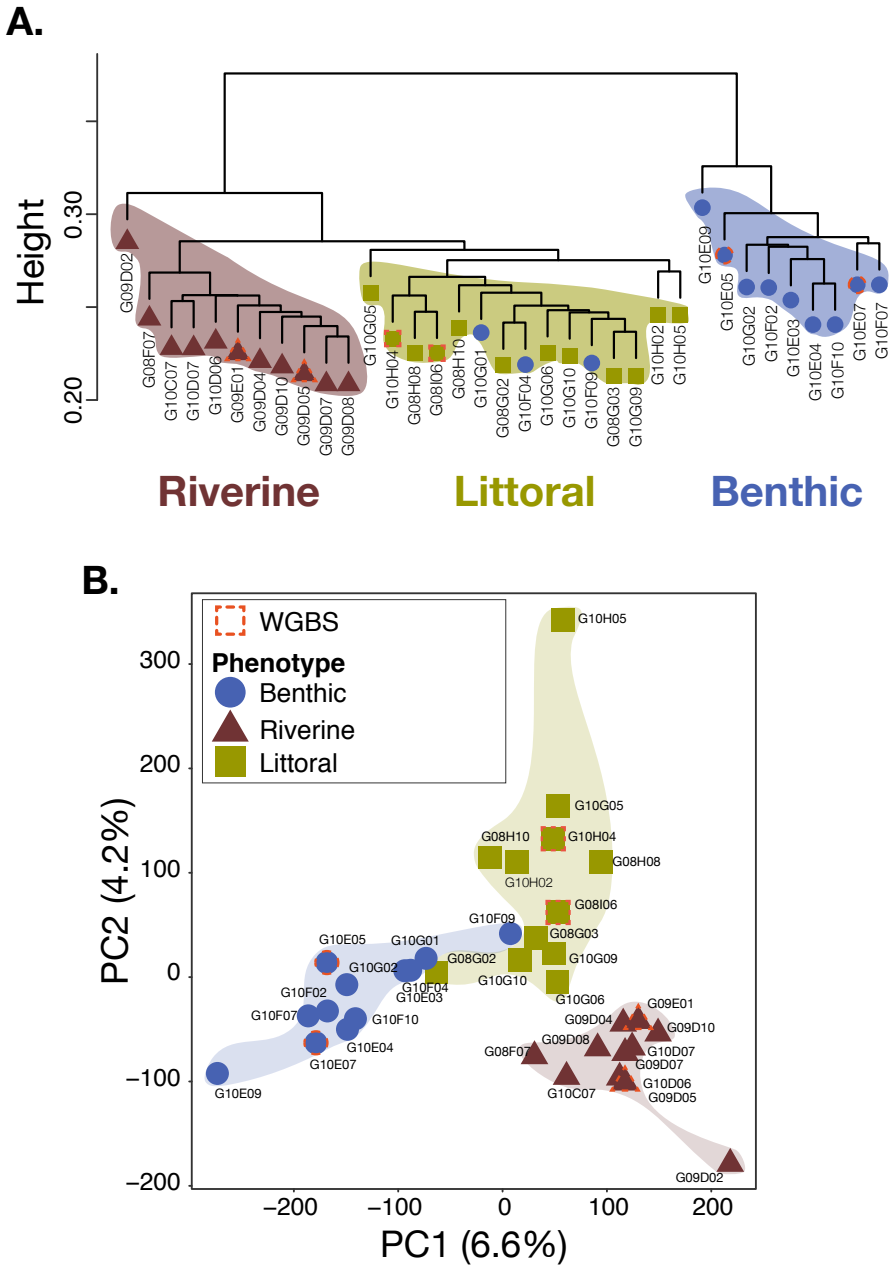


Fig. 3.4 Variation and patterns of liver methylomes in the two *A. calliptera* sp. Massoko ecomorph and the riverine *A. calliptera* populations. RRBS data of 35 *A. calliptera* of Lake Massoko. **A.** Unbiased hierarchical clustering of all the 35 *A. calliptera* male specimens based on liver methylome similarities (Spearman correlation scores). Red dotted outline indicates specimens for which WGBS has been generated as well. **B.** Principal component analysis (PC1 vs PC2) of liver methylome at conserved underlying CpG sequence. Percentage of explained variation for each PC in brackets. Specimens were classified into the benthic or littoral populations based on catchment depth and male breeding colours.

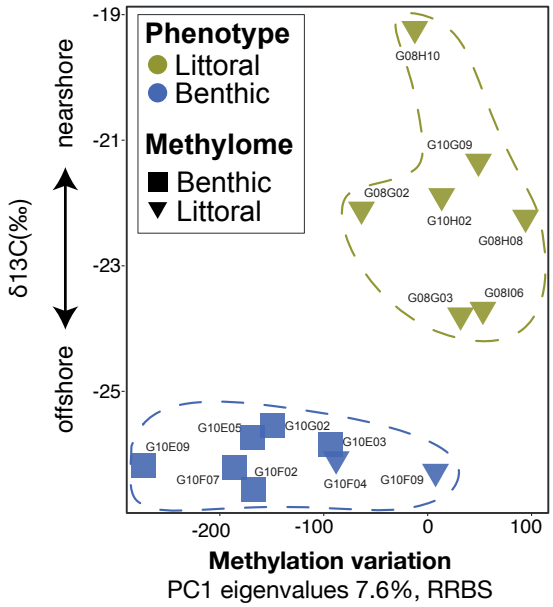


Fig. 3.5 Diet and methylome variation in Lake Massoko cichlids. Liver methylome variation (PC1) was plotted against stable isotope ratios measured in the livers of some individuals only (δ^{13}C, per mille). Isotope measurements carried out by Martin Genner and colleagues. Specimens were identified as belonging to either the benthic (blue) or littoral (yellow) populations based on male breeding colours and depth of capture. Methylome patterns specific to benthic (square) or littoral (triangle) populations derived from Fig.3.4b. Isotope measurement data only available for some individuals.

3.2.3 DMRs between the two *A. calliptera* populations of Lake Massoko and riverine *A. calliptera*

I then sought to identified differentially methylated regions (DMRs) showing consistent and significant change in single CpG methylation levels.

Most of the differentially methylated regions are hypermethylated in the benthic populations when compared to the riverine or littoral specimens (75% and 70% total DMRs, resp.; Fig. 3.6). Interestingly, the variation in methylome between the riverine and littoral is more balanced, with as many hypomethylated DMRs as hypermethylated ones. Moreover, the highest number of DMRs is observed between riverine-benthic populations, with 534 DMRs – this is 60% more DMRs than between either riverine-littoral and littoral-benthic. This highlights that the benthic population exhibits the most distinctive and unique liver methylome pattern. The littoral population shows a more intermediate methylome, in-between the ancestral-like riverine and the deep benthic populations, though sharing slightly more similarities with the ancestral-like methylome.

Strikingly, most of the methylation variation in Lake Massoko *A. calliptera* populations seems to be associated with increased methylation levels in the benthic group in particular, and also, to a lesser extent, in the shallow population, when compared with the ancestral-like riverine group.

The gain in DNAme in the benthic population might explain what both the PC and Spearman correlation analyses have highlighted: the benthic population formed a distinct group, clustering away from the shallow and riverine groups (Fig. 3.4a,b). The benthic population also shows a slightly increased methylation level genome-wide compared to shallow and riverine groups (Fig. 3.3b).

RRBS-DMRs were on average 170-220bp long (Fig. 3.7a). Of 1,209 DMRs found between the three Massoko populations, 1,068 (88.3%) are unique and 141 are common to some pairwise comparisons. Overall, 55% of all the DMRs show hypermethylated levels in benthic and/or littoral samples (with 37% hypermethylated in benthic exclusively) compared the riverine specimens. By contrast, 8% of all the DMRs are hypermethylated exclusively in the riverine specimens (Figs. 3.6b and 3.7b).

Current work aims at further characterising DMRs between the littoral and benthic groups, to investigate the epigenetic basis of this recent speciation.

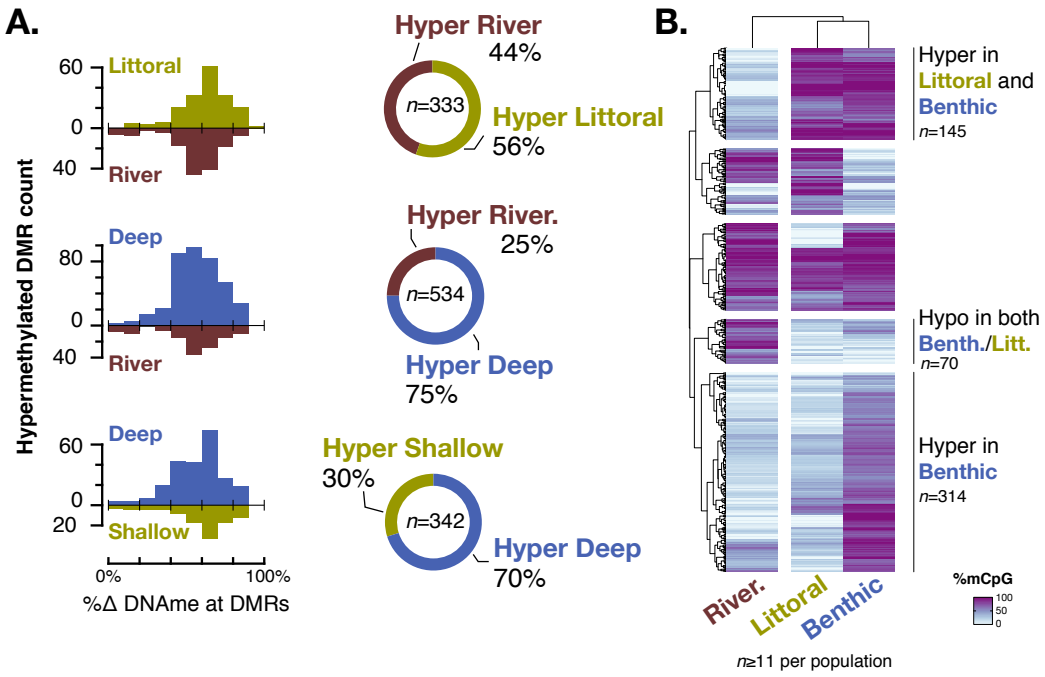


Fig. 3.6 Global increased in DNA methylation levels is associated with the littoral and benthic populations compared to the ancestral, riverine *A. calliptera* group. A. Histograms of the number of DMRs found between the liver methylome of the three ecomorphs with DNAme difference (hypermethylated DMRs plotted). The y-axis represents the number of hypermethylated DMRs in each group. The x-axis represents percentage change of methylation state at the DMR. Pie charts on the right summarise the number of hyper- vs. hypo-methylated DMRs in each comparison with total counts of DMR in the centre. **B.** Heatmap of DNAme with unsupervised clustering in each ecomorph population at each unique DMR found between at least one of the three comparisons (average mCpG in $\geq$1 individual in each population). Different clusters are found based on methylome patterns. In total, 770 DMRs are found in $\geq$ 1 specimen in each population. Sample size (n) for each group given at the bottom.

3.2.4 Genomic localisation of DMRs

The genomic localisation of the DMRs found was then determined. More than half (58%) of the hypermethylated DMRs in benthic/littoral lie in gene bodies (exons or/and introns) and only a small fraction (10%) in promoter regions[2] (defined as 500bp upstream and downstream of TSS, TSS $\pm$ 500). Slightly more than a fifth of these DMRs lie in transposon and repeat sequences, and almost a third are intergenic DMRs, which potentially include enhancer regions (Fig 3.7c).

[2]The RRBS technique is known to enrich for CG-rich fragments (CGI). In mammals at least, many promoter regions contain CGI and occupies a large fraction of RRBS reads [186, 44]. Malawi cichlids (*M. zebra* UMD2a) differ from mammals, in that only 32% of all the TSS $\pm$ 500bp contain CGI, and only 15% of all predicted CGIs lie in TSS (see method section 2.10.7 and Ref. [71]). Therefore, many CGIs in cichlids might be located outside TSS regions, in other biologically relevant loci ('orphan' CGI).

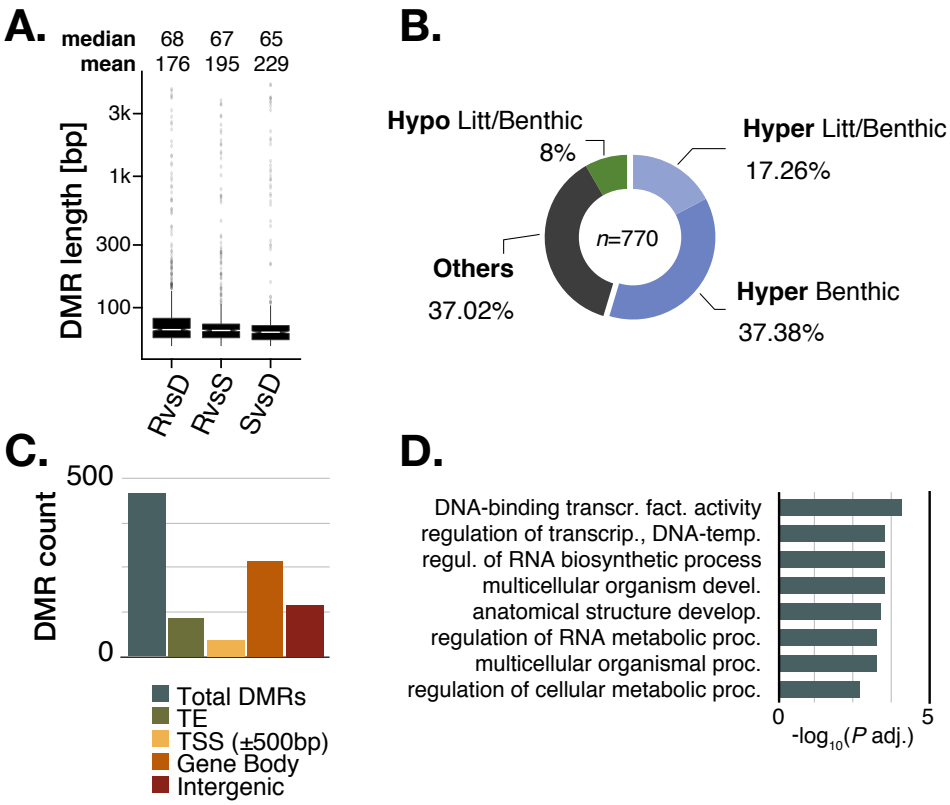


Fig. 3.7 Gain of DNA methylation is enriched in genes coding for DNA-binding proteins. A. Boxplots representing the length in bp of the DMRs between each comparison. Median and mean values are given above each boxplot. **B.** Pie chart summarising the different DNAme clusters shown in Fig. 3.6c. **C.** Genomic localisation of hypermethylated DMR (blue sections of pie chart in **b**). **D.** Significant GO terms for genes associated with DMRs located in either TSS, TE or gene bodies.

I then performed gene ontology enrichment analysis for the genes that could be associated with DNA methylation variation (DMRs).

Promisingly, many genes that show significant hypermethylation in benthic/littoral encode for DNA-binding factors. Many of these are transcription factors (TF) that are associated with anatomical structure development, embryo and tissue development, with major gene regulation functions (Fig. 3.7d). For example, many homeobox proteins (*hox-b7a, -b5a, -a5, -a7, -a5, -a3, -b3a*), or *fox1p*, as well as other TFs are hypermethylated in either the benthic or littoral specimens. Of note, some other genes, such as the visual system homeobox, *vsx1* (Fig. 3.8), or hypoxia-inducible factor 1-alpha, *hif1al*, might be relevant to adaptation to dimly lit and low oxygen environments. Of note, RRBS technique enriches for CG-dense portions of the genome, the so-called CpG island (see section 2.5.3). Mammalian promoters are enriched

for CGIs, in striking contrast with anamniotic vertebrates, where only less than a third of the promoters would be composed of CGIs. The RRBS technique might consequently enriched for CGIs in promoters even in fish, although most CGI are expected to be intergenic (orphan CGIs). They might therefore be a slightly biased enrichment for CGI in TSS in the analysis altogether.

The identified key developmental genes, showing differential methylation, appear to be hypermethylated and were found in liver tissues. This somewhat counter-intuitive observation might reflect embryonic memories of DNA methylation patterns (embryonic relics) [169]. Such patterns of methylation might be relevant and exhibit functions restricted to early development processes with little or no roles in a fully differentiated liver. Yet, their methylation patterns could have been passed on throughout differentiation and cell divisions, independently and without being erased. This provides a fruitful insight into methylation differences that may bear significant functions between the ecomorphs during embryogenesis.

Importantly, while there is a significant enrichment for developmental TFs being hypermethylated, other genes with functions related to diet and metabolism (in particular glucose and lipid metabolisms) are also hypermethylated. That might highlight a change in diet and in liver functions.

In addition, some epigenetic-related factors, such as *tet2* participating in DNA demethylation, the methionine synthase (*mtr*) involved in the S-Adenosyl Methionine cycle (SAM, the principal methyl donor for cytosine methylation) and the histone deacetylase 7 (*hdac7b*) are hypermethylated in the benthic populations. Their function in liver remains unclear, however some of them have being involved in liver regeneration [70]. Moreover, some cichlid species of Lake Malawi, in particular *P. genalutea*, show similar methylome variation at these genes as well (see 2.8.2).

Finally, hypomethylated DMRs in benthic/littoral populations account for less than 10% (n=70) of all DMRs in Lake Massoko cichlids (Fig. 3.6c). No significant enrichment for any GO terms were observed. Hypomethylated DMRs were localised in TSS (7.1%), gene body (58.6%), repeats (28.6%) and intergenic regions (34.3%); graph not shown.

In brief, the benthic specimens show increased methylation levels genome-wide compared to both the littoral and the ancestral-like, riverine specimens. Promoter regions of DNA-binding genes show in particular gain of methylation in the benthic population, which might affect gene expression. Increased methylation at promoters have been reported to be associated with lower transcriptional activity, in particular at early stages of development [78, 17]. Interestingly, many developmental genes show DNAme variation, hinting at a possible functional implication of methylation variation in embryonic development.

Although the RRBS technique allows for robust DNAme variation analysis in large sample size, between 20-25% only of the genome is covered. I then sought to confirm these results by generating whole-genome liver methylome landscapes at a single-base resolution for some selected specimens of each of the three populations.

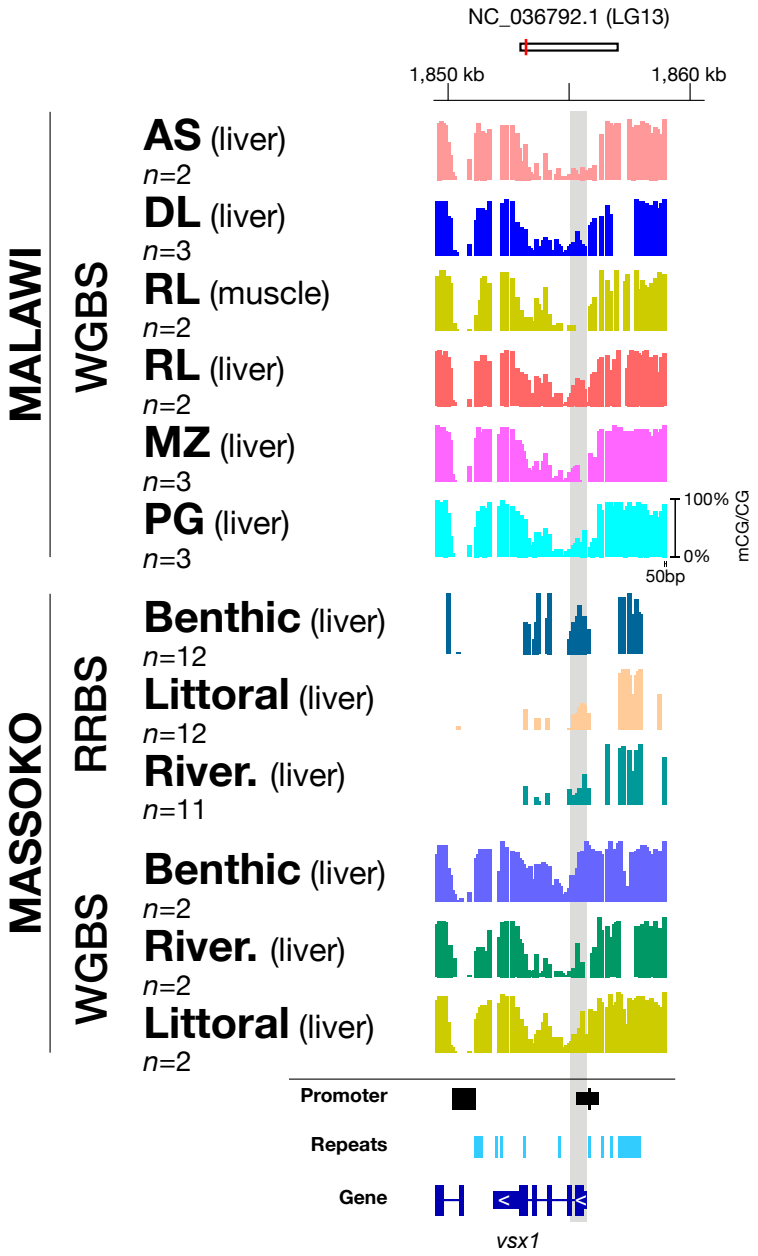


Fig. 3.8 The gene visual system homeobox 1, *vsx1*, shows benthic-specific hypermethylated state in promoter region. The promoter region of the gene visual system homeobox 1, *vsx1*, shows significant hypermethylation in the benthic population of *A. calliptera* in Lake Massoko. Here, liver methylome of different cichlid species of Lake Malawi is also given for comparison. This TSS-DMR overlaps with a CGI (CG-rich region), is predicted to be 206 bp-long and to be composed of 37 CpG sites. The methylation difference between the benthic and shallow/river populations is 47% across the region. Sample size for each morph/species is indicated (n), tissue in brackets. Average mCG in 50bp-long, non-sliding windows of all individuals of each species, genome-wide (percentage methylation).

3.3 Whole-genome liver methylomes at a single CpG resolution of the different populations of *A. calliptera* sp. Massoko

In order to obtain genome-wide liver methylome at a single-CpG resolution, two males from the RRBS dataset for each of the three *A. calliptera* populations were additionally sequenced genome-wide (WGBS); see Fig. 3.4 for sample selection (red outlines). This whole-genome approach has many advantages over RRBS, in that reads are three times longer (150-bp long) and are paired-end. In addition to the genome-wide benefit, this greatly improves mapping efficiency to the reference genome, in particular to repetitive elements (see Fig. 2.34). Sequencing costs is one of the major, if the only, drawback of WGBS – fewer samples were chosen to be sequenced for pecuniary reasons. While RRBS enables a large sample size, WGBS allows for a genome-wide sequencing at a single nucleotide resolution, with the trade-off of analysing only a few individuals.

Here, I use WGBS to get a genome-wide resolution of liver methylome in Lake Massoko cichlids and compare it to the results generated with the RRBS technique. The two datasets are then merged to draw a more accurate comparison of methylome dynamics between Lakes Massoko and Malawi.

3.3.1 WGBS - overall characterisation

Overall, 32.2 million CpGs were sequenced between the 6 individuals – of which 23.73 million CpG sites (74%) were present in all the specimens. This could be real biological sequence difference (SNP, indels for example) however I expect high genetic conservation between Massoko individuals, therefore technical differences, due to sequencing errors or lack of/low coverage, might be the primary reasons for this drop in count of shared CpG sites.

I then sought to quantify the variation in whole-genome liver methylome patterns between the benthic, littoral and riverine *A. calliptera* populations. To this aim, principal component analysis was performed using methylation divergence at conserved CpG sites.

At whole-genome resolution, the liver methylomes of the ancestral riverine population appears to be quite distinct to either the littoral and benthic groups, that show less methylome divergence (Fig. 3.9a), as previously observed in the RRBS dataset. However, PC analysis of WGBS variation reveals a less distinct liver methylome than the one observed in RRBS dataset between the benthic and littoral population. In addition, the WGBS approach also

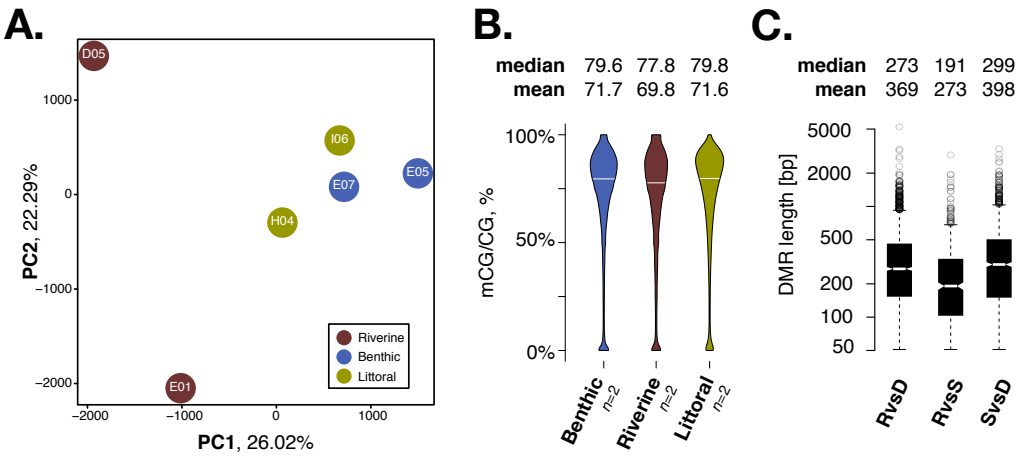


Fig. 3.9 Whole-genome liver methylomes of the two *A. calliptera* populations of Lake Massoko and of riverine *A. calliptera*.
A. PC analysis of liver methylome, genome-wide. PC1 and PC2 are shown, explaining 26% and 22.3% of the total variance, respectively. **B.** Violin plots of DNA methylation levels, genome wide. Average mCpG of the two biological replicates for each population (50bp-long non-sliding windows, sequencing coverage of 4-100 non-clonal reads). White lines represent respective median values. **C.** Box plots of the length of DMRs in the three comparisons (in bp). Mean and median values above graphs.

confirms that DNA methylation levels in the two riverine samples were slightly lower genome-wide, compared to both the littoral and the benthic populations, which share similar levels (Fig. 3.9b).

Between the three populations, there are 4,279 DMRs predicted, with 963 DMRs present in more than one pairwise comparison. In total, this means 3,296 unique DMRs, i.e. without duplicated DMRs (after duplicates being collapsed into unique DMRs). This is 2,228 DMRs more than with the RRBS approach (Table 3.2). The length of all DMRs was around 191-300 bp (Fig. 3.9c), larger than that in the RRBS dataset (probably due to read length. RRBS: single-end 50 bp-long reads, WGBS: paired-end 150 bp-long reads).

The highest number of DMRs is found between the riverine and benthic populations (n=2,164; 68% hypermethylated in benthic), followed by that of between the riverine and littoral populations (n=1,499; 69% hypermethylated in littoral). Finally, a smaller difference in liver methylomes was found between the benthic and the littoral specimens (n=616; with 55% hypermethylated in the littoral), reflecting PCA observations. Overall, 59.2% of all DMRs show hypermethylation in benthic or/and littoral (hyperDMRs), and 18.3% hypomethylation in the benthic specimens only (Fig. 3.10a, clusters indicated by * and $\diamond$,

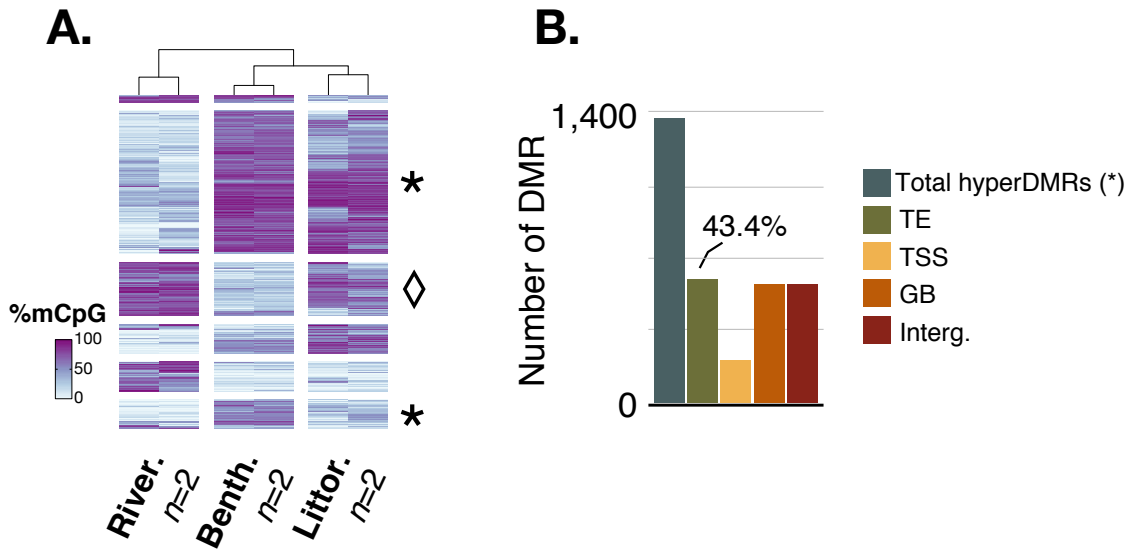


Fig. 3.10 Whole genome liver methylome of *A.calliptera* populations in Lake Massoko. A. Heatmap of CpG methylation levels at DMRs with CpG coverage of 4-100 seq reads present in all samples. *, hypermethylated DMRs in Benthic/Littoral population (59.2%); ◇, hypomethylated in the benthic population (18.3%). DMR count = 2,328. **B.** Genomic localisation of hypermethylated WGBS-DMR in benthic and/or littoral populations (* in c.). 51% of all WGBS-DMRs lie within TEs. Abbrev.: TE, transposable elements and repeats; TSS, 500 bp upstream and downstream of TSS; GB, gene body (exon and intron); Interg., intergenic if outside TSS and GB (according the aforementioned definitions).

respectively). Of note, there is a small fraction of DMRs that are only hypomethylated in the littoral ecomorphs.

Interestingly, 43.4% of the hypermethylated WGBS-DMRs in benthic/littoral lie within repeats (Fig. 3.10b). Compared to the RRBS hyperDMRs (23.5% of all hyperDMRs, see Fig. 3.6e), this increase in TE-DMRs could be linked to enhanced alignment of long paired-end reads produced with WGBS to repeat elements of the cichlid genome (see method, Fig. 2.34). Enrichment for any particular genomic regions will be assessed using DMR of both datasets combined (see below).

3.3.2 Overlap between RRBS- and WGBS-DMRs

Table 3.2 DMR count – WGBS vs. RRBS

	BSeq type	
Comparison	RRBS	WGBS
Riverine vs. Benthic	534	2,164
Riverine vs. Littoral	333	1,499
Littoral vs. Benthic	342	616
Total	**1,209**	**4,279**
Total unique [1]	**1,068**	**3,296**
Common	**168**	

[1] Unique, i.e. DMRs present between one or more pairwise comparisons (no collapsed duplicates)

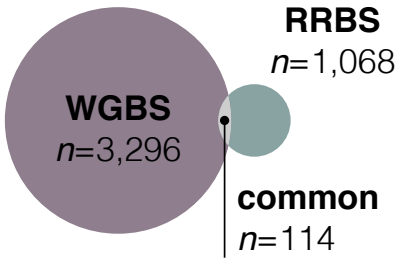


Fig. 3.11 Few DMRs are found using different bisulfite sequencing methods. Overlap between DMRs found using WGBS and RRBS (overlapping DMRs). Only 114 DMRs are found with both bisulfite sequencing techniques after collapsing DMR duplicates (several DMRs in one datasets for one larger DMR in the other dataset).

However, even though many more DMRs are observed in the WGBS dataset (probably due to the whole genome resolution, as well as a better alignment to repetitive elements), only a small fraction, 114 (10.7% or 3.5% of all DMRs in RRBS and WGBS, respectively), overlaps with the RRBS-DMR (Fig. 3.11 and Table 3.2). Furthermore, that statistical power (n=12 in each group) might be much higher using the RRBS approach, and could also partly explain the difference in DMRs found. The two methods are more complementary than similar, in that one enables a genome-wide resolution and a good coverage of repetitive regions while the other allows for a large number of replicates to be studied. Hence, DMRs from both datasets will be merged to capitalise on the strengths of both methods in the next sections, dealing with parallel dynamism of liver methylomes in the two lake systems.

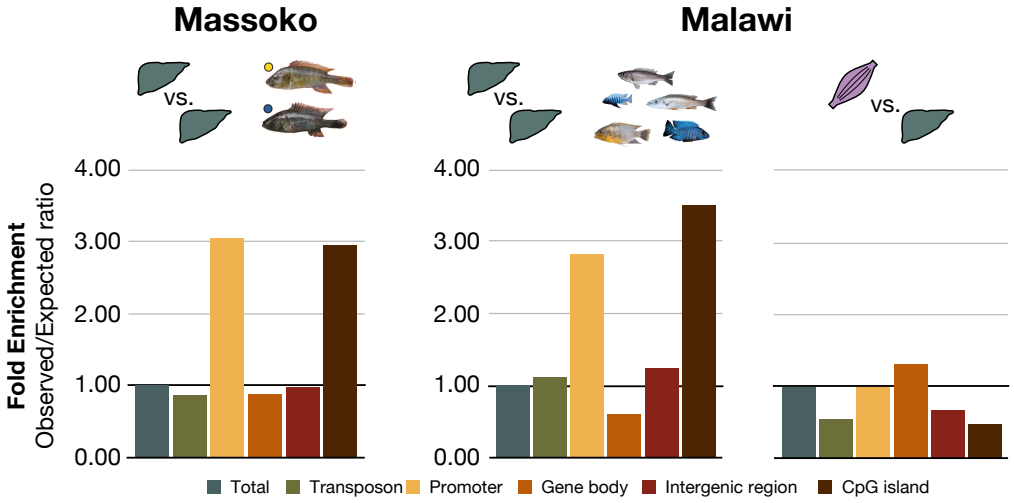


Fig. 3.12 DNAme variation is enriched at promoter regions and CGIs in livers of cichlids of both Lakes. Enrichment analysis for genomic localisation of DMRs found in liver of cichlids of Lakes Malawi and Massoko. Tissue-specific DMRs found in Lake Malawi cichlids are shown for comparison. See method for details about Observed Expected ratio formula.

Interestingly, although few DMRs in total are located in promoter regions, DNAme variation in liver between the three ecomorph populations shows strong enrichment for promoter regions (Fig. 3.12 and Table 3.3). In fact, promoter regions are 3 times more represented than what one could expect by chance. The same is also true for CpG islands, most of which are predicted to be located outside TSS regions (the 'orphan' CGI, see Ref. [44] and section 2.10.7). Other genomic regions such as transposable elements, gene bodies and intergenic regions are much less observed to exhibit DNAme variation – which could not rule out any biological functions of single DMR in these regions nevertheless.

Interestingly, the liver methylome in five species of Lake Malawi shows similar enrichment for variation at promoters and CGIs, suggesting that similar processes may be in action. The main differences between Lake Massoko and Malawi cichlids is that intergenic regions and transposable elements are enriched for DNAme variation in Lake Malawi only. This is in stark contrast with tissue-specific methylation variation that seems more located in gene bodies (Fig. 3.12).

As shown above, the example of the promoter region of the gene visual system homeobox 1 is striking: the TSS region of *vsx1* is significantly hypermethylated in the benthic population, consistently in both RRBS and WGBS datasets (Fig. 3.8). The transcription factor *vsx1* has been shown to play an important role in eye development in zebrafish and other vertebrates [187, 188]. In particular, during development, cells expressing *vsx1* enter a restricted

Table 3.3 Comparison of DMR genomic localisations between Lakes Massoko and Malawi cichlids

	Malawi		**Massoko**[†]	
Total DMRs[*]	10,988	*100.0%*	4,250	*100.0%*
TE/repeats	7,235	*65.8%*	1,900	*44.7%*
TSS	1,375	*12.5%*	504	*11.9%*
Gene Body	3,481	*31.7%*	1,857	*43.7%*
Intergenic	6,132	*55.8%*	1,889	*44.4%*

[*] Unique, collapsed DMRs, present between one or more pairwise comparisons (no DMR duplicates);
[†] Combined RRBS and WGBS DMR
Abbrev.: see caption of Fig.3.13.

differentiation to mainly give rise to retina bipolar cells, important population of cells in relaying signals from photoreceptor cells, such as cones and rods, to the ganglial cells. Differentially methylation of the gene *vsx1* might therefore participate in the visual adaptation to dimly lit parts of the lake. To test this hypothesis, possible further experimental work could quantify the methylation and transcription levels of the gene *vsx1* at an early developmental stage in the two ecomorphs (benthic and littoral) reared in different light levels.

Key epigenetic genes, such as *dnmt3b*, *tet2*, *tet3* are also differentially methylated and could be associated with cellular regenerative processes observed in liver [70].

3.4 Convergence in genomic localisation of shared DNAme variation in cichlids of Lakes Malawi and Massoko

Liver methylome of cichlids of both lakes show DNAme enriched in proximal *cis*-regulatory regions and in CpG islands. I then investigated whether the variation observed in liver methylome of the different cichlid species of Lake Massoko and Malawi shows convergence in genomic localisation – whether same genomic loci would be prone to show higher DNA methylation variation in the two different lake systems and whether they could be involved in the same biological processes in both lake systems. I hypothesise that some loci would show high DNAme variation in both lake systems, possibly facilitating phenotypic plasticity in response to environmental changes, which could promote phenotypic diversification.

To this end, I compared DMRs predicted in liver of the three populations of Lake Massoko *A. calliptera* with the five species of Lake Malawi, showing different eco-morphological and trophic adaptations. All individuals were sequenced using WGBS at a considerable sequencing depth, showed high DNA sequence conservation and mapped to the same reference genome, allowing for reliable comparison (Fig. 3.13a).

3.4.1 Shared DNAme variation is primarily located in TE sequences

In parallel in both lakes, 1,147 genomic loci show some level of convergence in their genomic localisation. This represents 10.4% or 27% of all DMRs found in Lakes Massoko and Malawi, respectively (Fig. 3.13b). Strikingly, 80% of these common DMR are associated with transposable elements (Fig. 3.13c), which might play a significant role in DNA methylation variability and might bear some regulatory functions [38]. In addition, intergenic regions, probably involving TEs as well, accounts for 40% of the total common variation. This might include intergenic promoters and enhancers. Unfortunately, no genomic annotation for cichlid intergenic regions (enhancers, in particular) exists at the moment, which warrants further genomic characterisation. And finally, a fifth of the common variation involves promoter regions (Fig. 3.13c).

Shared DNAme variation is located in genes involved in lipid metabolism and visual processes

I then performed gene ontology enrichment for genes that could be associated with common DMRs in order to characterise the regions that are consistently showing high DNAme

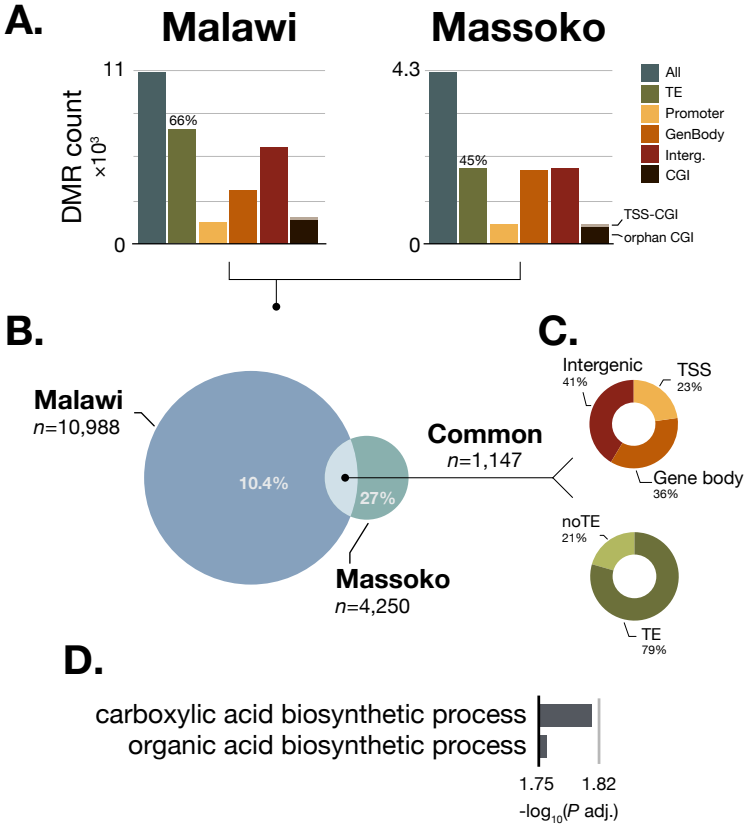


Fig. 3.13 Convergence in genomic localisations of liver methylome variation in Lakes Massoko and Malawi cichlid fishes. A. Summary of all DMR found in cichlids in Lakes Malawi (left; 5 species, triplicates, males) and Massoko (right: 3 populations of *A.calliptera*, triplicates, males). DMR counts are reported in details in Table 3.3. **B.** Overlap of DMRs between the two lake systems. *n*=1,147 common DMR. **C.** Genomic localisation of the common DMRs. **D.** GO terms enrichment analysis for the 843 genes associated with DMRs common to both systems. Abbrev.: TE, transposable elements and repeats; TSS, 500 bp upstream and downstream of TSS; GB, gene body (exon and intron); Interg., intergenic if outside TSS and GB (according to the aforementioned definitions).

variation between the two lake systems and that could possibly promote phenotypic variation in similar biological pathways.

Interestingly, genes targeted by similar DNAme variation are enriched in liver metabolism processes (Fig. 3.13d), in particular lipid and carboxyl metabolism (significant enrichment for genes related to carboxylic acid biosynthetic and organic acid biosynthetic processes [3] ; adj. *p*-value, p<0.02).

[3]e.g.: *shmt2*, serine hydroxymethyltransferase, mito; *glulc*, glutamine synthetase; *acacb*, acetyl-CoA carboxylase beta; *elovl6l*, elongation of very long chain fatty acids protein 6; ENSMZEG00005022331, cystathionine beta-synthase; *ass1*, argininosuccinate synthase 1; *adi1*, acireductone dioxygenase 1; *has1*, hyaluronan synthase 1; ENSMZEG00005022339, cystathionine beta-synthase.

Other genes, more related to development (general and eye/retina) are also enriched and might be particularly relevant at earlier developmental stages. A fish paralogues of the mammalian DNMT3b, *dnmt3bb.3* [4] shows DNAme variation in both lake systems.

Shared DNAme variation is associated with differential gene expression in Lake Malawi cichlids

In total, this common DNAme variation could be associated with 843 genes (either TSS-DMR, gene body-DMR or TE-DMR lying within 4kbp distance of one gene). In Lake Malawi cichlids, 96 of these genes show significant differential expression in liver and show enrichment for genes related to specific fatty acid metabolism (adj. p-value, p<0.02; *aacs*, *echsi*). Some of which show a strong patterns of expression carnivore vs herbivore (Fig. 3.14). Of note, *P. genalutea* has unique patterns of gene expression with many genes being upregulated only in this species (reflected in RNAseq graph showing unique DEG; see Chapter 2), and might be related to hepatic regeneration.

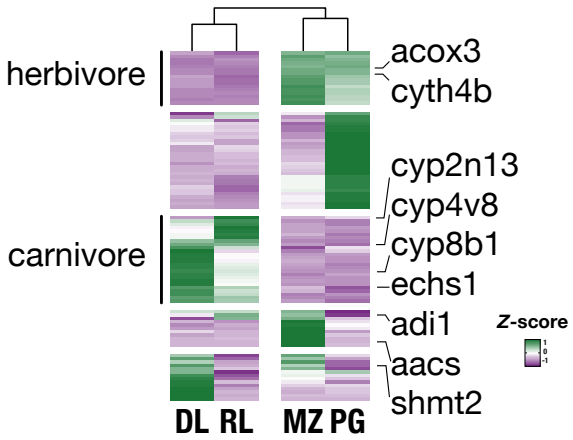


Fig. 3.14 Shared DNAme variation is associated with transcriptional changes. Heatmap of gene expression levels for DEG in liver of four Lake Malawi cichlids associated with common DMR from Fig. 3.13b (*Z*-score; averaged TPM expression per gene per species). DL, *D. limnothrissa* (*n*=3); RL, *R. longiceps* (*n*=2); MZ, *M.zebra* (*n*=3); PG, *P. genalutea* (*n*=3). Sample size (*n*) given in brackets. Cluster showing patterns of gene expression according to diets (carnivorous vs herbivorous diets) are indicated on the left.

A striking example of differentially expressed gene correlated with common DNAme variation at its promoter region is the cannabinoid receptor 1 (*cnr1*, see Fig. 3.15). This G-coupled protein receptor has been reported to be involved in energy balance and food

[4] ENSMZEG00005026841; predicted to be an isoform of *dnmt3b*; see section 1.1.

intake, particularly in hepatic lipogenesis. Furthermore, *cnr1* knock-out in mice has been linked to diet-induced obesity [189], suggesting a strong involvement in energy balance. Another fatty acid metabolism-related genes, such as the different subunits of the cytochrome P450 (Fig. 3.15, highlighted in pink), exhibit both significant differential methylation and gene expression changes, consistently in the cichlids of both lakes. This suggests that some genes are more prone to show epigenetic variability, accompanied with transcriptional changes, processes which might both participate in the phenotypic diversity observed in a context of adaptation to different diets and habitats and species diversification. Further work is required in order to characterise the direct role of DNAme in possibly regulating gene expression, in particular with regards to the set of genes showing shared DNAme variation. RNA sequencing of the livers of *A. calliptera* sp. Massoko is currently being carried out to address this aim.

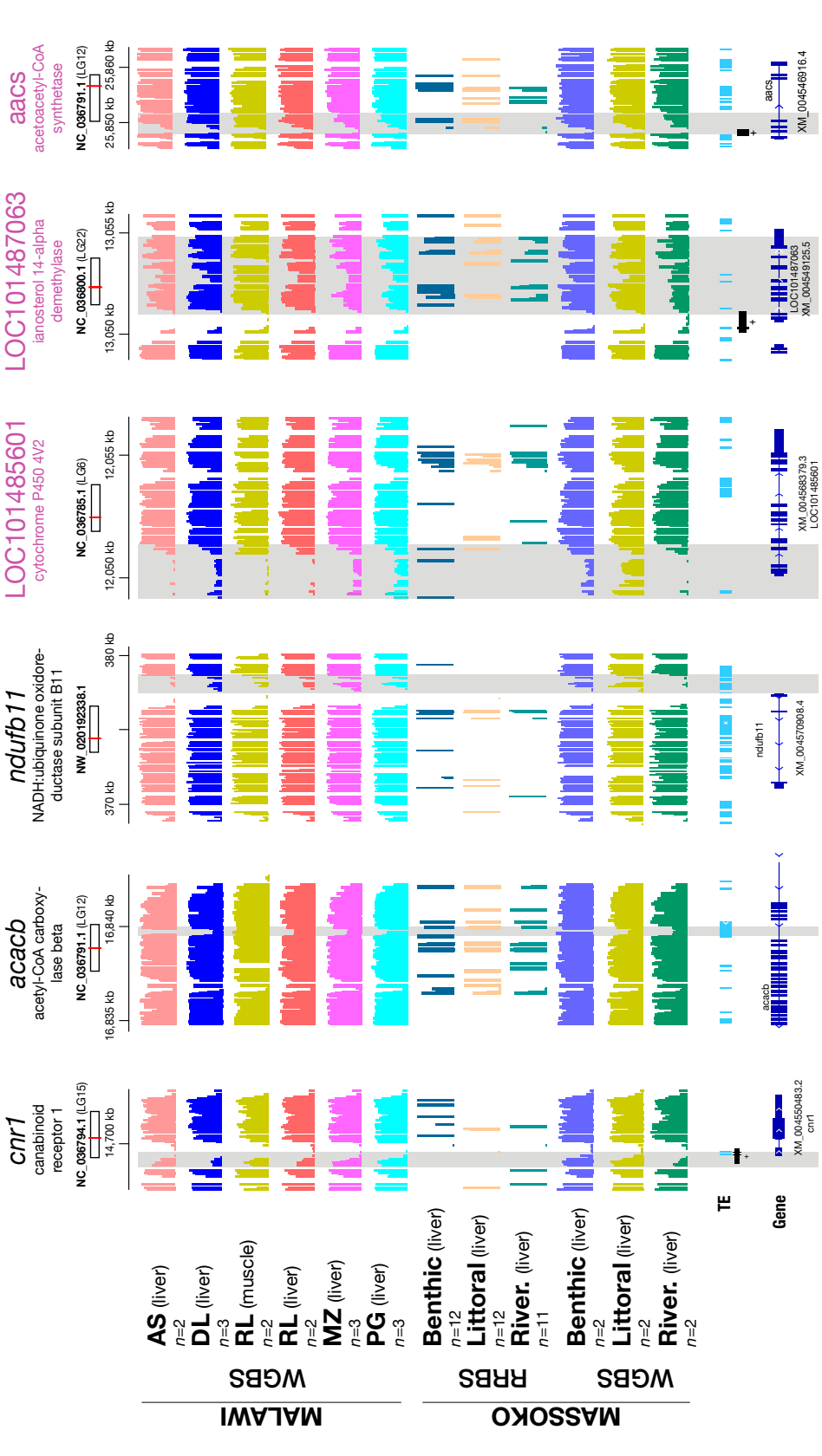


Fig. 3.15 Example of common DNAme in both Lake systems. Sample size (*n*) for each morph/species is indicated below the name of the species, tissue in brackets. Average DNAme over 50bp-long, non-sliding windows of all individuals of each species, genome-wide (percentage methylation). Lake Malawi species (WGBS): AS, *A. stuartgranti*, DL, *D. limnothrissa*; RL, *R. longiceps*; PG, *P. genalutea*. Lake Massoko *A. calliptera* population (RRBS and WGBS): Benthic (blue/deep) morphs; Littoral (yellow/shallow)morph; River, ancestral/riverine. All specimens are wild-caught, breeding males.

3.5 Discussion and future work

In chapter 3, I described DNA methylation variability in liver tissues of *A. calliptera* sp. Lake Massoko, in order to investigate the role of liver methylome in facilitating phenotypic plasticity. I sought to explore whether such DNAme variation could be shared between Lake Malawi and Lake Massoko cichlids, in that in both systems traits related to adaptation to different diets have been described and could play a significant role in early stages of speciation (Fig. 3.5 and Ref.[181, 143]). Same biological processes could exhibit epigenetic variation, hinting at the biological relevance of these pathways in short and long-term divergence. We hypothesise that DNAme might be important in shaping the transcriptome, and thus the phenotypes. The epigenetic landscape might underlie phenotypic plasticity of certain adaptive traits that might therefore be selectable. In this context, DNAme variability could be part of natural selection and might be relevant during the early phases of adaptive divergence, when genetic variation and genetic differentiation might impede phenotypic plasticity [11]. .

3.5.1 DNAme patterns unique to the benthic and littoral groups, compared to the ancestral patterns

In Chapter 3, I described the variability in liver methylomes of the two ecomorph populations of Lake Massoko and one riverine subspecies, probably closely related to the ancestral *A. calliptera* population of Lake Massoko, using two complementary methods. This allowed for the sequencing at both population and genome-wide levels.

Strikingly, the three populations exhibit significant methylome divergence at conserved underlying genetic sequences, sufficient to reconstruct the phylogeny based on liver methylome variation (Fig. 3.4). The benthic population shows the most distinct liver methylome patterns compared to the littoral and riverine population. Interestingly, most of the variation observed in the benthic population is associated with widespread increase in DNA methylation and is enriched at promoter regions of genes coding for DNA-binding factors (Fig. 3.7). Differential methylation levels at TF binding sites can be brought about both by differential occupancy of particular DNA-binding actors (i.e. lack of one particular TF might lead to increased methylation levels) or by upstream mechanisms leading to change in promoter methylation and therefore in TF occupancy [78]. This warrants further studies, directly testing the biological relevance of differential methylation state at key genes during development in both benthic and littoral populations.

These results suggest that the invasion and colonisation of the deeper part of the lake by an ancestral *A. calliptera* cichlid population might have been accompanied by a gain in CG methylation at key developmental and metabolic genes in the benthic population of Lake Massoko in particular. This might have facilitated rapid adaptation to this novel environment via DNAme-mediated differential liver metabolic functions and embryonic development.

3.5.2 Genes involved in lipid metabolism and visual system exhibit DNAme variation shared with Lake Malawi cichlids

Interestingly, one important example of biological process with high epigenetic variation is the visual homeobox *vsx1*, with hypermethylated promoter specifically in the benthic population. This could have considerable implication during development and might facilitate adaptation to dimly lit environments of the lake. Similarly, recent genetic analyses of Lake Massoko cichlids have highlighted increased genetic polymorphism in the benthic population in genes related not only to metabolic pathways (steroid synthesis specifically related to visual system, e.g. retinoic acid, the chromophore bound to rhodopsin) but also to the eye development and photoreceptors [181]. Other studies, focusing on different lakes, have shown that the genes involved in visual perception showed intense transcriptional dynamics in cichlids, with examples of change in opsin gene expression driven by differential binding of the transcription factor *tbxa* due to a single mutation [190]. Another study revealed considerable transcriptional changes related to the visual system (opsin genes among others), which could have accompanied the adaptation to dimly lit environment in fast speciating deep-water cichlids of the crater lake Barombi Mbo [191].

Altogether, this suggests that epigenetic and genetic processes might in concert facilitate visual adaptation to darker ecological habitats, with rapid evolution of gene structure (via genetic polymorphism or gene duplication for example) [181, 192] and differential transcriptional activities of genes part of the visual system [191], some possibly linked to TE activity, such as the DNA transposon subclass, Tc*Mariner* [190].

Epigenetic processes have been reported to be tightly linked to gene silencing - differential methylation at promoter levels might affect the interaction between DNA sequence and DNA-binding factors, such as transcription factors [78]. It would be therefore insightful to investigate further the biological relevance of the hypermethylated state of the visual homeobox *vsx1* in driving change in transcriptional activity in benthic fish, as such mechanisms could be responsible for rapid change in gene expression important for visual adaptation. Transcriptional activity of this homeobox gene could be quantified comparatively at dif-

ferent embryonic stages, in order to investigate their possible participation in differential development processes.

3.5.3 DNAme variation shared with Lake Malawi cichlids

Interestingly, most of DNAme variation in both Lake Malawi and Lake Massoko cichlids seem to be localised in similar genomic regions, in particular in promoters and CpG-islands (Fig. 3.9). This convergence in DNAme variation in regulatory genomic regions in both systems suggests that similar processes might play an important roles in short- and long-term divergence, possibly impacting transcriptome landscape in order to facilitate phenotypic plasticity. It is tempting to refer to these regions as *standing epigenetic variation*, exhibiting high epigenetic variation across many species, possibly promoting phenotypic plasticity (especially if located in regulatory regions) – further work is required to investigate the functional impact of such variation on transcriptional activity.

In addition, other genes related to vision show epigenetic variability in Lake Massoko cichlids, such as the ones involved in retinoic acid metabolism, important for the biogenesis of photoreceptor pigments. Interestingly, similar genes have been associated with high epigenetic variability in Lake Malawi cichlid radiation (see section 2.8.2). This suggests that epigenetic variability might be associated with similar biological processes in parallel in both lakes, facilitating adaptation to similar ecological habitats. It would interesting to draw a more accurate comparison of these molecular similarities, to understand, for example, if the deep-water *Diplotaxodon limnothrissa* of Lake Malawi share convergent epigenetic signatures with the benthic populations of Lake Massoko, both of which having colonised dimmer ecological niches in both lakes respectively.

Furthermore, it would be interesting to explore a possible link between epigenetic and genetic variations. In Lake Massoko, 'genomic islands of speciation' have been identified – these are regions of high genetic polymorphism (HDRs, or highly diverged regions) [181]. Epigenetic variability in the proximity of these HDRs might indicate that same loci might be under genetic and epigenetic pressures. Current efforts are being put to characterise this genetic-epigenetic overlap and investigate whether some genomic regions would show genetic and epigenetic diversity.

Further work would include RNAseq of livers of all the fish studied here, in order to characterise the regulatory functions of Massoko DMRs in modulating gene expression. Transcriptomic data are currently being generated.

3.6 Detailed methodology

3.6.1 Field sampling

In brief, wild specimens from Lake Massoko and proximate river were sampled by Prof. Martin Genner, Prof G. Turner and Alan Hudson and Alexandra M. Tyers. Liver tissues were preserved in RNA*later* before being stored in -80ºC.

The study presented in Chapter 3 is part of a collaboration with Profs. George Turner, Alan Hudson and Martin Genner. Genomic data for all wild individuals have been published [181].

Field sampling for genetic, epigenetic, morphological and stable isotope samples

Fish were collected using fixed gill nets and SCUBA. On being brought to the surface, fish were given an overdose of anaesthetic (MS-222). From each fish I collected a genetic sample (fin clip) that was stored in ethanol, and cut a fillet of the flank for stable isotope analyses that was sun-dried and stored with desiccant. Various tissues were dissected and stored in RNA*later*.

3.6.2 DNA isolation, NGS library preparation and sequencing

RRBS

As part of a collaborative project, collection of liver tissues in the field together with NGS library preparation in the readiness for RRBS were performed by Dr. Alan Hudson, University of Bristol, UK. These steps are briefly described below.

In brief, HMW-gDNA from liver tissues were isolated using column based approach. To produce RRBS libraries, the Premium RRBS kit (C02030032, Diagene) was used according to manufacturer's instructions. The method is similar to the one used for WGBS (see section 2.10.2, except for the restriction enzyme step, that replaces the sonication step). Briefly, HMW-gDNA were incubated with the restriction enzyme MspI. This enzyme binds double-stranded DNA and cleaves the phosphodiester bonds at 5′C.CGG 3′, irrespective of the methylation state. This results in enrichment for DNA fragments with high-CpG ratio. Digested fragments were then end-repaired and dA-tailed in order to ligate the methylated adaptor. Finally, single-end 50bp-long reads were generated on HiSeq2500.

WGBS

Some Massoko individuals were re-sequenced using WGBS in order to get a whole-genome resolution of liver methylome (See individuals in red, Fig. 3.4a,b and Table 3.4).

Genomic DNA extraction, quality control, WGBS library preparation and bioinformatic analyses thereof were carried out following the exact same protocol detailed in Chapter 2 (refer to section 2.10).

3.6.3 Analysis of RRBS data

Adaptor trimming and quality assessment (filtering of bases with Phred score <20) of RRBS reads were performed with `trim_galore --rrbs --fastqc` (v0.5.0, Babraham Inst.) – the `--rrbs` option ensures that the last two bases at 3′-end are removed as they were artificially added at the end-repair step. Reads were then mapped to *M. zebra* reference genome (UMD2a) with `bismark -p 3 -N 1` [161]. Clonal reads (PCR duplicates) were not removed, as recommended for RRBS data. Methylation status at each CpG site was called using `bismark_methylation_extractor -p --comprehensive --merge_non_CpG`. For downstream analyses, CpG sites with less than 4 or more than 100 read coverage were filtered out. PC analysis (`prcom`) as well as Spearman correlation (`cor`) and unsupervised clustering (`phetmap` and `hclust(dist)`) were performed using `R`. Methylation levels were averaged across non-sliding windows genome-wide using `bedtools` (`groupby` and `intersect`).

DMRs were called using DSS [177] ($p<0.05$, methylation difference $\geq 25\%$). DMR annotation files were merged and collapsed for the 3 pairwise comparisons to generate a single annotation file with unique DMR coordinates for both WGBS and RRBS dataset.

Gene ontology enrichment analysis was performed with g:Profiler [193].

Table 3.4 *A. calliptera* sp. Massoko and riverine *A. calliptera* - sample IDs

ID	Morph	RRBS	WBGS
G10E03	Benthic	x	
G10E04	Benthic	x	
G10E05	Benthic	x	x
G10E07	Benthic	x	x
G10E09	Benthic	x	
G10F02	Benthic	x	
G10F04	Benthic	x	
G10F07	Benthic	x	
G10F09	Benthic	x	
G10F10	Benthic	x	
G10G01	Benthic	x	
G10G02	Benthic	x	
G08F07	Riverine	x	
G09D02	Riverine	x	
G09D04	Riverine	x	
G09D05	Riverine	x	x
G09D07	Riverine	x	
G09D08	Riverine	x	
G09D10	Riverine	x	
G09E01	Riverine	x	x
G10C07	Riverine	x	
G10D06	Riverine	x	
G10D07	Riverine	x	
G08G02	Littoral	x	
G08G03	Littoral	x	
G08H08	Littoral	x	
G08H10	Littoral	x	
G08I06	Littoral	x	x
G10G05	Littoral	x	
G10G06	Littoral	x	
G10G09	Littoral	x	
G10G10	Littoral	x	
G10H02	Littoral	x	
G10H04	Littoral	x	x
G10H05	Littoral	x	

All wild-caught, *A. calliptera* males in full breeding colours. $n \geq 11$

Chapter 4

Plasticity and Heritability of DNA methylation in cichlids

4.1 Background

In chapters 2 and 3, I detailed the epigenetic variation observed in cichlids of Lake Malawi with distinct eco-morphological and trophic adaptations, as well the drastic epigenetic modifications possibly facilitating the adaptation to the different ecological niches of Lake Massoko.

In the first part of this chapter, I first aim to investigate the plasticity and dynamics of liver methylomes upon environmental perturbation. Riverine, littoral and benthic *A. calliptera* were caught in the wild, reared and bred in tanks, under controlled laboratory conditions.

The second part of chapter 4 deals with the study of the heritability of methylome patterns in cichlid hybrids. Hybridisation is common in Lake Malawi, possibly underlying the emergence of novel phenotypes, absent in the parental taxa [147, 140].

The results of chapter 4 are preliminary and are the focus of current investigation and experimental work.

4.2 Common Garden cichlids - Plasticity of liver methylome

4.2.1 Background

In chapter 3, I highlighted unique patterns of DNA methylation in two speciating monophyletic populations of *A. calliptera* sp. Massoko. Gain of methylation at promoter regions of key developmental and metabolic genes is a distinct feature of the methylome of these fish. Such epigenetic variation could have promoted invasion and adaptation to the unique ecological habitats of Lake Massoko, possibly facilitating phenotypic diversification. It is therefore crucial to characterise further the crosstalk between the environment and DNA methylation, in order to understand the biological function of such an interaction.

In this section, I aim to assess the extent to which environmental perturbations could affect and alter the methylome of liver tissues by breeding and rearing *A. calliptera* sp. Massoko in a controlled environment. I hypothesise that there is a direct crosstalk between the environment and DNA methylation, possibly accounting for changes in phenotype (such as gene expression patterns).

The liver methylomes of tank-reared G1 (first generation) offspring specimens bred from wild *A. calliptera* specimens from Lake Massoko were generated. This so-called common garden experiment ensures littoral and benthic morphs were reared and bred in the same controlled tank environment (same food source, same lighting conditions). This experiment allows us to investigate how plastic the methylation landscape is in response to environmental perturbations, such as domestication from wild to laboratory conditions. This will allow to assess the extent to which the methylome is shaped by the environment as opposed to fixed/innate ecomorph-specific methylome differences. Methylome variation associated with such an environmental perturbation will be identified and characterised.

Table 4.1 Sampling size - Common Garden experiment

Population	sampling size (n)
A. calliptera spp. Littoral Massoko (G1)	2
A. calliptera spp. Benthic Massoko (G1)	2
A. calliptera sp. River Itupi (Fn)	2

Liver tissues of full breeding-colour male specimens, all tank-reared.
All WGBS

In order to investigate and characterise the plasticity and dynamics of the liver methylome in response to environmental changes in these two natural populations of *A. calliptera* cichlids

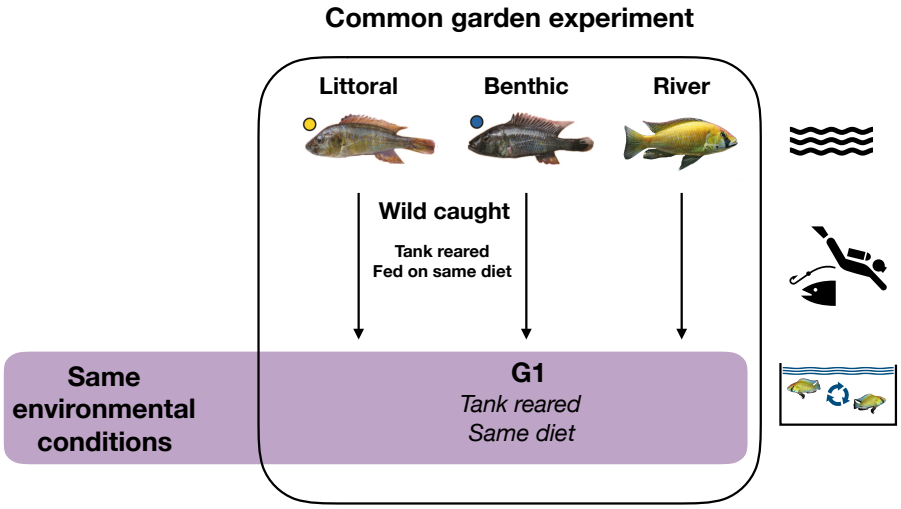


Fig. 4.1 Experimental design - Common Garden experiment. Specimens of the benthic (deep/blue), littoral (shallow/yellow) and riverine groups of *A. calliptera* sp. Massoko were caught in the wild. Specimens were brought to the UK by collaborators, Prof. G. Turner and colleagues. Caught specimens were then reared in the same tank conditions and eventually bred to produce G1 offspring. Note that the two different ecomorph populations were kept in separate tanks. Offspring (G1) were also isolated in separate tanks (same tank conditions). Note that G1 male individuals exhibit the same breeding colours as their the wild specimen counterparts (yellow or blue). Liver and muscle tissues of male G1 specimens of both groups were collected to generate WGBS libraries, with the aim to characterise DNA methylation dynamics associated to adaptation to a common environment. Wild Massoko specimens, sequenced and analysed in Chapter 4 are not the direct parental individuals of the G1 analysed hereafter.

of Lake Massoko, liver tissues of two G1 male specimens (reared in controlled-environment tanks) bred from wild individuals for both the benthic and littoral populations were sequenced. In addition, liver tissues of two tank-reared *A. calliptera* Itupi males (the river Itupi is a small stream directly connected to the Mbaka river system[1]) were also sequenced (WGBS). These are not part of Lake Massoko system but are tank-reared riverine subspecies, closely related to the ancestral riverine *A. calliptera* of Lake Massoko, just like *A. calliptera* sp. Mbaka in Chapter 3 (see Table 4.1).

4.2.2 WBGS of Common Garden *A. calliptera*

WGBS reads were generated from liver tissues and mapped to the same reference genome. Only conserved CG dinucleotide sequences between all samples showing high sequencing coverage were analysed. See section 4.6 for detailed methodology.

[1] *A. calliptera* samples collected from Itupi [here, chapter 4] and Mbaka rivers [chapter 3] are likely to be very closely related species [181].

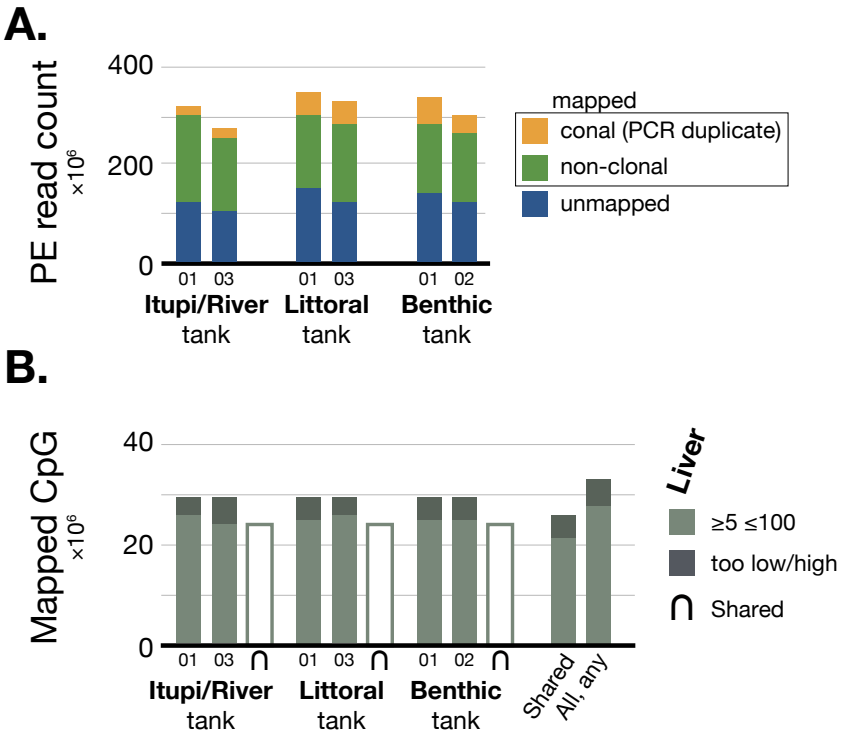


Fig. 4.2 Mapping efficiency of WGBS reads and CpG calling. A. Total number of WGBS paired-end reads generated (150bp long) from liver tissues for all common-garden experiment (tank) specimens. Individual samples were sequenced on a single HiSeq4000 lane each. Unmapped reads in blue, clonal mapped (PCR duplicates) reads in yellow and non-clonal mapped reads in green. All mapped to *M.zebra* reference genome (UMD2a). **B.** Total number of CpGs mapped to *M. zebra* reference genome (using clonal reads, in green in a.) for each sample. For each group (i.e. Itupi/River, Littoral and Benthic), the total of number of shared CpG sites are indicated by the intersection symbol ∩ (empty bar, green outline).

On average, 318.3 ± 26.7 million paired-end reads were generated. Similar to the Malawi dataset, sequenced reads were mapped to the Lake Malawi reference genome (on average, 58.8 ± 2.9% of all reads mapped uniquely, equivalent rates to the Lake Malawi dataset). Only uniquely mapped and non-clonal reads were analysed further (149.3 ± 16.4 million non-clonal, uniquely mapped paired-end reads; Fig. 4.2a.). Bisulfite conversion rate was high for all samples (97.5 ± 0.7%, not shown). Of all the 32.6 million CpGs sequenced among the six samples (see Table 4.1), 26.12 million CG dinucleotide sites were shared between all of them, and 21.4 million were mapped and shared in all samples at a sufficient coverage (5-100 uniquely mapped, non-clonal reads (Fig. 4.2b). DNA sequence conservation is therefore very high, as expected for different populations of *A. calliptera*. Any drop in CG counts could be due to sequence differences (unmapped), or too low/high a coverage.

Overall, cytosine methylation levels are high genome-wide, with high read coverage at CpG sites (Fig. 4.3a,b). The distribution of methylation is bimodal, with the majority of the genome being hypermethylated and a very small fraction of it not being methylated at all. Interestingly, riverine specimens show lower DNAme levels (both tank and wild, 75.5 $\pm$ 0.3%) in liver compared to Benthic and Littoral (78.4 $\pm$ 0.11%, mean $\pm$ SEM; Fig. 4.3a.). This higher level of methylation has been highlighted in the RRBS dataset as well – the wild specimens of the benthic population showed globally slightly higher methylation levels (see Chapter 3, Figs. 3.3b and 3.9b).

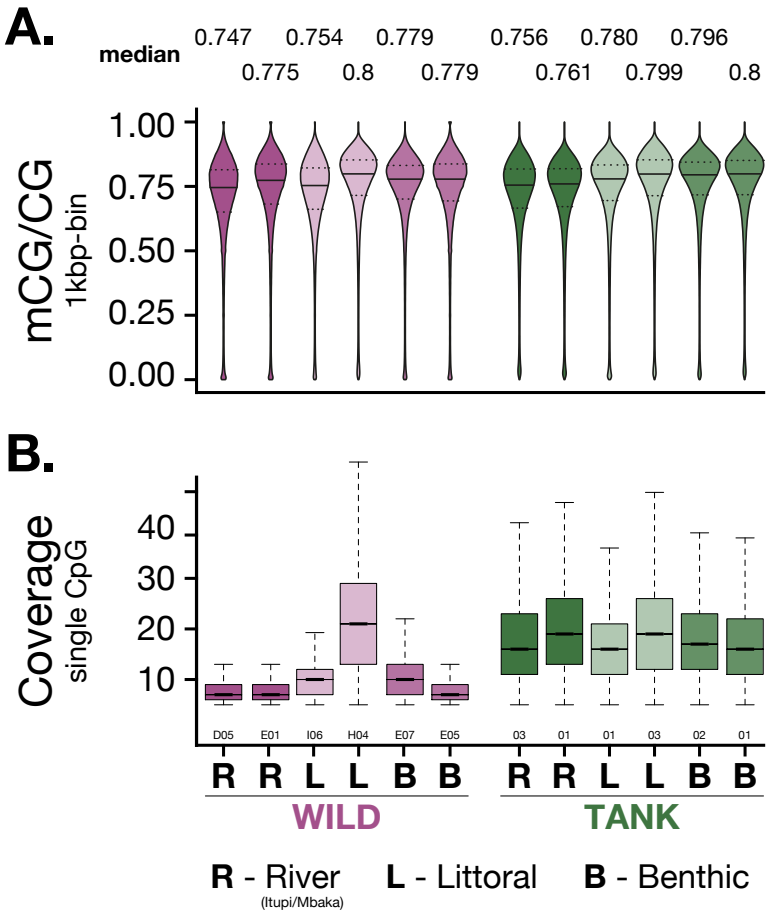


Fig. 4.3 Genome-wide liver methylomes of tank vs wild *A. calliptera* cichlid fishes. A. Violin plots representing the average DNAme levels in each liver sample (genome-wide non-sliding 1kbp bins [n=680,331]). Black lines and dotted lines represent median and 1st/3rd quantiles, respectively. **B.** Non-clonal read coverage at CpG sites (4-100 read coverage). Wild specimens highlighted in purple and tank (common garden) specimens in green. R, L and B: Riverine (wild specimens from River Mbaka, introduced in Chapter 3 [see section 4.2.1]; tank specimen from River Itupi), Littoral and Benthic (same wild specimens as the ones introduced in Chapter 3), respectively. w and t, wild and tank-reared and bred, respectively.

4.2.3 Liver methylome dynamics in Common Garden cichlids

Here, I then characterised methylome dynamics in the livers of different populations of wild-caught and tank-reared *A. calliptera* fish under controlled laboratory conditions. I aimed at identifying the regions that show DNAme variation upon environmental perturbation.

To this end, the variation in liver methylomes of wild vs tank-reared *A. calliptera* specimens from Lake Massoko was characterised by means of both principal component analysis and pairwise comparisons of Spearman correlation.

Widespread increase in DNA methylation levels at promoter regions possibly accompanying colonisation of Lake Massoko in *A. calliptera* cichlids

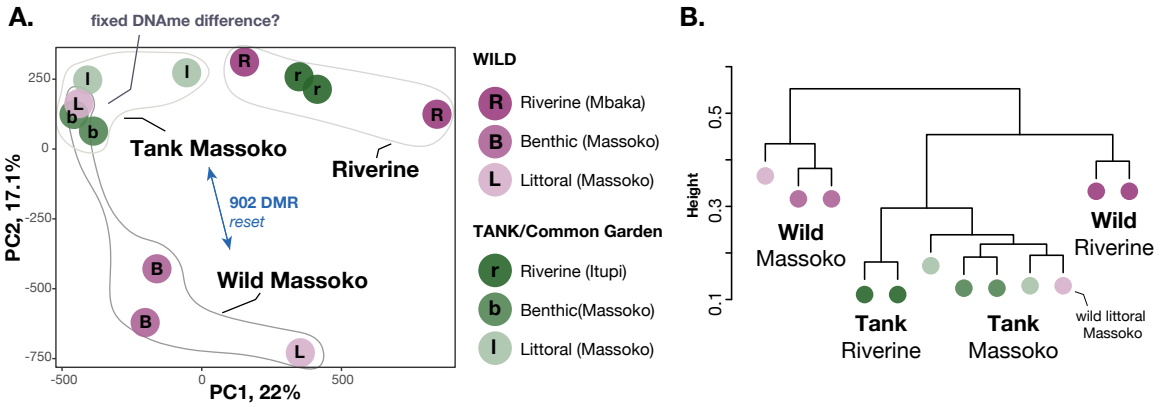


Fig. 4.4 Unique liver methylome patterns in wild *A. calliptera* specimens of Lake Massoko. A. PC analysis of liver methylome variation in wild versus tank-reared specimens. Wild specimens are shown in purple, tank-reared specimens (common garden, same controlled environment) in green. Benthic, Littoral and Riverine specimens are represented by upper-case and lower-letters, respectively, for wild and common garden individuals (B/b,L/l,R/r). The number of DMRs found between the wild and tank populations of benthic and littoral specimens is indicated in blue. **B.** Unsupervised clustering based on pairwise Spearman correlation scores between liver methylomes of all samples (Euclidean's distances are plotted).

Interestingly, wild individuals from Lake Massoko (both benthic and littoral) exhibit unique liver methylome patterns, clustering away from any tank-reared (benthic/littoral) specimens and any wild or tank-reared riverine populations (Fig. 4.4a,b). The analysis revealed also one putative outlier: one wild littoral specimen is clustering with tank-reared Massoko specimens, for unknown reasons. Moreover, the liver methylome of the riverine population reared in tanks (common garden) resembles that of the wild riverine population, suggesting that methylome plasticity is not only shaped by the artificial environment of

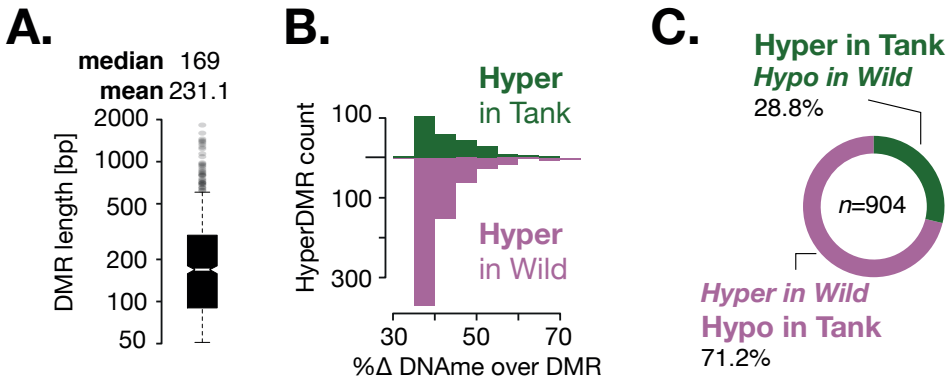


Fig. 4.5 Wild populations of benthic and littoral *A. calliptera* of Lake Massoko show widespread hypermethylation. A. Length [bp] of all DMRs found (902 in total) between tank-reared (Common Garden) and Wild specimens of both Benthic and Littoral populations. **B.** Histograms of the count of hypermethylated DMRs found between wild and tank-reared specimens (benthic and littoral only). **C.** Pie chart summarising the count of hyper- vs. hypomethylated DMRs between wild and tank-reared Benthic/Littoral specimens.

laboratory tanks, but still exhibits a degree of ecomorph-specificity. The tank-reared riverine population serves as an important control to assess the impact of laboratory-condition tanks on the liver methylome. Differences in the environment could cause DNAme variation. However our results seem to provide evidence against this hypothesis: although the direct environment (tank vs wild) seems to greatly influence the liver methylome (especially for the wild benthic and littoral specimens), the different populations exhibit enough variation to distinguish the riverine specimens from the littoral and from the benthic populations, regardless of whether they were reared in tanks or caught in the wild, suggesting that ecomorph-specific DNAme differences persist even when specimens are reared in the same controlled environment (Fig. 4.4b).

Strikingly, upon drastic environmental change, such as in this common garden experiments, only the wild populations of Lake Massoko appears to lose/reset wild-specific methylome patterns to resemble the ancestral-like riverine species (Fig. 4.4a, PC2). This could reflect the process of domestication, from wild to laboratory conditions, although wild riverine specimens do not show major differences in overall DNAme patterns compared to their tank-reared counteracts. It would be important to elucidate whether this change in DNAme specific to the wild benthic/littoral specimens is acquired during development in their respective wild habitats or whether they represent some fixed DNAme variation, divergent between the different wild populations of Lake Massoko. The widespread loss of DNAme observed in the wild specimens provides some evidence for the former.

I then characterised the differentially methylated regions between the wild and tank-reared benthic and littoral populations, in order to identify and characterise the genomic regions showing wild-specific methylation variation (Fig. 4.4a, PC2). In total, 902 DMRs have been identified between the wild and tank-reared population of *A. calliptera* of Lake Massoko (only littoral and benthic). DMRs are on average 230 bp long and show 40-60% difference in methylation levels overall (Fig. 4.5a,b). Tank-reared specimens specifically show reduced methylation states at more than 70% of total DMRs (Fig. 4.5c), suggesting that tank-reared individuals from the lake (benthic and littoral populations taken together) might reset some wild-specific hypermethylated DMRs to resemble the hypomethylated levels of the ancestral-like, riverine population (Figs. 2.19a and 4.5c).

Hypermethylation states at promoter region of metabolic genes is a feature of liver methylomes of wild *A.calliptera* sp. Massoko

The genomic localisation of the wild-specific hypermethylated DMRs were identified and gene ontology enrichment analysis was performed to investigate potential biological functions underlying the wild-specific increase of methylation observed in Lake Massoko cichlids.

Promoter regions and CpG islands (CGI) appear to be enriched for DNAme variation associated with the wild benthic and littoral populations, in particular at genes involved in metabolic pathways (Fig. 4.6a,b). This could suggest altered transcription activity at key genes due to differential methylation states, but not yet validated as no RNAseq has been done so far. Altered methylation levels at promoter regions might greatly influence the binding interaction of methyl-sensitive DNA-binding actors at specific TSS, thus resulting in modified gene expression patterns. DNA-binding proteins, such as some families of transcription factors, have been shown to be differentially methylated in *A. calliptera* sp. Massoko. They might in turn participate in modifying the methylation states at other promoters as well [80, 78] (see Fig. 3.6f). Surprisingly, transposable elements and intergenic regions show slight depletion in wild-specific DNAme variation.

Altogether this suggests a possible role of promoter regions and CGI in facilitating phenotypic plasticity in the context of adaptation to the ecological niches of Lake Massoko. Similar observations have been made in radiating cichlids of Lake Malawi, in that DNAme variation is also preferentially associated with promoters regions, although, in the case of Lake Malawi cichlids, intergenic regions and transposons are also enriched in DNAme variation. In Lake Massoko, the slight depletion in TE-associated DNAme variation does not rule out any adaptive role of such elements, as DMRs overlapping TEs are the most numerous in total (Fig. 4.6a). Wild-specific DMRs associated with gene bodies, also not showing

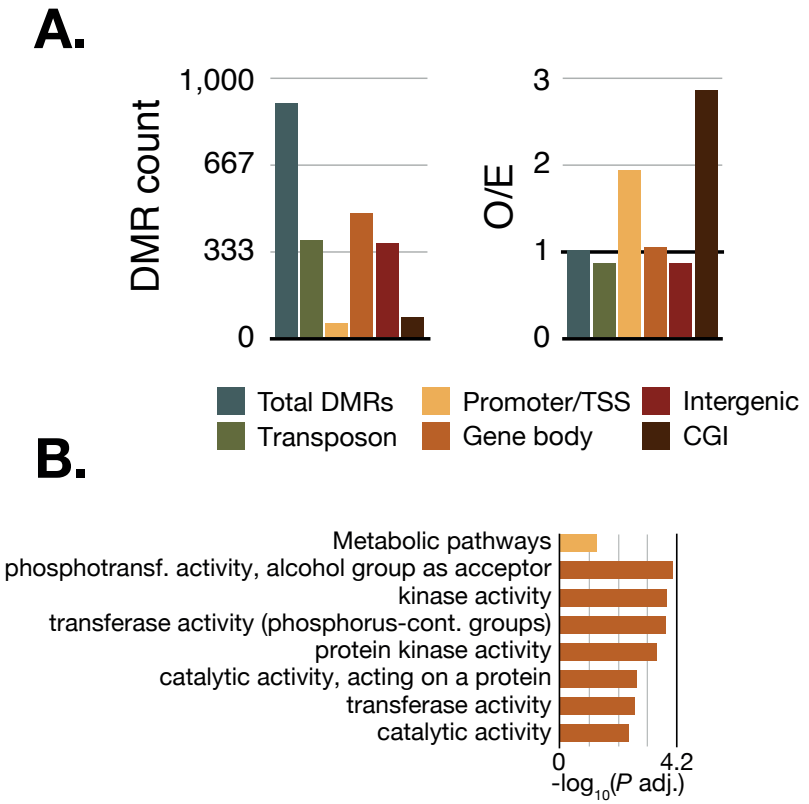


Fig. 4.6 Wild-specific hypermethylated DMRs are enriched in promoter and CGI regions. A. Genomic localisation (left) of the 902 wild-specific DMRs. These can be localised in promoter regions (TSS), repeats and transposons (transposon), exons or introns of genes (gene body, GB) and intergenic regions. Enrichment analysis for genomic elements (right) - 'expected' refers to the DMRs localisation as expected by chance if DMRs were to be randomly shuffled across the genome (10 iterations; see methods). **B.** GO enrichment analysis for hypermethylated DMRs located in promoter regions and gene bodies.

a particular enrichment, appear to be related to genes involved in enzymatic reactions, in particular for protein kinase activities (Fig. 4.6b, orange label), possibly involved in signal transduction or in kinase-mediated regulation of metabolic processes.

Furthermore, genes involved in metabolic pathways in particular appear to show considerable variation in liver methylomes of the populations having colonised the different ecological habitats of Lake Massoko. Adaptation to different sources of nutrients has been highlighted to be an important adaptive characteristic of Lake Massoko cichlids (see Fig. 3.5 and Malinsky *et al.* [181]). Interestingly, many cytochrome P450 genes, in particular the ones related to lipid metabolism, show considerable methylation variation in cichlids of Lake Malawi and Massoko. The benthic and littoral populations of *A. calliptera* in Lake Massoko specifically show increased cytosine methylation levels at the promoter of one cytochrome

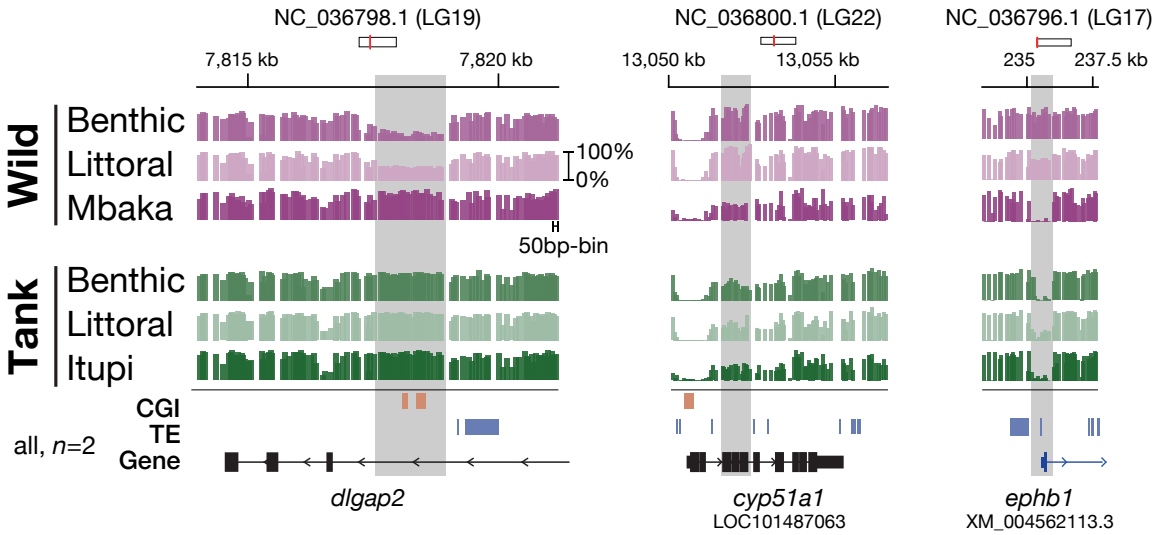


Fig. 4.7 Examples of wild-specific DMRs. Examples of hypomethylated and hypermethylated DMRs in wild *A. calliptera* of Lake Massoko. Averaged mCG/CG (0-100%) of two biological replicates in non-sliding 50 bp-long windows genome-wide for the six different populations. **Left**, DMR located in an intronic region of the gene *dlgap2*, disks large-associated protein 2; **middle**, DMR in three exons of the gene lanosterol 14α-demethylase, *cyp51a1*; **right**, DMR in the promoter region of the gene *ephb1*, ephrin type-B receptor 1). CGIs2, predicted CpG islands; TE, transposable elements and repeats. DMRs are boxed out in grey.

P450 (*cyp51A1*), that is the enzyme lanosterol 14 alpha-demethylase participating in the bio-synthesis of zymosterol from lanosterol, as part of the sterol biosynthetic pathway (Fig. 4.7, panel in the middle). Other metabolic genes[2] have a DMR at their promoter. This could suggest wild-specific altered lipogenesis, with putative impact on steroid-mediated signalling or cholesterol metabolic processes.

Not only genes involved in hepatic functions exhibit DNAme variation. For example, the intronic region of the gene *dlgap2*, with reported important functions in the molecular organisation of synapses and in neuronal cell signalling, shows a wild-specific hypermethylated state in wild *A. calliptera* cichlids (Fig. 4.7, left panel).

As previously mentioned, it is important to note that many DMRs might be species-specific, rather than tissue-specific, with limited relevance for liver functions but rather reflect embryonic memories (relics). Such DMRs could have borne important regulatory functions at earlier stages during development, and could have been passed on throughout cell divisions

[2]such as *ggt5a*, glutathione hydrolase 5 proenzyme-like; *plcd4b*, phospholipase C delta 4; ENSMZEG00005013864, UDP-glucuronosyltransferase 2A1; *agxt2*, alanine–glyoxylate aminotransferase 2; *pfkma*, ATP-dependent 6-phosphofructokinase; ENSMZEG00005015314, putative bifunctional UDP-N-acetylglucosamine transferase and deubiquitinase ALG13; *cers3a*, ceramide synthase 3; *c1galt1c1*, C1GALT1 specific chaperone 1

and differentiation – the machinery capable of interacting with embryonic DMRs being only present at specific embryonic times [194, 169].

Global DNA methylation resetting in tank-reared cichlids to resemble the riverine, ancestral liver methylome

The wild populations of *A. calliptera* from Lake Massoko present unique liver methylome patterns, with a distinctive gain in methylation at many genomic loci. Within one generation time, Lake Massoko specimens reared in tanks (in the same artificial environment) from wild-caught parents appear to lose these distinctive patterns (Fig. 4.5b), via a potential DNAme resetting/reprogramming. Their liver methylomes seem to resemble the ones characteristic of the ancestral-like riverine *A. calliptera* population.

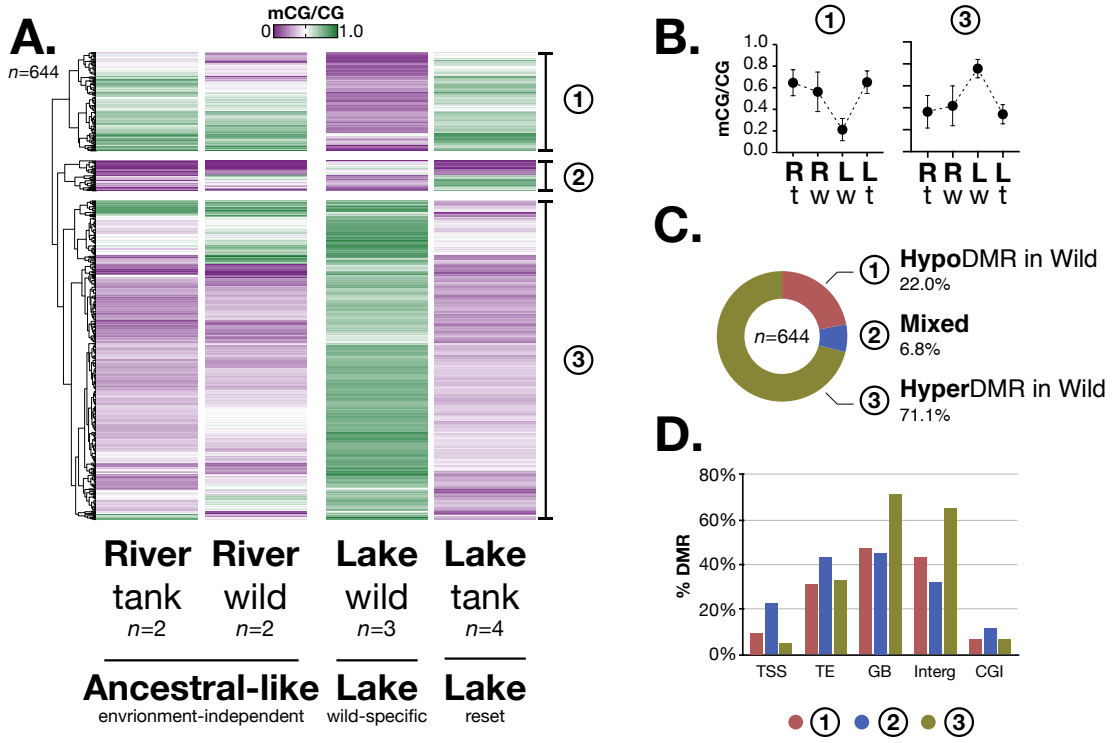


Fig. 4.8 Wild-specific liver methylome is reset to resemble ancestral methylome in common garden experiment. A. Heatmap showing DNA methylation states at each of the DMRs found between wild and tank-reared Benthic and Littoral populations (see Fig. 2.19a). Liver WGBS data, mean mCG/CG per group, *n* indicates biological replicates for each group. Three clusters based on methylome patterns. **B.** Plot of averaged DNAme levels at DMRs of two clusters. Error bar, mean±sd. **C.** Pie chart summarising the number of DMRs for each cluster. **D.** Genomic localisation of DMRs for each cluster, given in percentage total DMR per cluster. TSS, promoter region; TE, transposable element or repeats; GB, gene body; Interg., intergenic; CGI, CpG-island.

I further investigated the dynamics of DNAme in loci showing variation in wild specimens compared to the tank-reared offspring bred from wild-caught individuals. Strikingly, most (93%) of the DMRs unique to the wild benthic and littoral populations show changes in

methylation to resemble the riverine ancestral-like methylome patterns upon common garden experiment (Fig.4.8a). Hypermethylated DMRs (71%) found between wild and tank-reared benthic and littoral specimens show a decrease in methylation, matching lower methylation states seen in the ancestral-like, riverine specimens (Fig. 4.8a-c, cluster 3). Inversely, hypomethylated DMRs in the wild lake population show an increased methylation state in tank-reared lake fish (Fig. 4.8a-c, cluster 1). A small fraction of DMRs (6%) show more variable and unspecific methylome patterns (cluster 2). Note that DMRs were initially found and predicted without taking into account any riverine species – nevertheless patterns of DNAme in liver of tank-reared lake (benthic and littoral) specimens were similar to these of riverine individuals. Hypermethylated regions in wild-caught cichlids are particularly present in gene bodies and intergenic regions, possibly *cis*-regulatory regions (Fig. 4.8d).

Further work would focus on generating RNAseq data for wild and tank individuals in order to link DMRs found between wild and tank specimens to changes in transcriptional activity.

4.3 Discussion and future work - Common Garden experiment

In this section, I aimed at investigating the dynamics of DNA methylation in liver upon environmental perturbation. To address this aim, I quantified and compared the liver methylomes of wild and tank-reared (G1) cichlid individuals of the same subpopulation.

4.3.1 Plasticity of liver methylomes, resetting of DNAme patterns upon environmental perturbations

In chapter 3, I reported that the liver methylomes of wild *A. calliptera* sp. Massoko (both littoral and, in particular, benthic) specimens exhibit a widespread hypermethylated state in particular at promoter regions, compared to the riverine, putatively ancestral *A. calliptera* sp. Itupi population (Figs. 3.6 and 3.12). This suggests that specific DNAme patterns have accompanied the invasion of the deeper parts of Lake Massoko, possibly promoting phenotypic plasticity and adaptation (see Chapter 3). However, it remained unclear how much of the observed liver methylome divergence is specific to the benthic and littoral populations, as opposed to be exclusively shaped by the environment.

Strikingly, upon domestication from the wild to laboratory controlled conditions, most of the methylome observed in wild *A. calliptera* sp. Massoko specimens were lost and resembled DNAme patterns of the ancestral-like, riverine population (Fig. 4.8a). This observation suggests that the specific liver methylome patterns unique to wild *A. calliptera* of Lake Massoko are very plastic upon environmental perturbations. Adaptation to deeper habitats of the lake could have been accompanied by a global increase in DNA methylation, in particular at promoter regions of important metabolic genes (Figs. 4.6 and 4.7). It remains unclear when such remodelling happens, whether it occurred in the wild G0 specimens or during the development of G1 individuals in tanks, which warrants further investigation of DNAme patterns during development. Altogether, these preliminary results postulate that reset DNAme patterns observed in tank-reared individuals could have mediated the rewiring of the transcriptional network. Such a widespread hypermethylated state in the wild benthic populations could confer an important fitness advantage to the benthic population of the lake.

4.3.2 DNA methylation divergences in littoral and benthic populations

Nevertheless, one important question remains: how much of the variation in DNA methylation observed in the lake is fixed between the benthic and littoral populations, possibly underlying phenotypic divergence. Current work aims at addressing this by identifying DMRs that are retained in tank-reared and wild benthic populations independently of environmental perturbations. Such ecomorph-specific variation could be fixed and inherited within one specific population, thus promoting short- and long-term phenotypic divergence. Initial analysis suggests that the benthic and littoral ecomorphs have some distinct DNAme patterns in the wild and in laboratory conditions (PCA cluster both ecomorphs apart, Fig. 4.4), however this does not provide evidence that the same DMRs are involved here. The benthic and littoral populations of *A. calliptera* sp. Massoko are philopatric, in that males are territorial and females probably rear their fry in the same environment. Fry are likely to grow to adulthood in the same environment as well. Adaptation to the dark, zooplankton-rich waters of the benthic zone is therefore required. Genetic predispositions (such as the ones related to rhodopsin genes [181]) enable them to thrive, as such individuals are likely to be less fit in other ecological niches. Eventually, more traits can become fixed genetically, generating islands of divergence, described by Malinsky, Challis and colleagues, before whole genome fixation. I postulate that similarly DNA methylation at regulatory regions could act in concert with genetic polymorphism to generate selectable phenotypic plasticity, promoting adaptation, without more permanent alterations in the underlying DNA sequences.

4.4 Inter-species AxA hybrids - epigenetic inheritance in cichlids

4.4.1 Background

In the first part of chapter 4, I investigated the plasticity of methylome variation upon environmental perturbation and concluded that the unique liver methylome pertaining to wild *A. calliptera* specimens of Lake Massoko was mostly lost and reset to resemble the one of the ancestral-like, riverine lineage. This epigenetic plasticity is thought to have occurred within one generation time. However, it remains unclear where and when these drastic changes might have occurred: in the parental taxa, after being wild caught and brought back to laboratory tanks, or in the G1 specimens specifically when reared, or a combination of both. I concluded there is a tight crosstalk between the unique environment of Lake Massoko and the epigenetic landscape of liver tissues. Although it is tempting to postulate that such an epigenetic plasticity might be a common feature in cichlids, this mechanism might be unique to Lake Massoko cichlids as well. Furthermore, the upstream 'sensing' mechanisms underlying such changes in DNAme variation remain elusive and the heritability of such patterns is unknown. Future work would focus on generating hybrids between the wild populations of littoral and riverine species, to investigate the inheritance of different methylome patterns. Another experiment would be to artificially place wild littoral specimens back into the ancestral-like river (or some of the riverine specimens in the lake), to test our hypothesis that the liver methylome would undertake global reprogramming to resemble the ancestral riverine methylome.

In plants and mammals, there is evidence that methylome landscape is transmitted across generations, even in the absence of any original environmental trigger responsible for epigenetic alteration and even in the instance of global DNA methylation reprogramming [2, 1, 16]. In teleost fish, it remains unclear whether the lack of global DNA demethylation during development is conserved [3] and whether any alteration in epigenetic landscape could be intergenerationally inherited. Thus, it is important to distinguish all the aspects underlying DNAme variability and inheritance. In the first part of chapter 4, I concluded that the environment would greatly affect the methylome of genetically very similar cichlids. However, it remains unknown whether these altered patterns can be transmitted intergenerationally even if the absence of the initial environmental trigger.

Here in the second half of chapter 4, we have taken advantage of some of the more moderate cases of reproductive isolation observed in Lake Malawi fish reared in tanks in

order to generate inter-species hybrids (the reproductive isolation in Lake Malawi is usually strong in the wild). DNA methylation patterns of the parental taxa, as well as of the hybrids, would be compared to investigate the genetic vs non-genetic components of methylome inheritance. Genomic imprinting has not been observed in teleost fish, and no homologous gene of the mammalian *DNMT3C* underlying mammalian imprinting has been characterised, at least in zebrafish [61]. It would be interesting to understand whether any parental biases in methylome transmission would be observed, or instances of 'hybrid vigour', whereby hybrids would display transgressive phenotypes, not observed in the parental taxa, possibly increasing hybrid fitness compared to parental lines [195]. In plants, the methylome of hybrids have been shown to generate novel patterns, thought to be mainly mediated by small non-coding RNAs (RdDM), possibly underlying phenotypic diversity [30, 196, 197].

In addition, it has been reported in plants that hybridisation might constitute an 'epigenomic shock', in that upon inter-breeding of individuals exhibiting significant methylome divergence, novel patterns of methylation and, consequently, gene expression might emerge, producing phenotypic diversity (or transgressive phenotypes) [2, 196, 198]. DNA methylation associated with TEs might be in particular divergent, and some TE families, perhaps unique to one species, might become active in the hybrids [30]. This is particularly relevant in cichlid populations evolving in sympatry, with sometimes relaxed assortative mating and common hybridisation events [144, 140].

To this aim, I generated inter-species reciprocal cross in collaboration with Prof. George Turner at Bangor University. We selected *A. calliptera* spp. Itupi (the river Itupi is part of the Lake Malawi catchment) and *A. stuartgranti* spp. Usisya (named after the location of catchment in Lake Malawi) as the maternal and paternal taxa, receptively.

4.4.2 Inter-species hybrids - experimental design

Cichlid hybrids were bred from one *A. stuartgranti* spp. Usisya (AC) male with two *A. calliptera* spp. Itupi (AC) females (AxA hybrids) - two separate broods were produced using the same father but different mothers (Fig. 4.9 and Table 4.2). Unfortunately, the reciprocal cross (i.e. using AS females) has remained unsuccessful.

In terms of genetic divergence, both AC and AS show the average low sequence divergence representative of the cichlid flock of Lake Malawi (pairwise sequence divergence, 2.24 ± 0.008 SNP per kbp genome-wide; Hannes Svardal [102]).

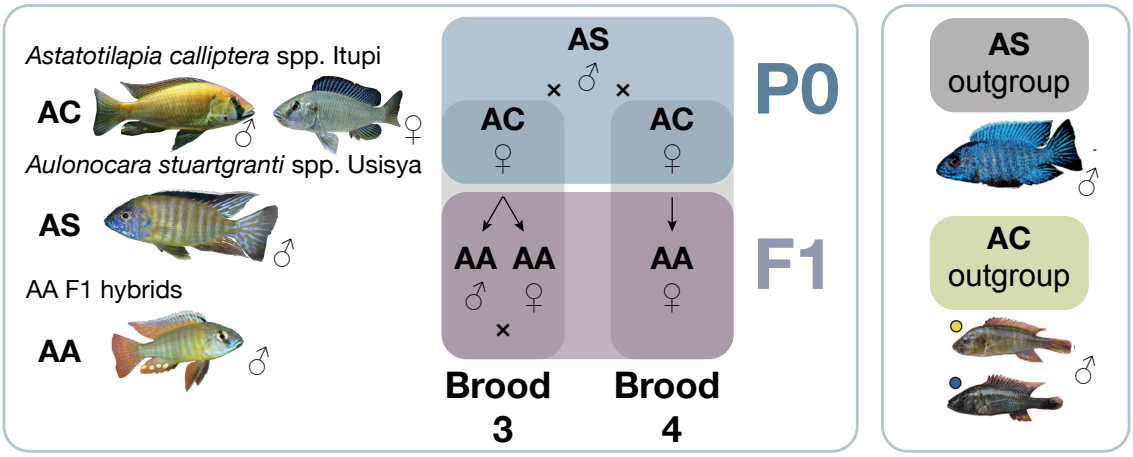


Fig. 4.9 Experimental design - inter-species cichlid hybrids. Two cichlid species, *Aulonocara stuartgranti* sp. Usisya (AS) and *Astatotilapia calliptera* sp. Itupi were inter-bred to produce cichlid hybrids in order to investigate the epigenetic landscape and inheritance of DNA methylation in hybrids. Two different broods were produced, both with the same AS male, and with two different AC females. The offspring (AA F1) are currently being bred to produce F2 progeny. In addition, two cross out-groups were used (*A. calliptera* sp. Massoko and *A. stuartgranti*) as a mean to assess the inheritance in F1 hybrids only.

Hybrid morphology and anatomy

This inter-species (AxA) cross project is part of a collaboration between the laboratories of Eric Miska, Emília Santos, Martin Genner and George Turner. A complete phenotyping of F1 and eventually F2 hybrid individuals is being conducted at the moment, to understand the genetic basis of particular Lake Malawi cichlid phenotypes (especially with regards to pigmentation patterns, behaviours and lateral line).

4.4.3 Global characterisation of liver methylome in hybrids

In order to characterise the global methylome landscape in cichlid hybrids, WGBS data of liver tissues for both the parental and hybrid specimens were first generated. All the sequencing reads were mapped to the same reference genome (*M. zebra*, UMD2a) and only conserved CpG sites among all individuals with high read coverage (5-100 non-clonal paired-end 150 bp-long reads) were further analysed.

Female and male hybrids as well as the parents similarly exhibit overall high methylation levels (78.0 $\pm$ 2.3%, median$\pm$sd; Fig. 4.10), akin to what has been described in other Lakes Malawi and Massoko cichlid populations (see Chapters 2 and 3). Furthermore, specimens (both parents and hybrids) show 97.6% sequence conservation at mapped CpG sites genome-

Table 4.2 Sampling size - hybrid experiment.

	sample size (*n*)
A. stuartgranti sp. Usisya	1
A. calliptera sp. Itupi	2
AxA hybrids (F1)	3
A. stuartgranti, cross-independent	2
A. calliptera sp. Massoko, cross-independent	4

Liver tissue of sexually mature specimens were used for WGBS, all tank-reared

wide, reflecting the high level of sequence conservation, even between the two parental taxa. In addition, such high DNAme levels were comparable to both control AS specimens, unrelated to the cross. This suggests that hybrid methylomes, globally, do not deviate much from the parental and cross-independent specimens methylomes.

4.4.4 Hybrid-specific DNA methylation patterns

I then characterised epigenetic variability between F1 hybrids and parental taxa, including cross-independent specimens (control) reared in tanks to account for tank- and species-specific variability.

As expected based on results from previous chapters, a phylogeny (or clustering) could be created solely based on methylome divergence at conserved CG dinucleotide sequences, suggesting species-specific methylome patterns (Fig. 4.11). Note that species-specific patterns are observed even though all individuals were reared in tanks; two subpopulations of tank-reared *A. calliptera* from Lake Massoko (benthic and littoral) and River Itupi (AC mothers) show enough methylome divergence to cluster separately. Furthermore, AC individuals all cluster away from AS specimens (Fig. 4.11). This suggests that, even though the local environment might alter the epigenetic landscape (see first section of chapter 4), there are still distinct methylome patterns unique to even a subpopulation of genetically closely-related fish. Species-specific patterns of DNA methylation is therefore transmitted in a common garden experiment (independently of the environment). Importantly, this does not rule out any genetic components mediating such methylome divergence and inheritance.

I then compared liver methylomes based on pairwise Spearman correlation scores. Liver methylomes of the two parental taxa are similar at 76-78% at conserved underlying sequences (Fig. 4.12). Intra-species liver methylome similarities are higher, respectively 86% and 80-84% between *A. calliptera* spp. Itupi and *A. stuartgranti* spp. Usisya individuals. F1

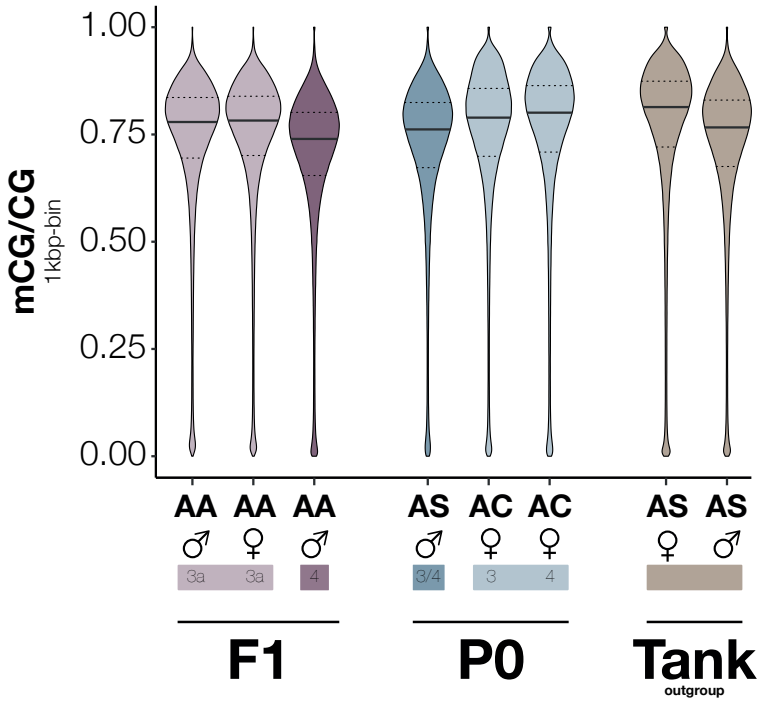


Fig. 4.10 High genome-wide DNAme levels in livers of both parental and hybrid specimens. Genome-wide DNAme levels in liver tissues of the parental (P0) specimens (AS, females and AC, male; in red) and the three (F1, in blue) AxA hybrid individuals from both broods. Two tank-reared AS specimens, independent of the hybrid cross were used as AS outgroup ("Tank", in beige). Average mCG/CG in 1kbp-windows. 4-100 unique read coverage at conserved CG dinucleotide DNA sequence only.

hybrids shared the highest levels of liver methylome patterns, in particular for siblings (88% liver methylome correlation for brood3; inter-brood, 83-84%). Inter-species and hybrids methylomes are very similar, possibly with major epigenetic divergence (>30% DNAme difference) localised at a small number of large genomic loci (Fig. 4.13). This suggests that although most of the liver methylome patterns might be similar, there are significant differences that may be unique to one taxon or to the hybrids, and could participate in phenotypic diversity.

Interestingly, F1 hybrids show some unique liver methylome patterns and cluster away from both parental taxa based on genome-wide methylome pairwise correlations and principal component analysis (Figs. 4.11 and 4.14). The methylome of F1 hybrids share nevertheless more similarities with the methylomes of parental taxa, as opposed to the control outgroup specimens (AS individuals independent of the cross). Similarly, PC analysis reveals that the methylome of F1 hybrids show intermediate patterns, resembling both parental liver

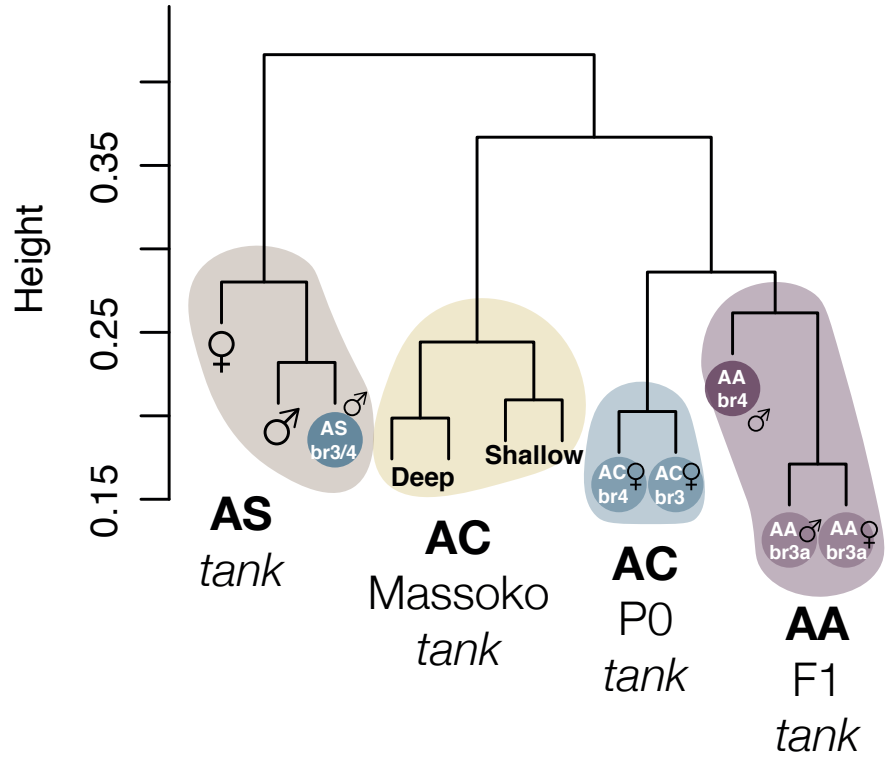


Fig. 4.11 Clustering of liver methylome variation in AxA hybrids. Phylogeny based on genome-wide methylome divergence at conserved underlying DNA sequences in AA hybrid experiment. Unsupervised hierarchical clustering based on Euclidean distances (inferred from pairwise Spearman's correlation matrix). All sequenced specimens were tank-reared. AS was the paternal taxa and included two cross-independent specimens (one male one female) and are highlighted in beige (male of broods 3 and 4 indicated in white text). AC was the maternal taxa as well as some cross-independent Lake Massoko individuals (all male individuals, in yellow). AC mothers of brood 3 and 4 are highlighted in blue. F1 AA hybrids are highlighted in purple.

methylome patterns (Fig. 4.14), suggesting a strong genetic basis of DNAme inheritance. Strikingly, some DNAme patterns seems to be unique to the hybrids (in particular hybrid AA male from brood4; Fig. 4.14). Such unique patterns of methylation could perhaps be described as a transgressive phenotypes, in that F1 inter-species hybrids could show traits not observed in the parental lines, which could promote adaptation (see Discussion in section 4.5).

Importantly, additional WGBS data of a larger population of F1 hybrids (possibly reciprocal cross) will provide stronger evidence for such transgressive DNAme patterns observed in hybrids.

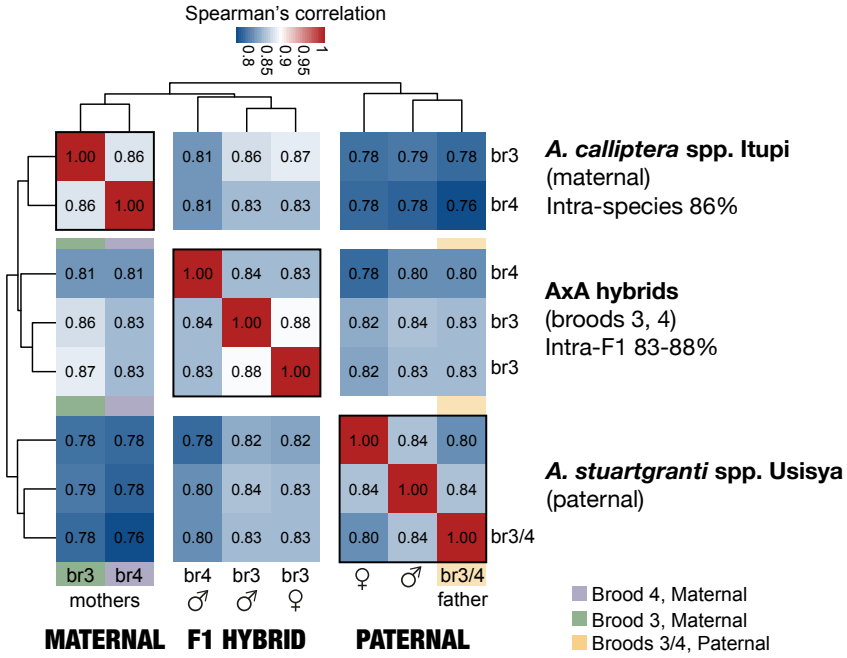


Fig. 4.12 Heatmap of genome-wide pairwise Spearman correlations of liver methylome variation in hybrids. Heatmap representing pairwise Spearman correlation scores of liver methylome divergence in hybrids and parental taxa. Based on average mCG/CG in 1kbp-long windows genome-wide. Hybrids consist of two broods (broods 3 and 4), were bred from the same AS father and from two different AC mothers. All individuals were reared and bred in laboratory tanks. Only cytosine methylation variation at conserved CG dinucleotide DNA sequence was analysed. CG sequencing coverage of 4-100 non-clonal paired-end reads. Single pairwise Spearman correlation (given in percentage of correlation) score is indicated in each heatmap square.

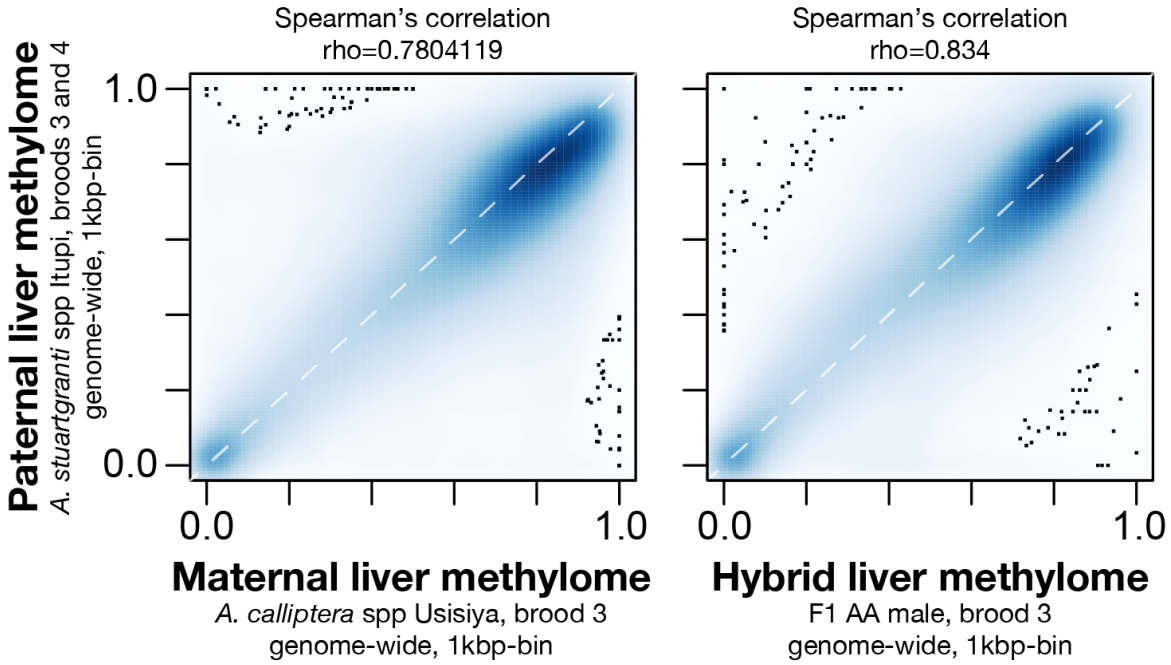


Fig. 4.13 Comparison of liver methylome landscape in hybrids. Biplots representing pairwise liver methylome comparison at 1kbp-bins genome-wide. y-axis represents the average cytosine methylation levels in paternal taxa (AS), x-axes represent the maternal liver methylome of AS brood 3 (left) and the AA F1 hybrid methylome (right). Only cytosine methylation variation at conserved CG dinucleotide DNA sequence was analysed. CG sequencing coverage of 4-100 non-clonal paired-end reads. Average mCG/CG in non-sliding 1kbp-long windows, genome-wide. White dotted lines represent 100% correlation or the same liver methylome.

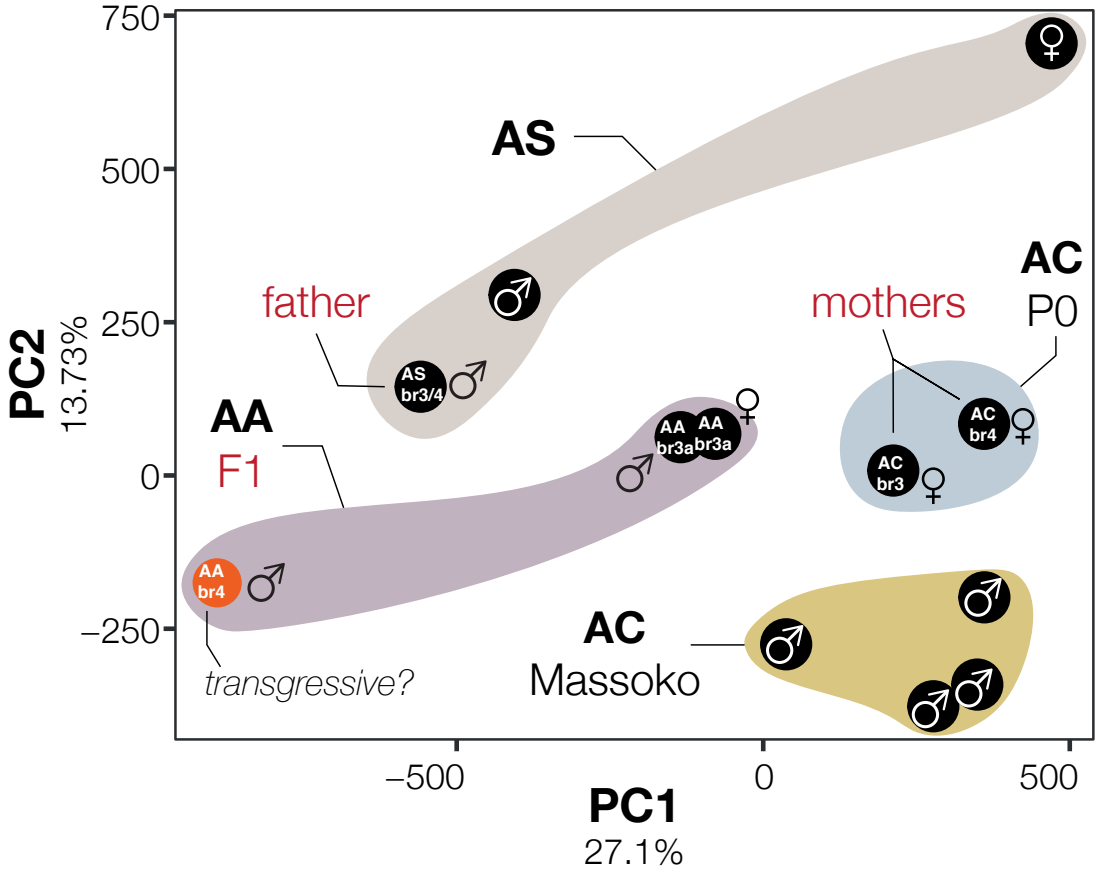


Fig. 4.14 Liver methylomes of hybrids resemble parental patterns. PC analysis of genome-wide liver methylome patterns in maternal and cross-independent AC, paternal and cross-independent AS and F1 hybrids of broods 3 and 4. Sex of each individual is shown. P0 stands for parental individuals. F1 hybrid specimen showing putative transgressive DNAme patterns is indicated in orange. AC Massoko specimens are used for comparison (as independent of the cross), *A. calliptera* sp. Massoko (benthic and littoral). All individuals were reared and bred in laboratory tanks. Only cytosine methylation variation at conserved underlying CG dinucleotide DNA sequences was analysed. CG sequencing coverage of 4-100 non-clonal paired-end reads were analysed only.

Loss of methylation at promoter regions in cichlid hybrids

Even though methylation at single CG dinucleotides might also show variability between parents and offspring, I focused on differentially methylated regions (DMRs) that are 50 bp-long in size at least, with an average methylation difference of $\geq$25% at CG sites between the parental taxa and F1 hybrids - these regions are likely to exert biological functions as opposed to single CpG sites.

I identified 380 DMRs between liver methylomes of the parental taxa and one F1 hybrid for brood 3, covering a total length of 82.9kbp (median length= 129bp, mean length=219.8bp). Interestingly, 79% of all the DMRs show hypomethylation in F1 hybrids, i.e. loss of methylation compared to parental taxa (hypermethylated genomic loci specific to the paternal taxon are visible at the top left corner of the left panel, Fig. 4.13). The genomic localisation of the hybrid-specific DMRs were then characterised.

Promoter regions and CGIs are extremely enriched for DMRs (respectively 2.1-fold and 7.1-fold enrichment compared over chance; Fig 4.15) – loss of cytosine methylation at promoters is expected to alter transcriptional activity [78], and might promote phenotypic differences, possibly leading to transgressive segregation.

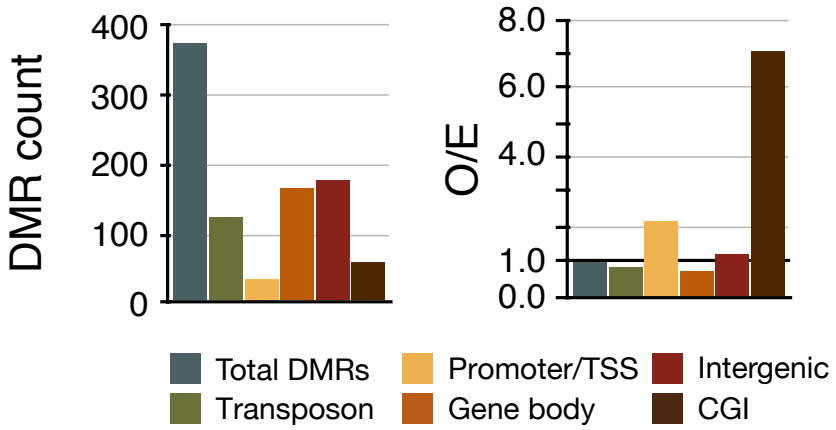


Fig. 4.15 Hybrid-specific methylome variation is enriched at promoter regions and CpG islands. 380 DMRs are predicted between the liver methylomes of the parental taxa and F1 hybrid of brood 3. Enrichment analysis for the genomic localisation of hybrid-specific DMR. Left, absolute DMRs count for each genomic region. Right, observed/expected ratios based on 10 random iterations using hybrid-specific DMR coordinates and features (length).

The liver methylomes of AxA hybrids reveal DNA methylation patterns unique to the F1 offspring (Fig.4.16, left). For example, the gene carbonic anhydrase 15-like (CA-XV

like), participating in systemic acid–base regulation in fish [199, 200], shows a specific hypomethylated state in two exons of the F1 hybrid (Fig. 4.16, left). Strikingly, both the parental taxa exhibit hypermethylation state at this locus, which possibly suggests that this is a hybrid-specific methylome pattern. The enzyme CA-XV like plays an important function in kidney and gills, where most direct exchanges of acid-base equivalents with the environment occur [199]. Therefore, this change in methylation at the gene CA-XV like observed in liver methylome of hybrids might possibly be present in other tissues or in early embryogenesis, where it might exert specific functions.

Another example of DMR unique to cichlid hybrids is found in the CpG-rich region of one intragenic transposon (Fig. 4.17b). The three F1 hybrids, bred from two different mothers and one father, show specific hypomethylation levels compared to both parental taxa, hinting a TE-derived regulatory functions (see Discussion).

Other patterns of methylation inheritance is be also observed. For example, the gene *CD276*, coding for a transmembrane protein involved in cell immune signalling, shows hypomethylation in both the paternal and hybrid specimens specifically compared to the maternal taxon (Fig. 4.16, right). This DMR spans two exons and one exon-intron junction, hinting at a role of DNA methylation in promoting alternative splicing of this gene – characterisation of specific transcript expression via RNA sequencing is required to validate such hypothesis. However, we cannot rule out that this DMR could also be sex-specific, as the hypomethylated state is only observed in the paternal and male hybrid specimens. Another example of DMR putatively inherited from the paternal line is found in one of the intron (close to the exon-intro junction) of the gene *prkag2* (Fig. 4.17a), encoding for the enzyme 5'-AMP-activated protein kinase subunit gamma-2, involved in lipogenesis regulation (in particular fatty acid and cholesterol). F1 hybrids of both sexes show the same hypermethylated states as the paternal specimens compared to both maternal specimens, hinting at the transmission of epigenetic states in a parent-of-origin manner. Reciprocal cross experiment (using the same parental taxa in a reverse fashion) would be required to distinguish methylation patterns transmitted in a species-specific manner from the ones inherited from one specific parent.

Finally, the genomic region directly upstream of one DMR appears to be missing in the paternal genome (Fig. 4.16 right; lack of read coverage), suggesting a possible genomic insertion or deletion event in the paternal line. In addition, the maternal methylation patterns at this regions might not be inherited in the offspring. Similar observations at other genomic regions have been made using this inter-species cross and are been characterised at the moment. Such genomic regions are being compared with whole-genome data generated for

many individuals of the same species, in order to validate any insertion/deletion events. The methylation states at such regions will be compared between the parental individuals and F1 hybrids.

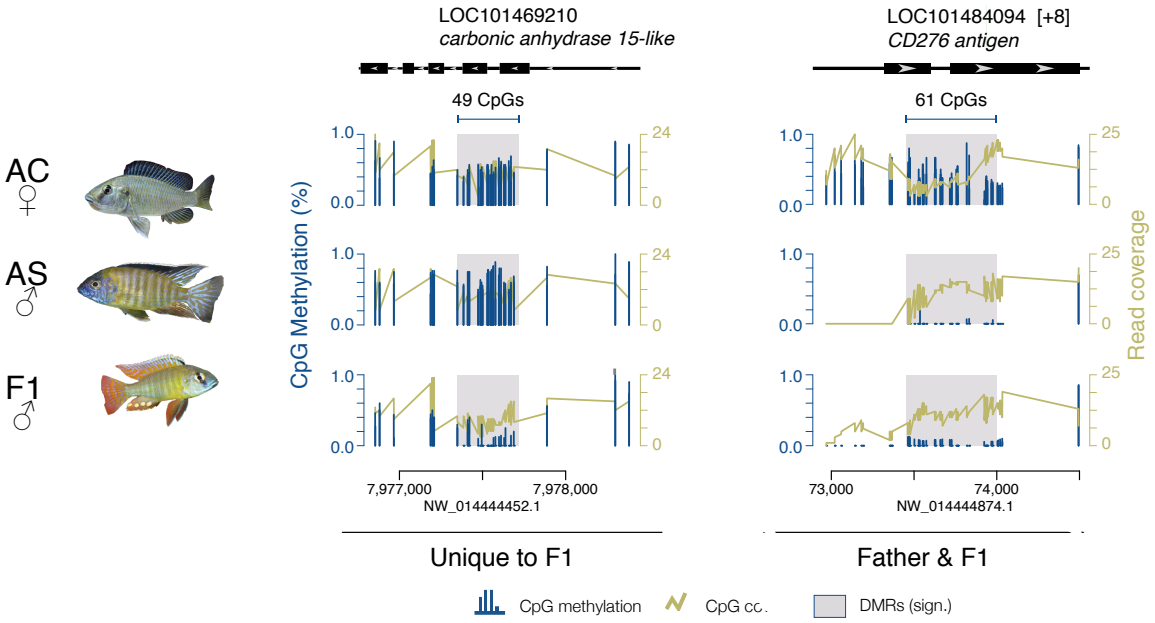


Fig. 4.16 Examples of DMRs in hybrids and parental taxa. Example of two DMRs found in hybrids and parental taxa. Left: the gene *CA-XV-like* show hypermethylation state (DMR: 54 CpGs, 541bp-long, 41%) in two exons, and is specific to the F1 hybrid. Both parental taxa show hypermethylated state at that locus. Right: Hypomethylated DMR spanning two exons and one intro region (DMR: 61 CpG, 464bp-long, 41.5%) of the gene *CD276 antigen* observed in the maternal and hybrid methylomes only. Paternal liver methylome shows hypermethylation at that locus.

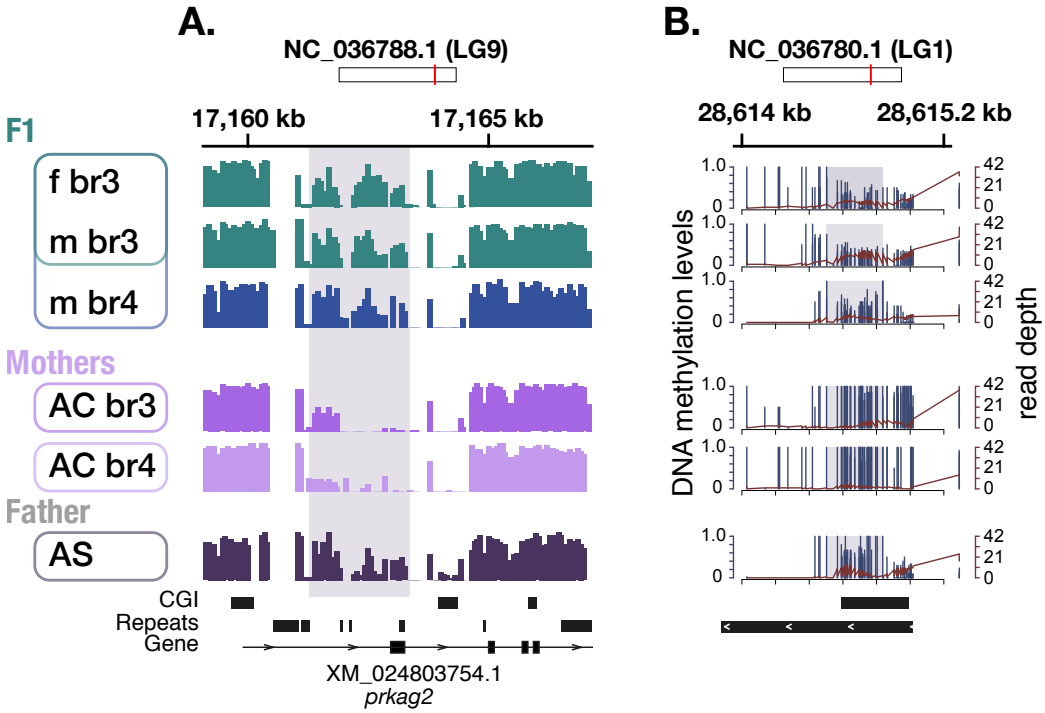


Fig. 4.17 Examples of DMRs in hybrids associated with metabolic genes and intragenic TE sequences. A. The gene body (one intron and one exon) of the gene *prkag2* shows significantly higher methylation levels in all the three AxA F1 hybrids and in paternal line (AS) compared to both maternal (AC) specimens. The gene *prkag2* encodes for the enzyme 5'-AMP-activated protein kinase subunit gamma-2 and is involved in regulating lipogenesis (in particular fatty acid and cholesterol). **B.** An example of DMR located in a intragenic CpG-rich TE sequence. The three hybrid specimens show hypomethylation compared to both parental taxa. CGI, for CpG island.

4.5 Discussion and future work - Cichlid hybrids

In the second part of chapter 4, I characterised the liver methylomes of inter-species hybrids, bred from two different cichlid species (Fig. 4.9). The aim was to investigate the inheritance of DNAme variation in F1 hybrids. From studies carried out in zebrafish, epigenetic reprogramming in teleost fish is thought to be very different from the ones observed in other amniotic vertebrates, in that no global DNA demethylation takes place, instead the sperm methylome is retained in the embryo after fertilisation and the maternal methylome is specifically reprogrammed to resemble the paternal configuration (see Fig. 1.5 and Ref. [14]). This is in stark contrast with mammalian epigenetic reprogramming where DNA methylation is reset during embryogenesis to achieve totipotency. However, it is still unknown whether the unique DNAme reprogramming seen in zebrafish is conserved in other teleost fish.

4.5.1 Hybrid methylome

The first observation is that overall DNAme levels in fully differentiated liver cells of F1 hybrids were very similar to that of both parental lines as expected (Fig. 4.10 and 4.13). Moreover, hybrid methylomes resembled more the methylomes of both parental specimens, compared to other cross-independent specimens from the same taxa as the parents (Figs. 4.10 and 4.12). This hints at a genetic basis in the transmission of DNAme patterns as parental methylomes seem to be transmitted to F1 offspring along with the parental alleles without major obvious parent-specific bias. The sequence divergence between the two parental genomes is very low (0.2%, or 2 SNP per kbp) and represents the average sequence divergence observed between other Lake Malawi cichlids. At the epigenetic level, parental methylomes (or between the two parental species) at conserved underlying DNA sequences correlate at 76-79% (Fig. 4.12). This correlation is higher between any parental specimens and F1 hybrids (80-87%), in particular between parental and hybrid individuals of the same brood (83-86%). Principal component analysis could further provide evidence for the genetic basis of DNAme patterns, whereby hybrid methylomes resemble more that of parents, compared to other specimens, that are part of the same species as the parents but are not related to the cross.

4.5.2 Patterns of inheritance in hybrids

Although most of the hybrid methylome appear to inherited from the parents, the patterns of DNAme at certain genomic regions hints at different types of epigenetic inheritance.

Earlier in chapter 4, I provided examples of DMRs showing DNA methylation patterns similar in one of the parent taxa and hybrid specimens, suggesting either sex-specific or paternal transmission of DNAme variation. For examples, the genes *CD2676 antigen* (Fig. 4.16, right) and *prkag2* (Fig. 4.17a) show similar methylation levels specifically in hybrids and the paternal line, compared to both maternal individuals. Interestingly, both DMRs span intron-exon junctions, hinting at a role of DNAme in alternative splicing events. However, the sequencing of more F1 hybrid individuals is needed to confirm that these DMRs are actually paternally inherited.

Future work would identify not only species-specific, but also individual-specific DMRs that are transmitted to hybrids over the other taxa.

4.5.3 Transgressive DNAme patterns in hybrids

Strikingly, some patterns of DNA methylation appear to be only observed in hybrid individuals, in that they are very distinct from both parental methylomes (Figs. 4.14, 4.16 left and 4.17b). Such hybrid-specific patterns could be interpreted as being transgressive, in that they were significantly different from both parental lines.

In chapter 4, two examples of putative transgressive DMRs are provided. The first one is found in one CpG-rich regions of an intragenic TE sequence and is hypomethylated in all hybrid individuals compared to parents, irrespective of the species. Interestingly, this hybrid-specific DMR spans a CpG-rich region (also known as CpG islands, CGIs) that are known to exert important regulatory regions, at least in mammals [43, 44]. In mammals, most promoters (ca70%) are associated with CGIs (and are usually unmethylated), in contrast to zebrafish and cichlid where less than 30% of all the promoters have CGIs (see Fig. 2.13). The functions of CGIs localised outside promoter regions (orphan CGIs) are unclear, and could be linked to regulatory functions (such as TE-derived ectopic promoter; see section 1.4.5). It is therefore tempting to speculate about a possible regulatory function associated with this hybrid-specific DMR localised in a TE-derived CGI. To confirm this hypothesis, differential gene expression analysis of neighbouring genes could be assessed. In addition to this, the genomic annotation of cichlid promoters and enhancer should be generated to further identify the functions of any hybrid-specific intragenic DMRs.

The second example of hybrid-specific DMRs is found in the exon-intron junction of the gene *CA-XV* like, coding for the enzyme carbonic anhydrase 15-like, participating in systemic acid–base regulation in fish [199, 200]. Many isoforms have been found for this gene in fish and can be expressed in a tissue-specific and even species-specific manner in

teleost fish [199]. Differentially methylation levels at intron-exon junctions could participate in alternative splicing of this gene. Such a role for DNAme has been characterised previously [86]. The inclusion of one exon part of the gene *CD45* for example has been reported to be regulated by differential methylation levels at the exon-intron junction. This is thought to happen via the interaction of a methyl-sensitive CTCF factor with RNApol II (Pol II), resulting in a possible cell-type-specific expression of different isoforms of the gene *CD45* [87]. Such DNA-mediated mechanism of alternative splicing could exist as well. RNAseq of liver hybrids could further validate this.

Such transgressive DNAme patterns observed in hybrids could be in line with what has been observed in plant hybrids, where F1 hybrids can show transgressive phenotypes (called hybrid vigour in plants) possibly associated with dissimilar DNAme patterns from parental taxa, possibly affecting transcriptional activity [198, 30, 196, 201, 202]. Furthermore, a recent common-garden (lake conditions) and transplant (from river to lake) experiment with Lake Tanganyika *Astatotilapia burtoni* revealed that F1 individuals bred from river and lacustrine parents showed high levels of phenotypic plasticity, resulting in slightly higher fitness (faster growth), compared to either the lacustrine or riverine parents, even competing pure-bred lacustrine specimens reared in their native environment [195]. This suggests that phenotypic plasticity could promote adaptation to new environments and that DNAme variation observed in F1 hybrids could facilitate such adaptive plasticity in certain instances.

Note that, only a few number of hybrid individuals were analysed in the second part of chapter 4 (more are currently being sequenced at the moment). It is therefore impossible to conclude whether the hybrid-specific methylation patterns observed are unique to F1 hybrid individuals or to this particular inter-breeding setup. They could could also arise from stochastic epigenetic variation and be linked to particular developmental times. In addition, it would be important to generate the reciprocal cross in order to characterise the specificity of novel, hybrid-specific methylome patterns. Unfortunately, the reciprocal cross has remained unsuccessful.

Current work focus on generating WGBS data for more F1 hybrids of the same crosses, to understand how frequent hybrid-specific or parent-specific DMRs are observed in particular this cross setup. Gene expression analysis (RNAseq) using F1 hybrid liver tissues will be generated to explore the functional link between hybrid-specific DMRs and transcriptional regulation. In addition, the recent technical advances in genome editing [203] might represent a tool in order to precisely and locally manipulate the DNA methylation state at particular loci in the embryo (using modified CRISPR-dCas9 system to bring epigenetic erasers and modifiers to a particular genomic locus). This might provide a way to expand our knowledge

on the heritability of modified epigenetic patterns in subsequent generations. Considerable efforts are being made at the moment to develop such technique in cichlids in Cambridge, UK in collaboration with the laboratory of Emília Santos.

4.6 Detailed methodology

4.6.1 Common Garden and AA hybrids - 'husbandry'

All tank-reared cichlid individuals were reared and bred in laboratory-condition tanks, in the same room (same light/dark cycles), fed on the same standard aquarium fish flake food.

Wild specimens of Lake Massoko

Wild specimens from Lake Massoko were sampled by Prof. Martin Genner, Prof G. Turner, Alan Hudson and Alexandra M. Tyers. Liver tissues were preserved in RNA*later* before being stored in -80ºC. Refer to section 3.6.1 and ref [181] for detailed methodology. Live fish were collected in November 2011 by a team of professional aquarium fish collectors, under the supervision of Profs. George Turner and Martin Genner, for shipment to UK. These were collected from depths of >20m or < 5m to ensure good representation of both ecomorphs. Deep-water fish were decompressed overnight in keep-nets at depths of 5-10m.

Liver methylome - tissue extraction, NGS library preparation and methylome analysis

In order to generate the liver methylome for both G1/F1 specimens and Tank-Itupi control individuals, the same method as the one described in Chapter 2 (see section 2.10.2) was used. The only difference is that immediately after dissection tank-reared tissues were snap-frozen (dry ice) and not preserved, like for wild individuals, in RNA*later*. In brief, liver tissues of tank-reared specimens were snap-frozen upon dissection. WGBS data were generated according to the same protocol as the one used in Chapter 2. Paired-end reads, 150bp-long, were sequenced on HiSeq4000. All sequencing reads were then mapped to the same reference genome (*M. zebra*, UMD2a), and only conserved CG dinucleotide sequences between all samples showing high sequencing coverage were analysed (coverage of 5-100 non-clonal paired-end reads).

Chapter 5

Conclusion

5.1 Conclusive remarks and perspectives

In this thesis, I first investigated the epigenetic basis of adaptive phenotypic plasticity observed in cichlids of Lake Malawi. To address this aim, I generated and analysed whole-genome bisulfite and RNA sequencing data from two different homogenous tissues, namely liver and muscle, of five different cichlid species, covering four distinct ecological clades of Lake Malawi. In parallel to this work, I generated WGBS data for the two ecomorph populations of *A. calliptera* sp. Massoko in early stages of speciation as well as for a riverine *A. calliptera*, putatively closely related to the ancestral *Astatotilapia* of Lake Massoko. This work represents the first study investigating the epigenetic basis for species diversification and adaptation in natural populations of East African cichlids, and provides some initial evidence for a role of DNA methylation in rewiring the transcriptional network in a context of trophic adaptation.

In chapter 2, I described the main characteristics of the methylome of Lake Malawi cichlid *Maylandia zebra*, many of which are shared with other vertebrates. In brief, **(i)** the genome of Lake Malawi cichlids is highly methylated at CG dinucleotide sequences; **(ii)** transposable elements are in particular highly methylated (more than the genome-wide methylation average), hinting at a DNAme-based mechanism of TE transcriptional silencing; **(iii)** methylation at promoters is negatively correlated with transcription activity, while being more weakly, but positively correlated in gene bodies, which hints at important regulatory functions of DNAme at such regions. Then, I characterised the considerable variation at conserved underlying DNA sequences of the liver methylomes between five eco-morphologically different cichlid species. Species phylogeny based solely on methylome

variation in natural populations of cichlids could be reconstructed, possibly even reflecting diet adaptation. Furthermore, I revealed that the variation in DNA methylation associated with significant changes in transcription activity was mainly located in the promoters and in intragenic TE sequences of genes involved in lipogenesis, hormone signalling, hepatic metabolic processes and, importantly, in development. Altogether, these results postulate an important role for DNA methylation at genomic regulatory regions in promoting phenotypic plasticity associated with adaptive traits related to development and liver functions.

In chapter 3, together with Prof. Martin Genner and colleagues I generated WGBS data from liver tissues of two *A. calliptera* populations of Lake Massoko in early stages of speciation and showing distinct trophic and morphological adaptation. Liver methylome of the two ecomorphs exhibited distinct, ecomorph-specific variation. In brief, the benthic population of *A. calliptera* sp. Lake Massoko showed widespread hypermethylated levels, in particular at promoter regions, compared to a riverine *A. calliptera* sp. Mbaka population, closely related to the ancestral *Astatotilapia* of Lake Massoko. Genes associated with important liver metabolic functions, as well as key developmental genes, displayed increased levels of methylation. I therefore postulated that DNA methylation could underlie some selectable phenotypic plasticity, which could facilitate invasion and adaptation of the benthic population to the darker, zooplankton-rich parts of the lake. Furthermore, I showed that cichlids of Lake Malawi and Lake Massoko shared some variation in liver methylomes, almost exclusively at TE-derived sequences and promoters. This implies that similar biological processes might be subject to epigenetic variability in both lake systems and might participate in the short- and long-term divergence of adaptive traits.

In chapter 4, I first analysed the dynamics of liver methylome in response to environmental perturbation, and then characterised the inheritance of DNAme patterns in cichlid hybrids. To quantify the extent to which the environment may shape liver methylome, I generated WGBS data from livers of tank-reared (common garden experiment) benthic, littoral (both from Massoko) and riverine ancestral *Astatotilapia* specimens. Surprisingly, most of the variation in liver methylome associated with the invasion and colonisation of Lake Massoko was rest upon domestication to resemble the ancestral-like, riverine liver methylome. Specifically, the majority of the hypermethylated regions observed in the wild individuals showed drastic reduction in methylation in the tank-reared specimens. Inversely, the few hypomethylated regions in the wild individuals compared to tank specimens were re-methylated in the tank-read specimens, just like in the ancestral-like, riverine population. This suggests that upon common-garden experiment, most of the methylome variation associated with adaptation to

the deep ecological habitats of Lake Massoko could have been erased, although my analysis could not rule out any methylation patterns retained in both the wild and tank-reared benthic and littoral populations, which could represent fixed heritable epigenetic divergence between the two populations. In the second part of chapter 4, I described the inheritance of DNAme in F1 cichlid hybrids. I provided some evidence for a potential genetic basis of the inheritance of DNA methylation, in that hybrid methylomes were very similar to those of both parents. However, at certain loci, some different patterns of inheritance were observed: in some instances, the paternal methylome only appeared to be transmitted to F1 hybrids, while in other cases hybrid cichlids showed transgressive methylome patterns, in that they were dissimilar to these of both parental taxa. I postulate that such transgressive patterns could participate in phenotypic divergence, potentially leading to transgressive segregation and adaptation. The results presented in chapter 4 are very preliminary, which warrants further experimental work. Nevertheless they provided initial evidence for a highly dynamic DNA methylation remodelling in response to environmental cues, as well as for some levels of transgressive segregation in cichlid hybrids, potentially generating selectable phenotypic divergence facilitating speciation.

Altogether, the findings presented in this thesis provide the first line of evidence that methylome variation might underlie selectable phenotypic plasticity in response to the environment, potentially rewiring the transcriptional landscape. Such mechanisms might eventually lead to phenotypic divergence and adaptation in natural populations of cichlids. Importantly, this work has set the basis for future projects investigating further the role of DNAme variation in underlying selectable phenotypic divergence in East African cichlid fishes.

One of the major findings of this work is that at conserved underlying DNA sequences liver methylomes exhibit high levels of species- and possibly diet-specificity, suggesting that some adaptative traits could be mediated by DNA methylation variation. I revealed as well that some of this variation was localised in key developmental genes, possibly underlying long-lasting phenotypic divergence between cichlid species. Another important finding of this research is that some methylome variation is shared between cichlids of Lake Malawi and Massoko, suggesting that similar biological processes might be subject to epigenetic variation. This sort of *standing epigenetic variation* could underlie important phenotypic plasticity associated with adaptive traits essential for the colonisation of similar ecological niches.

Future work will be required to validate and characterise further the main findings of this work, and have been described in detail in each result chapter.

In brief, future experimental projects could aim at **(i)** characterising the importance of DNAme variation during development, in particular in inter-species hybrids. Such an experiment would allow to investigate when specific-specific DNAme patterns are established as well as to study their potential link with transcription changes; **(ii)** generating RNAseq of Lake Massoko cichlids to investigate whether adaption to similar ecological niches and diets (e.g. dimly lit, zooplankton-rich habitats) is correlated with similar methylome patterns and transcriptional activity in both lake systems; **(iii)** fully characterising the inheritance and reprogramming of DNAme patterns at different developmental times, and to draw a comparison with the unique reprogramming processes seen in zebrafish . This could be of particular interest in hybrid cross, in order to elucidate the extent to which DNAme patterns are erased and re-established in the offspring and to assess any biases in the inheritance of patterns of methylation (paternal vs maternal transmission). Ideally, WGBS data from sperm, egg and FAC-sorted (fluorescence activated cell sorting) cells of the embryo (somatic vs PGCs) should be generated; **(iv)** characterising further and with more biological replicates the unique transgressive methylome patterns observed in inter-species hybrids, to eventually correlate such transgressive variation with differential transcription activity in hybrids; **(v)** using a more targeted approach to understand the functional importance of some inter-specific DMRs. For example, the methylation levels at one candidate gene, as well at the associated gene expression activity, could be determined in larger populations of fish presenting similar diet adaptation, or thriving in the same habitat but feeding on different diets (pyrosequencing, qPCR). In addition, one could set up a live fish experiment where specific DMRs and gene expression could be monitored upon specific environmental perturbations (light condition, e.g.) - for example, the gene *vsx1* is observed to be highly methylated in fish having colonised deep, dimly lit parts of the lake. In a common garden experiment set up, the dynamics of DNAme at the promoter of this gene could be monitored, to understand whether epigenetic changes could allow for phenotypic plasticity ; **(vi)** validating the widespread DNAme reprogramming observed upon environmental perturbation in a natural context. An experiment could be designed to understand whether such mechanisms can be seen in a wild population of Lake Massoko cichlids artificially placed in their putative ancestral environment, namely the river nearby (Mbaka); **(vii)** characterising species-divergent and environmental-independent DMRs between the benthic and littoral populations of Lake Massoko to elucidate whether some DNAme variation may be fixed in a population, possibly underlying short-term phenotypic diversification; **(viii)** generating the genomic annotations of regulatory regions

(promoter, enhancers) in cichlid, as a considerable amount of variable DNAme has been predicted to be located in such regions. One could perform ChIP-sequencing (enhancers and promoters; chromatin immunoprecipitation followed by sequencing) and chromosome conformation capture (HiC; especially relevant to identify promoter-enhancer interactions).

References

[1] Giacomo Cavalli and Edith Heard. Advances in epigenetics link genetics to the environment and disease. *Nature*, 571(7766):489–499, jul 2019.

[2] Leandro Quadrana, Amanda Bortolini Silveira, George F. Mayhew, Chantal LeBlanc, Robert A. Martienssen, Jeffrey A. Jeddeloh, and Vincent Colot. The Arabidopsis thaliana mobilome and its impact at the species level. *eLife*, 5(JUN2016):1–25, 2016.

[3] Ksenia Skvortsova, Katsiaryna Tarbashevich, Martin Stehling, Ryan Lister, Manuel Irimia, Erez Raz, and Ozren Bogdanovic. Retention of paternal DNA methylome in the developing zebrafish germline. *Nature Communications*, 10(1):3054, 2019.

[4] Conway Zirkle. The Inheritance of Acquired Characters and the Provisional Hypothesis of Pangenesis. *American Naturalist*, 69:417–445, 1935.

[5] Jean-Baptiste de Lamarck. *Philosophie zoologique ou exposition des considérations relatives à l'histoire naturelle des animaux*. Dentu, Musée d'Histoire Naturelle (Jardin des Plantes), 1809.

[6] Charles Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London, 1859.

[7] Edith Heard and Robert A. Martienssen. Transgenerational epigenetic inheritance: myths and mechanisms. *Cell*, 157(1):95–109, mar 2014.

[8] Kevin Laland, Tobias Uller, Marc Feldman, Kim Sterelny, Gerd B. Müller, Armin Moczek, Eva Jablonka, John Odling-Smee, Gregory A. Wray, Hopi E. Hoekstra, Douglas J. Futuyma, Richard E. Lenski, Trudy F. C. Mackay, Dolph Schluter, and Joan E. Strassmann. Does evolutionary theory need a rethink? *Nature*, 514(7521):161–164, oct 2014.

[9] Eva Jablonka and Marion J. Lamb. The inheritance of acquired epigenetic variations. *Journal of Theoretical Biology*, 139(1):69–83, jul 1989.

[10] Eric A. Miska and Anne C. Ferguson-Smith. Transgenerational inheritance: Models and mechanisms of non-DNA sequence-based inheritance. *Science*, 354(6308):59–63, oct 2016.

[11] Mary Jane West-Eberhard. Developmental plasticity and the origin of species differences. *Proceedings of the National Academy of Sciences*, 102(Supplement 1):6543–6549, may 2005.

[12] Vincenzo E. A. Russo, Robert A. Martienssen, and Arthur D Riggs. *Epigenetic Mechanisms of Gene Regulation*. Cold Spring Harbor Laboratory Press, Woodbury, russo, vin edition, 1996.

[13] Adrian Bird. Perceptions of epigenetics. *Nature*, 447(7143):396–398, may 2007.

[14] Ksenia Skvortsova, Nicola Iovino, and Ozren Bogdanović. Functions and mechanisms of epigenetic inheritance in animals. *Nature Reviews Molecular Cell Biology*, 19(12):774–790, dec 2018.

[15] Ksenia Skvortsova, Nicola Iovino, and Ozren Bogdanović. Functions and mechanisms of epigenetic inheritance in animals. *Nature Reviews Molecular Cell Biology*, 19(12):774–790, dec 2018.

[16] Johannes Bohacek and Isabelle M Mansuy. Molecular insights into transgenerational non-genetic inheritance of acquired behaviours. *Nature Reviews Genetics*, 16(11):641–652, nov 2015.

[17] Maxim V. C. Greenberg and Deborah Bourc'his. The diverse roles of DNA methylation in mammalian development and disease. *Nature Reviews Molecular Cell Biology*, 20(10):590–607, oct 2019.

[18] Zachary D Smith and Alexander Meissner. DNA methylation: roles in mammalian development. *Nature Reviews Genetics*, 14(3):204–220, mar 2013.

[19] Assaf Zemach and Daniel Zilberman. Evolution of Eukaryotic DNA Methylation and the Pursuit of Safer Sex. *Current Biology*, 20(17):R780–R785, sep 2010.

[20] Özgen Deniz, Jennifer M. Frost, and Miguel R Branco. Regulation of transposable elements by DNA modifications. *Nature Reviews Genetics*, 20(7):417–431, jul 2019.

[21] Adrian Bird. Epigentic Memory. *Genes and Development*, 16:16–21, 2002.

[22] Vicky W. Zhou, Alon Goren, and Bradley E. Bernstein. Charting histone modifications and the functional organization of mammalian genomes. *Nature Reviews Genetics*, 12(1):7–18, 2011.

[23] Craig L Peterson and Marc-André Laniel. Histones and histone modifications. *Current Biology*, 14(14):R546–R551, jul 2004.

[24] Tony Kouzarides. Chromatin Modifications and Their Function. *Cell*, 128(4):693–705, feb 2007.

[25] Jiamu Du, Lianna M. Johnson, Steven E. Jacobsen, and Dinshaw J. Patel. DNA methylation pathways and their crosstalk with histone methylation. *Nature Reviews Molecular Cell Biology*, 16(9):519–532, 2015.

[26] Kevin V Morris and John S Mattick. The rise of regulatory RNA. *Nature Reviews Genetics*, 15(6):423–437, jun 2014.

[27] Rafael Galupa and Edith Heard. X-Chromosome Inactivation: A Crossroads Between Chromosome Architecture and Gene Regulation. *Annual Review of Genetics*, 52(1):535–566, nov 2018.

[28] Julie A. Law and Steven E. Jacobsen. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews Genetics*, 11(3):204–220, mar 2010.

[29] Katharina Gapp and Eric A. Miska. tRNA fragments: novel players in intergenerational inheritance. *Cell Research*, 26(4):395–396, apr 2016.

[30] Donna M. Bond and David C. Baulcombe. Small RNAs and heritable epigenetic variation in plants. *Trends in Cell Biology*, 24(2):100–107, feb 2014.

[31] John B. Gurdon. The developmental capacity of nuclei taken from differentiating endoderm cells of Xenopus laevis. *Journal of embryology and experimental morphology*, 8(December):505–26, dec 1960.

[32] Eva Hörmanseder, Angela Simeone, George E. Allen, Charles R. Bradshaw, Magdalena Figlmüller, John Gurdon, and Jerome Jullien. H3K4 Methylation-Dependent Memory of Somatic Cell Identity Inhibits Reprogramming and Development of Nuclear Transfer Embryos. *Cell Stem Cell*, 21(1):135–143.e6, 2017.

[33] Deepak Srivastava and Natalie DeWitt. In Vivo Cellular Reprogramming: The Next Generation. *Cell*, 166(6):1386–1396, sep 2016.

[34] Timothy H. Bestor. DNA Methylation: Evolution of a Bacterial Immune Function into a Regulator of Gene Expression and Genome Structure in Higher Eukaryotes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 326(1235):179–187, jan 1990.

[35] Assaf Zemach, Ivy E. McDaniel, Pedro Silva, and Daniel Zilberman. Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation. *Science*, 328(5980):916–919, may 2010.

[36] Nina V Fedoroff. Transposable Elements, Epigenetics, and Genome Evolution. *Science*, 338(6108):758–767, nov 2012.

[37] Timothy H. Bestor. Cytosine methylation mediates sexual conflict. *Trends in Genetics*, 19(4):185–190, apr 2003.

[38] Diwash Jangam, Cédric Feschotte, and Esther Betrán. Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. *Trends in Genetics*, 33(11):817–831, 2017.

[39] Günter Raddatz, Paloma M. Guzzardo, Nelly Olova, Marcelo Rosado Fantappié, Markus Rampp, Matthias Schaefer, Wolf Reik, Gregory J. Hannon, and Frank Lyko. Dnmt2-dependent methylomes lack defined DNA methylation patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 110(21):8627–8631, 2013.

[40] Silvana Rošić, Rachel Amouroux, Cristina E. Requena, Ana Gomes, Max Emperle, Toni Beltran, Jayant K. Rane, Sarah Linnett, Murray E. Selkirk, Philipp H. Schiffer, Allison J. Bancroft, Richard K. Grencis, Albert Jeltsch, Petra Hajkova, and Peter Sarkies. Evolutionary analysis indicates that DNA alkylation damage is a byproduct

of cytosine DNA methyltransferase activity. *Nature Genetics*, 50(3):452–459, mar 2018.

[41] Suhua Feng, Shawn J. Cokus, Xiaoyu Zhang, Pao-Yang Chen, Magnolia Bostick, Mary G. Goll, Jonathan Hetzel, Jayati Jain, Steven H. Strauss, Marnie E. Halpern, Chinweike Ukomadu, Kirsten C. Sadler, Sriharsa Pradhan, Matteo Pellegrini, and Steven E. Jacobsen. Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences*, 107(19):8689–8694, may 2010.

[42] Yupeng He and Joseph R. Ecker. Non-CG Methylation in the Human Genome. *Annual Review of Genomics and Human Genetics*, 16(1):55–77, aug 2015.

[43] Hao Wu, Tianlei Xu, Hao Feng, Li Chen, Ben Li, Bing Yao, Zhaohui Qin, Peng Jin, and Karen N. Conneely. Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Research*, 43(21):gkv715, jul 2015.

[44] Aimée M. Deaton and Adrian Bird. CpG islands and the regulation of transcription. *Genes & development*, 25(10):1010–22, may 2011.

[45] Adrian Bird. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research*, 8(7):1499–1504, 1980.

[46] Robin Holliday and Geoffrey W. Grigg. DNA methylation and mutation. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 285(1):61–67, jan 1993.

[47] Vassili Kusmartsev and Tobias Warnecke. Cytosine methylation affects the mutability of neighbouring nucleotides in human, Arabidopsis, and rice. *bioR*, ((doi.org/10.1101/764753.)), 2019.

[48] Rahul M. Kohli and Yi Zhang. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature*, 502(7472):472–479, 2013.

[49] R Holliday and J. Pugh. DNA modification mechanisms and gene activity during development. *Science*, 187(4173):226–232, jan 1975.

[50] Arthur D. Riggs. X inactivation, differentiation, and DNA methylation. *Cytogenetic and Genome Research*, 14(1):9–25, 1975.

[51] Zachary D. Smith, Michelle M. Chan, Kathryn C. Humm, Rahul Karnik, Shila Mekhoubad, Aviv Regev, Kevin Eggan, and Alexander Meissner. DNA methylation dynamics of the human preimplantation embryo. *Nature*, 511(7511):611–615, jul 2014.

[52] Stella M. K. Glasauer and Stephan C. F. Neuhauss. Whole-genome duplication in teleost fishes and its evolutionary consequences. *Molecular Genetics and Genomics*, 289(6):1045–1060, dec 2014.

[53] David Brawand, Catherine E. Wagner, Yang I. Li, Milan Malinsky, Irene Keller, Shaohua Fan, Oleg Simakov, Alvin Y. Ng, Zhi Wei Lim, Etienne Bezault, Jason Turner-Maier, Jeremy Johnson, Rosa Alcazar, Hyun Ji Noh, Pamela Russell, Bronwen Aken, Jessica Alföldi, Chris Amemiya, Naoual Azzouzi, Jean-François Baroiller, Frederique Barloy-Hubler, Aaron Berlin, Ryan Bloomquist, Karen L. Carleton, Matthew A. Conte, Helena D'Cotta, Orly Eshel, Leslie Gaffney, Francis Galibert, Hugo F. Gante, Sante Gnerre, Lucie Greuter, Richard Guyon, Natalie S. Haddad, Wilfried Haerty, Rayna M. Harris, Hans A. Hofmann, Thibaut Hourlier, Gideon Hulata, David B. Jaffe, Marcia Lara, Alison P. Lee, Iain MacCallum, Salome Mwaiko, Masato Nikaido, Hidenori Nishihara, Catherine Ozouf-Costaz, David J. Penman, Dariusz Przybylski, Michaelle Rakotomanga, Suzy C.P. P. Renn, Filipe J. Ribeiro, Micha Ron, Walter Salzburger, Luis Sanchez-Pulido, M. Emilia Santos, Steve Searle, Ted Sharpe, Ross Swofford, Frederick J. Tan, Louise Williams, Sarah Young, Shuangye Yin, Norihiro Okada, T. D. Kocher, Eric A. Miska, Eric S. Lander, Byrappa Venkatesh, Russell D. Fernald, Axel Meyer, Chris P. Ponting, J. Todd Streelman, Kerstin Lindblad-Toh, Ole Seehausen, and Federica Di Palma. SUPPL - The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, 513(7518):375–381, sep 2014.

[54] Frank Lyko. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nature Reviews Genetics*, 19(2):81–92, feb 2018.

[55] Tamara H L Smith, Terry Mark Collins, and Ross A. McGowan. Expression of the dnmt3 genes in zebrafish development: Similarity to Dnmt3a and Dnmt3b. *Development Genes and Evolution*, 220(11-12):347–353, 2011.

[56] Pawat Seritrakul and Jeffrey M. Gross. Expression of the de novo DNA methyltransferases (dnmt3 - dnmt8) during zebrafish lens development. *Developmental Dynamics*, 243(2):350–356, 2014.

[57] Nobuyoshi Shimoda, Kimi Yamakoshi, Akimitsu Miyake, and Hiroyuki Takeda. Identification of a gene required for de novo DNA methylation of the zebrafish no tail gene. *Developmental Dynamics*, 233(4):1509–1516, 2005.

[58] Catarina Campos, Luisa M.P. Valente, and Jorge M.O. Fernandes. Molecular evolution of zebrafish dnmt3 genes and thermal plasticity of their expression during embryonic development. *Gene*, 500(1):93–100, may 2012.

[59] Julia Arand, David Spieler, Tommy Karius, Miguel R Branco, Daniela Meilinger, Alexander Meissner, Thomas Jenuwein, Guoliang Xu, Heinrich Leonhardt, Verena Wolf, and Jörn Walter. In Vivo Control of CpG and Non-CpG DNA Methylation by DNA Methyltransferases. *PLoS Genetics*, 8(6):e1002750, jun 2012.

[60] Kunal Rai, L. D. Nadauld, S. Chidester, E. J. Manos, S. R. James, A. R. Karpf, Bradley R. Cairns, and D. A. Jones. Zebra Fish Dnmt1 and Suv39h1 Regulate Organ-Specific Terminal Differentiation during Development. *Molecular and Cellular Biology*, 26(19):7077–7085, oct 2006.

[61] Carol Best, Heather Ikert, Daniel J. Kostyniuk, Paul M. Craig, Laia Navarro-Martin, Lucie Marandel, and Jan A. Mennigen. Epigenetics in teleost fish: From molecular mechanisms to physiological phenotypes. *Comparative Biochemistry and Physiology*

Part B: Biochemistry and Molecular Biology, 224(September 2017):210–244, oct 2018.

[62] Joan Barau, Aurélie Teissandier, Natasha Zamudio, Stéphanie Roy, Valérie Nalesso, Yann Hérault, Florian Guillou, and Deborah Bourc'his. The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. *Science*, 354(6314):909–912, nov 2016.

[63] Kunal Rai, Itrat F. Jafri, Stephanie Chidester, Smitha R. James, Adam R. Karpf, Bradley R. Cairns, and David A. Jones. Dnmt3 and G9a Cooperate for Tissue-specific Development in Zebrafish. *Journal of Biological Chemistry*, 285(6):4110–4121, feb 2010.

[64] Rudolf Jaenisch and Adrian Bird. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33(S3):245–254, mar 2003.

[65] Xiaoji Wu and Yi Zhang. TET-mediated active DNA demethylation: Mechanism, function and beyond. *Nature Reviews Genetics*, 18(9):517–534, 2017.

[66] Miguel R Branco, Gabriella Ficz, and Wolf Reik. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nature Reviews Genetics*, 13(1):7–13, jan 2012.

[67] Zachary D. Smith, Michelle M. Chan, Tarjei S. Mikkelsen, Hongcang Gu, Andreas Gnirke, Aviv Regev, and Alexander Meissner. A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature*, 484(7394):339–344, 2012.

[68] Joel Hrit, Leeanne Goodrich, Cheng Li, Bang-An Wang, Ji Nie, Xiaolong Cui, Elizabeth Allene Martin, Eric Simental, Jenna Fernandez, Monica Yun Liu, Joseph R Nery, Rosa Castanon, Rahul M Kohli, Natalia Tretyakova, Chuan He, Joseph R Ecker, Mary Goll, and Barbara Panning. OGT binds a conserved C-terminal domain of TET1 to regulate TET1 activity and function in development. *eLife*, 7(c):2–6, oct 2018.

[69] Zebulon G. Levine and Suzanne Walker. The Biochemistry of O-GlcNAc Transferase: Which Functions Make It Essential in Mammalian Cells? *Annual Review of Biochemistry*, 85(1):631–657, jun 2016.

[70] Luigi Aloia, Mikel Alexander McKie, Grégoire Vernaz, Lucía Cordero-Espinoza, Niya Aleksieva, Jelle van den Ameele, Francesco Antonica, Berta Font-Cunill, Alexander Raven, Riccardo Aiese Cigliano, German Belenguer, Richard L. Mort, Andrea H. Brand, Magdalena Zernicka-Goetz, Stuart J. Forbes, Eric A. Miska, and Meritxell Huch. Epigenetic remodelling licences adult cholangiocytes for organoid formation and liver regeneration. *Nature Cell Biology*, 21(11):1321–1333, nov 2019.

[71] Magdalena E. Potok, David A. Nix, Timothy J. Parnell, and Bradley R. Cairns. Reprogramming the maternal zebrafish genome after fertilization to match the paternal methylation pattern. *Cell*, 153(4):759–772, 2013.

[72] Rimple D. Almeida, Matthew Loose, Virginie Sottile, Elena Matsa, Chris Denning, Lorraine Young, Andrew D. Johnson, Martin Gering, and Alexey Ruzov. 5-hydroxymethyl-cytosine enrichment of non-committed cells is not a universal feature of vertebrate development. *Epigenetics*, 7(4):383–389, apr 2012.

[73] Lan Jiang, Jing Zhang, Jing-Jing Wang, Lu Wang, Li Zhang, Guoqiang Li, Xiaodan Yang, Xin Ma, Xin Sun, Jun Cai, Jun Zhang, Xingxu Huang, Miao Yu, Xuegeng Wang, Feng Liu, Chung-I Wu, Chuan He, Bo Zhang, Weimin Ci, and Jiang Liu. Sperm, but not oocyte, DNA methylome is inherited by zebrafish early embryos. *Cell*, 153(4):773–84, may 2013.

[74] Ozren Bogdanović, Arne H. Smits, Elisa de la Calle Mustienes, Juan J. Tena, Ethan Ford, Ruth Williams, Upeka Senanayake, Matthew D. Schultz, Saartje Hontelez, Ila van Kruijsbergen, Teresa Rayon, Felix Gnerlich, Thomas Carell, Gert Jan C. Veenstra, Miguel Manzanares, Tatjana Sauka-Spengler, Joseph R. Ecker, Michiel Vermeulen, José Luis Gómez-Skarmeta, and Ryan Lister. Active DNA demethylation at enhancers during the vertebrate phylotypic period. *Nature Genetics*, 48(4):417–426, apr 2016.

[75] Cheng Li, Yahui Lan, Lianna Schwartz-Orbach, Evgenia Korol, Mamta Tahiliani, Todd Evans, and Mary G. Goll. Overlapping Requirements for Tet2 and Tet3 in Normal Development and Hematopoietic Stem Cell Emergence. *Cell Reports*, 12(7):1133–1143, aug 2015.

[76] Yan Li, Suresh Kumar, and Weiqiang Qian. Active DNA demethylation: mechanism and role in plant development. *Plant Cell Reports*, 37(1):77–85, jan 2018.

[77] Yimeng Yin, Ekaterina Morgunova, Arttu Jolma, Eevi Kaasinen, Biswajyoti Sahu, Syed Khund-Sayeed, Pratyush K. Das, Teemu Kivioja, Kashyap Dave, Fan Zhong, Kazuhiro R. Nitta, Minna Taipale, Alexander Popov, Paul A. Ginno, Silvia Domcke, Jian Yan, Dirk Schübeler, Charles Vinson, and Jussi Taipale. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*, 356(6337):eaaj2239, may 2017.

[78] Dirk Schübeler. Function and information content of DNA methylation. *Nature*, 517(7534):321–326, 2015.

[79] Ronan C. O'Malley, Shao-shan Carol Huang, Liang Song, Mathew G. Lewsey, Anna Bartlett, Joseph R. Nery, Mary Galli, Andrea Gallavotti, and Joseph R. Ecker. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell*, 165(5):1280–1292, may 2016.

[80] Heng Zhu, Guohua Wang, and Jiang Qian. Transcription factors as readers and effectors of DNA methylation. *Nature Reviews Genetics*, 17(9):551–565, 2016.

[81] Bart Deplancke, Daniel Alpern, and Vincent Gardeux. The Genetics of Transcription Factor DNA Binding Variation. *Cell*, 166(3):538–554, jul 2016.

[82] Nicholas E. Banovich, Xun Lan, Graham McVicker, Bryce van de Geijn, Jacob F. Degner, John D. Blischak, Julien Roux, Jonathan K. Pritchard, and Yoav Gilad. Methylation QTLs Are Associated with Coordinated Changes in Transcription Factor Binding, Histone Modifications, and Gene Expression Levels. *PLoS Genetics*, 10(9):e1004663, sep 2014.

[83] Aaron Taudt, Maria Colomé-Tatché, and Frank Johannes. Genetic sources of population epigenomic variation. *Nature Reviews Genetics*, 17(6):319–332, 2016.

[84] Na Hu, Pablo H. Strobl-Mazzulla, and Marianne E. Bronner. Epigenetic regulation in neural crest development. *Developmental Biology*, 396(2):159–168, dec 2014.

[85] William a Pastor, L Aravind, and Anjana Rao. TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nature Reviews Molecular Cell Biology*, 14(6):341–356, jun 2013.

[86] Galit Lev Maor, Ahuvi Yearim, and Gil Ast. The alternative role of DNA methylation in splicing regulation. *Trends in Genetics*, 31(5):274–280, may 2015.

[87] Sanjeev Shukla, Ersen Kavak, Melissa Gregory, Masahiko Imashimizu, Bojan Shutinoski, Mikhail Kashlev, Philipp Oberdoerffer, Rickard Sandberg, and Shalini Oberdoerffer. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, 479(7371):74–79, nov 2011.

[88] Valter Tucci, Anthony R. Isles, Gavin Kelsey, Anne C. Ferguson-Smith, Marisa S. Bartolomei, Nissim Benvenisty, Deborah Bourc'his, Marika Charalambous, Catherine Dulac, Robert Feil, Juliane Glaser, Lisa Huelsmann, Rosalind M. John, Gráinne I. McNamara, Kim Moorwood, Francoise Muscatelli, Hiroyuki Sasaki, Beverly I. Strassmann, Claudius Vincenz, and Jon Wilkins. Genomic Imprinting and Physiological Processes in Mammals. *Cell*, 176(5):952–965, feb 2019.

[89] Xiaowu Chen, Zhipeng Wang, Shoujie Tang, Yan Zhao, and Jinliang Zhao. Genome-wide mapping of DNA methylation in Nile Tilapia. *Hydrobiologia*, 791(1):247–257, may 2017.

[90] Austin T Hilliard, Dan Xie, Zhihai Ma, Michael P Snyder, and Russell D Fernald. Genome-wide effects of social status on DNA methylation in the brain of a cichlid fish, Astatotilapia burtoni. *BMC Genomics*, 20(1):699, dec 2019.

[91] Timothy H. Bestor and Deborah Bourc'his. Transposon Silencing and Imprint Establishment in Mammalian Germ Cells. *Cold Spring Harbor Symposia on Quantitative Biology*, 69:381–388, jan 2004.

[92] Diwash Jangam, Cédric Feschotte, and Esther Betrán. Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. *Trends in Genetics*, 33(11):817–831, 2017.

[93] Domitille Chalopin, Magali Naville, Floriane Plard, Delphine Galiana, and Jean-Nicolas Volff. Comparative Analysis of Transposable Elements Highlights Mobilome Diversity and Evolution in Vertebrates. *Genome Biology and Evolution*, 7(2):567–580, feb 2015.

[94] Jonathan F. Wendel, Johann Greilhuber, Jaroslav Doležel, and Ilia J. Leitch. *Plant Genome Diversity Volume 1*. Springer Vienna, Vienna, 2012.

[95] Jeffrey A. Yoder, Colum P. Walsh, and Timothy H. Bestor. Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics*, 13(8):335–340, aug 1997.

[96] Guillaume Bourque, Kathleen H. Burns, Mary Gehring, Vera Gorbunova, Andrei Seluanov, Molly Hammell, Michaël Imbeault, Zsuzsanna Izsvák, Henry L. Levin, Todd S. Macfarlan, Dixie L. Mager, and Cédric Feschotte. Ten things you should know about transposable elements. *Genome Biology*, 19(1):199, dec 2018.

[97] Eva-Maria Weick and Eric A. Miska. piRNAs: from biogenesis to function. *Development*, 141(18):3458–3471, sep 2014.

[98] Deniz M. Ozata, Ildar Gainetdinov, Ansgar Zoch, Dónal O'Carroll, and Phillip D. Zamore. PIWI-interacting RNAs: small RNAs with big functions. *Nature Reviews Genetics*, 20(2):89–108, 2019.

[99] Y.H. Jin, A. Davie, and H. Migaud. Expression pattern of nanos, piwil, dnd, vasa and pum genes during ontogenic development in Nile tilapia Oreochromis niloticus. *Gene*, 688(November 2018):62–70, mar 2019.

[100] Saskia Houwing, Leonie M. Kamminga, Eugene Berezikov, Daniela Cronembold, Angélique Girard, Hans van den Elst, Dmitri V. Filippov, Heiko Blaser, Erez Raz, Cecilia B. Moens, Ronald H.A. Plasterk, Gregory J. Hannon, Bruce W. Draper, and René F. Ketting. A Role for Piwi and piRNAs in Germ Cell Maintenance and Transposon Silencing in Zebrafish. *Cell*, 129(1):69–82, apr 2007.

[101] Saskia Houwing, Leonie M Kamminga, Eugene Berezikov, and Daniela Cronembold. Supplemental Data A Role for Piwi and piRNAs in Germ Cell Maintenance and Transposon Silencing in Zebrafish. 129, 2005.

[102] Milan Malinsky, Hannes Svardal, Alexandra M. Tyers, Eric A. Miska, Martin J. Genner, George F Turner, and Richard Durbin. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nature Ecology & Evolution*, 2(12):1940–1955, dec 2018.

[103] Michaël Imbeault, Pierre-Yves Helleboid, and Didier Trono. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*, 543(7646):550–554, mar 2017.

[104] Gabriela Ecco, Michael Imbeault, and Didier Trono. KRAB zinc finger proteins. *Development*, 144(15):2719–2729, aug 2017.

[105] Rachel L Cosby, Ni-Chen Chang, and Cédric Feschotte. Host–transposon interactions: conflict, cooperation, and cooption. *Genes & Development*, 33(17-18):1098–1116, sep 2019.

[106] Edward B. Chuong, Nels C. Elde, and Cédric Feschotte. Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews Genetics*, 18(2):71–86, feb 2017.

[107] Barbara McClintock. The significance of responses of the genome to challenge. *Science (New York, N.Y.)*, 226(4676):792–801, 1984.

[108] Shengfeng Huang, Xin Tao, Shaochun Yuan, Yuhang Zhang, Peiyi Li, Helen A. Beilinson, Ya Zhang, Wenjuan Yu, Pierre Pontarotti, Hector Escriva, Yann Le Petillon, Xiaolong Liu, Shangwu Chen, David G. Schatz, and Anlong Xu. Discovery of an Active RAG Transposon Illuminates the Origins of V(D)J Recombination. *Cell*, 166(1):102–114, jun 2016.

[109] Elissa D. Pastuzyn, Cameron E. Day, Rachel B. Kearns, Madeleine Kyrke-Smith, Andrew V. Taibi, John McCormick, Nathan Yoder, David M. Belnap, Simon Erlendsson, Dustin R. Morado, John A.G. Briggs, Cédric Feschotte, and Jason D. Shepherd. The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. *Cell*, 172(1-2):275–288.e18, jan 2018.

[110] M. Emília Santos, Ingo Braasch, Nicolas Boileau, Britta S. Meyer, Loïc Sauteur, Astrid Böhne, Heinz-Georg Belting, Markus Affolter, and Walter Salzburger. The evolution of cichlid fish egg-spots is linked with a cis-regulatory change. *Nature Communications*, 5:5149, 2014.

[111] Hugh D. Morgan, Heidi G.E. Sutherland, David I.K. Martin, and Emma Whitelaw. Epigenetic inheritance at the agouti locus in the mouse. *Nature Genetics*, 23(3):314–318, nov 1999.

[112] Pilar Cubas, Coral Vincent, and Enrico Coen. An epigenetic mutation responsible for natural variation in floral symmetry. *Nature*, 401(6749):157–161, sep 1999.

[113] Walfred W.C. Tang, Sabine Dietmann, Naoko Irie, Harry G. Leitch, Vasileios I. Floros, Charles R. Bradshaw, Jamie A. Hackett, Patrick F. Chinnery, and M. Azim Surani. A Unique Gene Regulatory Network Resets the Human Germline Epigenome for Development. *Cell*, 161(6):1453–1467, jun 2015.

[114] Aoi Hosaka, Raku Saito, Kazuya Takashima, Taku Sasaki, Yu Fu, Akira Kawabe, Tasuku Ito, Atsushi Toyoda, Asao Fujiyama, Yoshiaki Tarutani, and Tetsuji Kakutani. Evolution of sequence-specific anti-silencing systems in Arabidopsis. *Nature Communications*, 8(1):2161, dec 2017.

[115] Andrea D. McCue, Saivageethi Nuthikattu, and R. Keith Slotkin. Genome-wide identification of genes regulated in trans by transposable element small interfering RNAs. *RNA Biology*, 10(8):1379–1395, aug 2013.

[116] Arturo Marí-Ordóñez, Antonin Marchais, Mathilde Etcheverry, Antoine Martin, Vincent Colot, and Olivier Voinnet. Reconstructing de novo silencing of an active plant retrotransposon. *Nature Genetics*, 45(9):1029–1039, sep 2013.

[117] Heather J Lee, Timothy a Hore, and Wolf Reik. Reprogramming the Methylome: Erasing Memory and Creating Diversity. *Cell stem cell*, 14(6):710–719, jun 2014.

[118] Ferdinand von Meyenn and Wolf Reik. Forget the Parents: Epigenetic Reprogramming in Human Germ Cells. *Cell*, 161(6):1248–1251, jun 2015.

[119] Melanie A. Eckersley-Maslin, Celia Alda-Catalinas, and Wolf Reik. Dynamics of the epigenetic landscape during the maternal-to-zygotic transition. *Nature Reviews Molecular Cell Biology*, 19(7):436–450, 2018.

[120] Ke Zheng and P. Jeremy Wang. Blockade of Pachytene piRNA Biogenesis Reveals a Novel Requirement for Maintaining Post-Meiotic Germline Genome Integrity. *PLoS Genetics*, 8(11):e1003038, nov 2012.

[121] Patrick J. Murphy, Shan Fu Wu, Cody R. James, Candice L. Wike, and Bradley R. Cairns. Placeholder Nucleosomes Underlie Germline-to-Embryo DNA Methylation Reprogramming. *Cell*, 172(5):993–1006.e13, feb 2018.

[122] Jamie a Hackett and M. Azim Surani. Beyond DNA: programming and inheritance of parental methylomes. *Cell*, 153(4):737–9, may 2013.

[123] Vincent Colot and Jean-Luc Rossignol. Eukaryotic DNA methylation as an evolutionary device. *BioEssays*, 21(5):402–411, may 1999.

[124] S. Cortijo, R. Wardenaar, M. Colome-Tatche, A. Gilly, M. Etcheverry, K. Labadie, E. Caillieux, F. Hospital, J.-M. Aury, P. Wincker, F. Roudier, R. C. Jansen, V. Colot, and F. Johannes. Mapping the Epigenetic Basis of Complex Traits. *Science*, 343(6175):1145–1148, mar 2014.

[125] Taiji Kawakatsu, Shao-shan Carol Huang, Florian Jupe, Eriko Sasaki, Robert J Schmitz, Mark A Urich, Rosa Castanon, Joseph R Nery, Cesar Barragan, Yupeng He, Huaming Chen, Manu Dubin, Cheng-Ruei Lee, Congmao Wang, Felix Bemm, Claude Becker, Ryan O'Neil, Ronan C O'Malley, Danjuma X Quarless, Nicholas J Schork, Detlef Weigel, Magnus Nordborg, Joseph R. Ecker, Carlos Alonso-Blanco, Jorge Andrade, Joy Bergelson, Karsten Borgwardt, Eunyoung Chae, Todd Dezwaan, Wei Ding, Moisés Expósito-Alonso, Ashley Farlow, Joffrey Fitz, Xiangchao Gan, Dominik G Grimm, Angela Hancock, Stefan R Henz, Svante Holm, Matthew Horton, Mike Jarsulic, Randall A Kerstetter, Arthur Korte, Pamela Korte, Christa Lanz, Chen-Ruei Lee, Dazhe Meng, Todd P Michael, Richard Mott, Ni Wayan Muliyati, Thomas Nägele, Matthias Nagler, Viktoria Nizhynska, Polina Novikova, F Xavier Picó, Alexander Platzer, Fernando A Rabanal, Alex Rodriguez, Beth A Rowan, Patrice A A. Salomé, Karl Schmid, Ümit Seren, Felice Gianluca Sperone, Mitchell Sudkamp, Hannes Svardal, Matt M Tanzer, Donald Todd, Samuel L Volchenboum, George Wang, Xi Wang, Wolfram Weckwerth, and Xuefeng Zhou. Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. *Cell*, 166(2):492–505, jul 2016.

[126] Matthias Benoit, Hajk-Georg Drost, Marco Catoni, Quentin Gouil, Sara Lopez-Gomollon, David C. Baulcombe, and Jerzy Paszkowski. Environmental and epigenetic regulation of Rider retrotransposons in tomato. *PLOS Genetics*, 15(9):e1008370, sep 2019.

[127] Katharina Gapp, Saray Soldado-Magraner, María Alvarez-Sánchez, Johannes Bohacek, Grégoire Vernaz, Huan Shu, Tamara B. Franklin, David Wolfer, and Isabelle M. Mansuy. Early life stress in fathers improves behavioural flexibility in their offspring. *Nature communications*, 5(1):5466, jan 2014.

[128] Katharina Gapp, Ali Jawaid, Peter Sarkies, Johannes Bohacek, Pawel Pelczar, Julien Prados, Laurent Farinelli, Eric A. Miska, and Isabelle M Mansuy. Implication of sperm RNAs in transgenerational inheritance of the effects of early trauma in mice. *Nature Neuroscience*, 17(5):667–669, may 2014.

[129] Elizabeth J. Radford. Exploring the extent and scope of epigenetic inheritance. *Nature Reviews Endocrinology*, 14(6):345–355, 2018.

[130] K. Gapp, G. van Steenwyk, P. L. Germain, W. Matsushima, K. L. M. Rudolph, F. Manuella, M. Roszkowski, Grégoire Vernaz, T. Ghosh, P. Pelczar, I. M. Mansuy, and Eric A. Miska. Alterations in sperm long RNA contribute to the epigenetic inheritance of the effects of postnatal trauma. *Molecular Psychiatry*, oct 2018.

[131] Oliver J. Rando and Rebecca A. Simmons. I'm Eating for Two: Parental Dietary Effects on Offspring Metabolism. *Cell*, 161(1):93–105, 2015.

[132] Minoo Rassoulzadegan, Valérie Grandjean, Pierre Gounon, Stéphane Vincent, Isabelle Gillot, and François Cuzin. RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. *Nature*, 441(7092):469–74, may 2006.

[133] Taewoo Ryu, Heather D. Veilleux, Jennifer M. Donelson, Philip L. Munday, and Timothy Ravasi. The epigenetic landscape of transgenerational acclimation to ocean warming. *Nature Climate Change*, 8(6):504–509, 2018.

[134] Emma Whitelaw and David I. K. Martin. Retrotransposons as epigenetic mediators of phenotypic variation in mammals. *Nature Genetics*, 27(4):361–365, 2001.

[135] Anastasiya Kazachenka, Tessa M. Bertozzi, Marcela K. Sjoberg-Herrera, Nic Walker, Joseph Gardner, Richard Gunning, Elena Pahita, Sarah Adams, David Adams, and Anne C. Ferguson-Smith. Identification, Characterization, and Heritability of Murine Metastable Epialleles: Implications for Non-genetic Inheritance. *Cell*, 175(5):1259–1271.e13, nov 2018.

[136] David Gokhman, Eitan Lavi, Kay Prüfer, Mario F. Fraga, José A. Riancho, Janet Kelso, Svante Pääbo, Eran Meshorer, and Liran Carmel. Reconstructing the DNA Methylation Maps of the Neandertal and the Denisovan. *Science*, 344(6183):523–527, may 2014.

[137] Irene Hernando-Herraez, Javier Prado-Martinez, Paras Garg, Marcos Fernandez-Callejo, Holger Heyn, Christina Hvilsom, Arcadi Navarro, Manel Esteller, Andrew J. Sharp, and Tomas Marques-Bonet. Dynamics of DNA Methylation in Recent Human and Great Ape Evolution. *PLoS Genetics*, 9(9), 2013.

[138] Antoine Molaro, Emily Hodges, Fang Fang, Qiang Song, W. Richard McCombie, Gregory J. Hannon, and Andrew D. Smith. Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell*, 146(6):1029–1041, 2011.

[139] George W. Barlow. *The Cichlid Fishes: Nature's Grand Experiment in Evolution*. Perseus Publishing, Cambridge, Massachusetts, 2000.

[140] George F. Turner. Adaptive radiation of cichlid fish. *Current Biology*, 17(19):R827–R831, oct 2007.

[141] Martin J. Genner, Ole Seehausen, David H. Lunt, Domino A. Joyce, Paul W. Shaw, Gary R. Carvalho, and George F Turner. Age of cichlids: New dates for ancient lake fish radiations. *Molecular Biology and Evolution*, 24(5):1269–1282, 2007.

[142] Matthew D. McGee, Samuel R. Borstein, Russell Y. Neches, Heinz H. Buescher, Ole Seehausen, and Peter C. Wainwright. A pharyngeal jaw evolutionary innovation facilitated extinction in Lake Victoria cichlids. *Science*, 350(6264):1077–1079, nov 2015.

[143] Thomas D. Kocher. Adaptive evolution and explosive speciation: the cichlid fish model. *Nature Reviews Genetics*, 5(4):288–298, apr 2004.

[144] Walter Salzburger. Understanding explosive diversification through cichlid fish genomics. *Nature Reviews Genetics*, 19(11):705–717, nov 2018.

[145] Walter Salzburger, Bert Van Bocxlaer, and Andrew S Cohen. Ecology and Evolution of the African Great Lakes and Their Faunas. *Annual Review of Ecology, Evolution, and Systematics*, 45(1):519–545, 2014.

[146] Thomas D. Kocher, Janet A. Conroy, Kenneth R. McKaye, and Jay R. Stauffer. Similar Morphologies of Cichlid Fish in Lakes Tanganyika and Malawi Are Due to Convergence. *Molecular Phylogenetics and Evolution*, 2(2):158–165, jun 1993.

[147] Ole Seehausen, Roger K Butlin, Irene Keller, Catherine E Wagner, Janette W Boughman, Paul a Hohenlohe, Catherine L Peichel, Glenn-Peter Saetre, Claudia Bank, Ake Brännström, Alan Brelsford, Chris S Clarkson, Fabrice Eroukhmanoff, Jeffrey L Feder, Martin C Fischer, Andrew D Foote, Paolo Franchini, Chris D Jiggins, Felicity C Jones, Anna K Lindholm, Kay Lucek, Martine E Maan, David a Marques, Simon H Martin, Blake Matthews, Joana I Meier, Markus Möst, Michael W Nachman, Etsuko Nonaka, Diana J Rennison, Julia Schwarzer, Eric T Watson, Anja M Westram, and Alex Widmer. Genomics and the origin of species. *Nature reviews. Genetics*, 15(3):176–92, 2014.

[148] Robert P Lyons, Christopher A Scholz, Andrew S Cohen, John W King, Erik T Brown, Sarah J Ivory, Thomas C Johnson, Alan L Deino, Peter N Reinthal, Michael M. McGlue, and Margaret W Blome. Continuous 1.3-million-year record of East African hydroclimate, and implications for patterns of evolution and biodiversity. *Proceedings of the National Academy of Sciences*, 112(51):201512864, dec 2015.

[149] Milan Malinsky and Walter Salzburger. Environmental context for understanding the iconic adaptive radiation of cichlid fishes in Lake Malawi. *Proceedings of the National Academy of Sciences*, 113(42):11654–11656, oct 2016.

[150] David A. Marques, Joana I. Meier, and Ole Seehausen. A Combinatorial View on Speciation and Adaptive Radiation. *Trends in Ecology & Evolution*, 34(6):531–544, jun 2019.

[151] Sarah J. Ivory, Margaret W. Blome, John W. King, Michael M. McGlue, Julia E. Cole, and Andrew S. Cohen. Environmental change explains cichlid adaptive radiation at Lake Malawi over the past 1.2 million years. *Proceedings of the National Academy of Sciences*, 113(42):11895–11900, oct 2016.

[152] Britta S. Meyer, Michael Matschiner, and Walter Salzburger. Disentangling Incomplete Lineage Sorting and Introgression to Refine Species-Tree Estimates for Lake Tanganyika Cichlid Fishes. *Systematic Biology*, 0(0):syw069, aug 2016.

[153] Kathlyn M. Stewart and Alison M. Murray. Earliest fish remains from the Lake Malawi Basin, Malawi, and biogeographical implications. *Journal of Vertebrate Paleontology*, 33(3):532–539, may 2013.

[154] George L. Mutter, David Zahrieh, Chunmei Liu, Donna Neuberg, David Finkelstein, Heather E. Baker, and Janet A. Warrington. Comparison of frozen and RNALater solid tissue storage methods for use in RNA expression microarrays. *BMC Genomics*, 5:1–7, 2004.

[155] Ad Konings. *Malawi Cichlids in their natural habitat*. Cichlid Press, 5th edition, 2016.

[156] Martin J. Genner, Paul Nichols, Gary R. Carvalho, Rosanna L. Robinson, Paul W. Shaw, Alan Smith, and George F Turner. Evolution of a cichlid fish in a Lake Malawi satellite lake. *Proceedings of the Royal Society B: Biological Sciences*, 274(1623):2249–2257, sep 2007.

[157] Oliver Hahn, Sebastian Grönke, Thomas M. Stubbs, Gabriella Ficz, Oliver Hendrich, Felix Krueger, Simon Andrews, Qifeng Zhang, Michael J. Wakelam, Andreas Beyer, Wolf Reik, and Linda Partridge. Dietary restriction protects from age-associated DNA methylation and induces epigenetic reprogramming of lipid metabolism. *Genome Biology*, 18(1):56, dec 2017.

[158] Duncan E. Edgley and Martin J. Genner. Adaptive Diversification of the Lateral Line System during Cichlid Fish Radiation. *iScience*, 16:1–11, jun 2019.

[159] Martin J. Genner and George F. Turner. Ancient Hybridization and Phenotypic Novelty within Lake Malawi's Cichlid Fish Radiation. *Molecular Biology and Evolution*, 29(1):195–206, jan 2012.

[160] Matthew A. Conte, Rajesh Joshi, Emily C. Moore, Sri Pratima Nandamuri, William J. Gammerdinger, Reade B. Roberts, Karen L. Carleton, Sigbjørn Lien, and Thomas D. Kocher. Chromosome-scale assemblies reveal the structural evolution of African cichlid genomes. *GigaScience*, 8(4):1–20, apr 2019.

[161] Felix Krueger and Simon R Andrews. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11):1571–1572, jun 2011.

[162] Jian Feng, Ningyi Shao, Keith E Szulwach, Vincent Vialou, Jimmy Huynh, Chun Zhong, Thuc Le, Deveroux Ferguson, Michael E Cahill, Yujing Li, Ja Wook Koo, Efrain Ribeiro, Benoit Labonte, Benjamin M Laitman, David Estey, Victoria Stockman, Pamela Kennedy, Thomas Couroussé, Isaac Mensah, Gustavo Turecki, Kym F Faull, Guo-li Ming, Hongjun Song, Guoping Fan, Patrizia Casaccia, Li Shen, Peng Jin, and Eric J Nestler. Role of Tet1 and 5-hydroxymethylcytosine in cocaine action. *Nature Neuroscience*, 18(4):536–544, 2015.

[163] Adam J. Bewick, Brigitte T. Hofmeister, Rob A. Powers, Stephen J. Mondo, Igor V. Grigoriev, Timothy Y. James, Jason E. Stajich, and Robert J. Schmitz. Diversity of cytosine methylation across the fungal tree of life. *Nature Ecology & Evolution*, 3(3):479–490, mar 2019.

[164] Hao Wu, Brian Caffo, Harris A. Jaffee, Rafael A. Irizarry, and Andrew P. Feinberg. Redefining CpG islands using hidden Markov models. *Biostatistics*, 11(3):499–514, jul 2010.

[165] Ikuo Ashikawa. Gene-associated CpG islands in plants as revealed by analyses of genomic sequences. *The Plant Journal*, 26(6):617–625, dec 2001.

[166] Hannah K. Long, David Sims, Andreas Heger, Neil P. Blackledge, Claudia Kutter, Megan L. Wright, Frank Grützner, Duncan T. Odom, Roger Patient, Chris P. Ponting, and Robert J. Klose. Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *eLife*, 2013(2):1–19, 2013.

[167] Michael Weber, Ines Hellmann, Michael B. Stadler, Liliana Ramos, Svante Pääbo, Michael Rebhan, and Dirk Schübeler. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genetics*, 39(4):457–466, apr 2007.

[168] Matthew A. Conte, William J. Gammerdinger, Kerry L. Bartie, David J. Penman, and T. D. Kocher. A high quality assembly of the Nile Tilapia (Oreochromis niloticus) genome reveals the structure of two sex determination regions. *BMC Genomics*, 18(1):341, dec 2017.

[169] Gary C. Hon, Nisha Rajagopal, Yin Shen, David F. McCleary, Feng Yue, My D. Dang, and Bing Ren. Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nature Genetics*, 45(10):1198–1206, oct 2013.

[170] A Sarkar, Cheolho Sim, Y. S. Hong, J. R. Hogan, M. J. Fraser, H. M. Robertson, and F. H. Collins. Molecular evolutionary analysis of the widespread piggyBac transposon family and related "domesticated" sequences. *Molecular Genetics and Genomics*, 270(2):173–180, oct 2003.

[171] E. Albertsson, A. Rad, J. Sturve, D.G.J. Larsson, and L. Förlin. Carbonyl reductase mRNA abundance and enzymatic activity as potential biomarkers of oxidative stress in marine fish. *Marine Environmental Research*, 80:56–61, sep 2012.

[172] Chongyuan Luo, Petra Hajkova, and Joseph R. Ecker. Dynamic DNA methylation: In the right place at the right time. *Science*, 361(6409):1336–1340, sep 2018.

[173] Robert S. Illingworth, Ulrike Gruenewald-Schneider, Shaun Webb, Alastair R.W. Kerr, Keith D. James, Daniel J. Turner, Colin Smith, David J. Harrison, Robert Andrews, and Adrian Bird. Orphan CpG Islands Identify numerous conserved promoters in the mammalian genome. *PLoS Genetics*, 6(9), 2010.

[174] Alan F. Hofmann, Lee R. Hagey, and Matthew D. Krasowski. Bile salts of vertebrates: structural variation and possible evolutionary significance. *Journal of Lipid Research*, 51(2):226–246, feb 2010.

[175] A. Dereeper, V. Guignon, G. Blanc, S. Audic, S. Buffet, F. Chevenet, J.-F. Dufayard, S. Guindon, V. Lefort, M. Lescot, J.-M. Claverie, and O. Gascuel. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research*, 36(Web Server):W465–W469, may 2008.

[176] Michael J Booth, Miguel R Branco, Gabriella Ficz, David Oxley, Felix Krueger, Wolf Reik, and Shankar Balasubramanian. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science (New York, N.Y.)*, 336(6083):934–7, may 2012.

[177] Yongseok Park and Hao Wu. Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics*, 32(10):1446–1453, 2016.

[178] James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26, jan 2011.

[179] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, may 2016.

[180] Harold Pimentel, Nicolas L. Bray, Suzette Puente, Páll Melsted, and Lior Pachter. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods*, 14(7):687–690, jul 2017.

[181] Milan Malinsky, Richard J Challis, Alexandra M Tyers, Stephan Schiffels, Yohey Terai, Benjamin P Ngatunga, Eric A. Miska, Richard Durbin, Martin J Genner, and George F Turner. Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science*, 350(6267):1493–1498, dec 2015.

[182] Philip Barker, David Williamson, Françoise Gasse, and Elisabeth Gibert. Climatic and volcanic forcing revealed in a 50,000-year diatom record from Lake Massoko, Tanzania. *Quaternary Research*, 60(3):368–376, 2003.

[183] George F. Turner, Benjamin P. Ngatunga, and Martin J. Genner. The Natural History of the Satellite Lakes of Lake Malawi. *EcoEvoRxiv*, (10.32942/osf.io/sehdq), 2019.

[184] Ole Seehausen, Roger K. Butlin, Irene Keller, Catherine E. Wagner, Janette W. Boughman, Paul A. Hohenlohe, Catherine L. Peichel, Glenn-Peter Saetre, Claudia Bank, Åke Brännström, Alan Brelsford, Chris S. Clarkson, Fabrice Eroukhmanoff, Jeffrey L. Feder, Martin C. Fischer, Andrew D. Foote, Paolo Franchini, Chris D. Jiggins, Felicity C. Jones, Anna K. Lindholm, Kay Lucek, Martine E. Maan, David A. Marques, Simon H. Martin, Blake Matthews, Joana I. Meier, Markus Möst, Michael W. Nachman, Etsuko Nonaka, Diana J. Rennison, Julia Schwarzer, Eric T. Watson, Anja M. Westram, and Alex Widmer. Genomics and the origin of species. *Nature Reviews Genetics*, 15(3):176–192, mar 2014.

[185] Jochen B. W. Wolf and Hans Ellegren. Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*, 18(2):87–100, feb 2017.

[186] Hongcang Gu, Zachary D Smith, Christoph Bock, Patrick Boyle, Andreas Gnirke, and Alexander Meissner. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nature Protocols*, 6(4):468–481, apr 2011.

[187] Marta Vitorino, Patricia R. Jusuf, Daniel Maurus, Yukiko Kimura, Shin-ichi Higashijima, and William A. Harris. Vsx2 in the zebrafish retina: restricted lineages through derepression. *Neural Development*, 4(1):14, 2009.

[188] Erin A. Bassett and Valerie A. Wallace. Cell fate determination in the vertebrate retina. *Trends in Neurosciences*, 35(9):565–573, sep 2012.

[189] Douglas Osei-Hyiaman, Michael DePetrillo, Pál Pacher, Jie Liu, Svetlana Radaeva, Sándor Bátkai, Judith Harvey-White, Ken Mackie, László Offertáler, Lei Wang, and George Kunos. Endocannabinoid activation at hepatic CB1 receptors stimulates fatty acid synthesis and contributes to diet-induced obesity. *Journal of Clinical Investigation*, 115(5):1298–1305, may 2005.

[190] Benjamin A. Sandkam, Laura Campello, Conor O'Brien, Sri Pratima Nandamuri, William Gammerdinger, Matthew Conte, Anand Swaroop, and Karen L. Carleton. Tbx2a modulates switching of opsin gene expression. *bioRxiv*, (doi.org/10.1101/676478.):1–30, 2019.

[191] Zuzana Musilova, Adrian Indermaur, Arnold Roger Bitja-Nyom, Dmytro Omelchenko, Monika Kłodawska, Lia Albergati, Kateřina Remišová, and Walter Salzburger. Evolution of visual sensory system in cichlid fishes from crater lake Barombi Mbo in Cameroon. *Molecular Ecology*, 112(483):mec.15217, aug 2019.

[192] Zuzana Musilova, Fabio Cortesi, Michael Matschiner, Wayne I. L. Davies, Jagdish Suresh Patel, Sara M. Stieb, Fanny de Busserolles, Martin Malmstrøm, Ole K. Tørresen, Celeste J. Brown, Jessica K. Mountford, Reinhold Hanel, Deborah L. Stenkamp, Kjetill S. Jakobsen, Karen L. Carleton, Sissel Jentoft, Justin Marshall, and Walter Salzburger. Vision using multiple distinct rod opsins in deep-sea fishes. *Science*, 364(6440):588–592, may 2019.

[193] Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1):W191–W198, jul 2019.

[194] Hyung Joo Lee, Rebecca F Lowdon, Brett Maricque, Bo Zhang, Michael Stevens, Daofeng Li, Stephen L Johnson, and Ting Wang. Developmental enhancers revealed by extensive DNA methylome maps of zebrafish early embryos. *Nature communications*, 6:6315, 2015.

[195] Jelena Rajkov, Alexandra Anh-Thu Weber, Walter Salzburger, and Bernd Egger. Adaptive phenotypic plasticity contributes to divergence between lake and river populations of an East African cichlid fish. *Ecology and Evolution*, 8(15):7323–7333, aug 2018.

[196] Mélanie Rigal, Claude Becker, Thierry Pélissier, Romain Pogorelcnik, Jane Devos, Yoko Ikeda, Detlef Weigel, and Olivier Mathieu. Epigenome confrontation triggers immediate reprogramming of DNA methylation and transposon silencing in Arabidopsis thaliana F1 epihybrids. *Proceedings of the National Academy of Sciences*, 113(14):E2083–E2092, apr 2016.

[197] William T Jordan and Robert J Schmitz. The shocking consequences of hybrid epigenomes. *Genome Biology*, 17(1):85, dec 2016.

[198] Mélanie Dapp, Jon Reinders, Alexis Bédiée, Crispulo Balsera, Etienne Bucher, Gregory Theiler, Christine Granier, and Jerzy Paszkowski. Heterosis and inbreeding depression of epigenetic Arabidopsis hybrids. *Nature Plants*, 1(7):15092, jul 2015.

[199] Kathleen M. Gilmour and S. F. Perry. Carbonic anhydrase and acid-base regulation in fish. *Journal of Experimental Biology*, 212(11):1647–1661, jun 2009.

[200] Mika Hilvo, Martti Tolvanen, Amy Clark, Bairong Shen, Gul N. Shah, Abdul Waheed, Piia Halmi, Milla Hänninen, Jonna M. Hämäläinen, Mauno Vihinen, William S. Sly, and Seppo Parkkila. Characterization of CA XV, a new GPI-anchored form of carbonic anhydrase. *Biochemical Journal*, 392(1):83–92, 2005.

[201] Xinxin Zhao, Yang Chai, and Bao Liu. Epigenetic inheritance and variation of DNA methylation level and pattern in maize intra-specific hybrids. *Plant Science*, 172(5):930–938, may 2007.

[202] Quentin Gouil and David C. Baulcombe. Paramutation-like features of multiple natural epialleles in tomato. *BMC Genomics*, 19(1):203, dec 2018.

[203] Mazhar Adli. The CRISPR tool kit for genome editing and beyond. *Nature Communications*, 9(1):1911, dec 2018.

Appendix A

Supporting figures

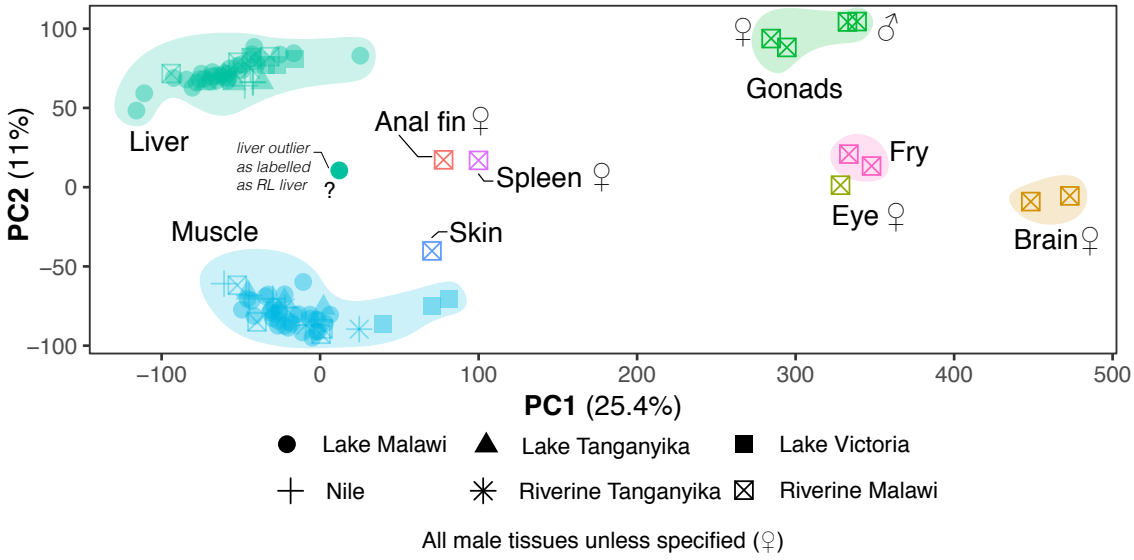


Fig. A.1 Transcriptomic data of several tissues in East African cichlids. Principal component analysis (PCA) of transcriptomic divergences in several tissues of cichlids of Lakes Malawi, Tanganyika, Victoria and of the rivers part of the Lake Malawi and Lake Tanganyika catchments and included as well the riverine cichlid Nile Tilapia. All sequencing data (paired-end 75 bp-long reads) were mapped to Lake Malawi *M. zebra* reference genome (UMD2a). Tissues and sexes are indicated in the graph directly. Liver and muscle tissues were sequenced for all samples, apart from *A. calliptera* sp. Itupi (Riverine Malawi) for which anal fin, skin, spleen, eye, fry, brain and gonads were sequenced as well (used to functionally annotate the reference genome of *A. calliptera* [fAstCal1.2., NCBI, published in 2018]). All samples were wild-caught. Importantly, most of the variance is explained by the tissue of origin, and not by the Lake of origin.

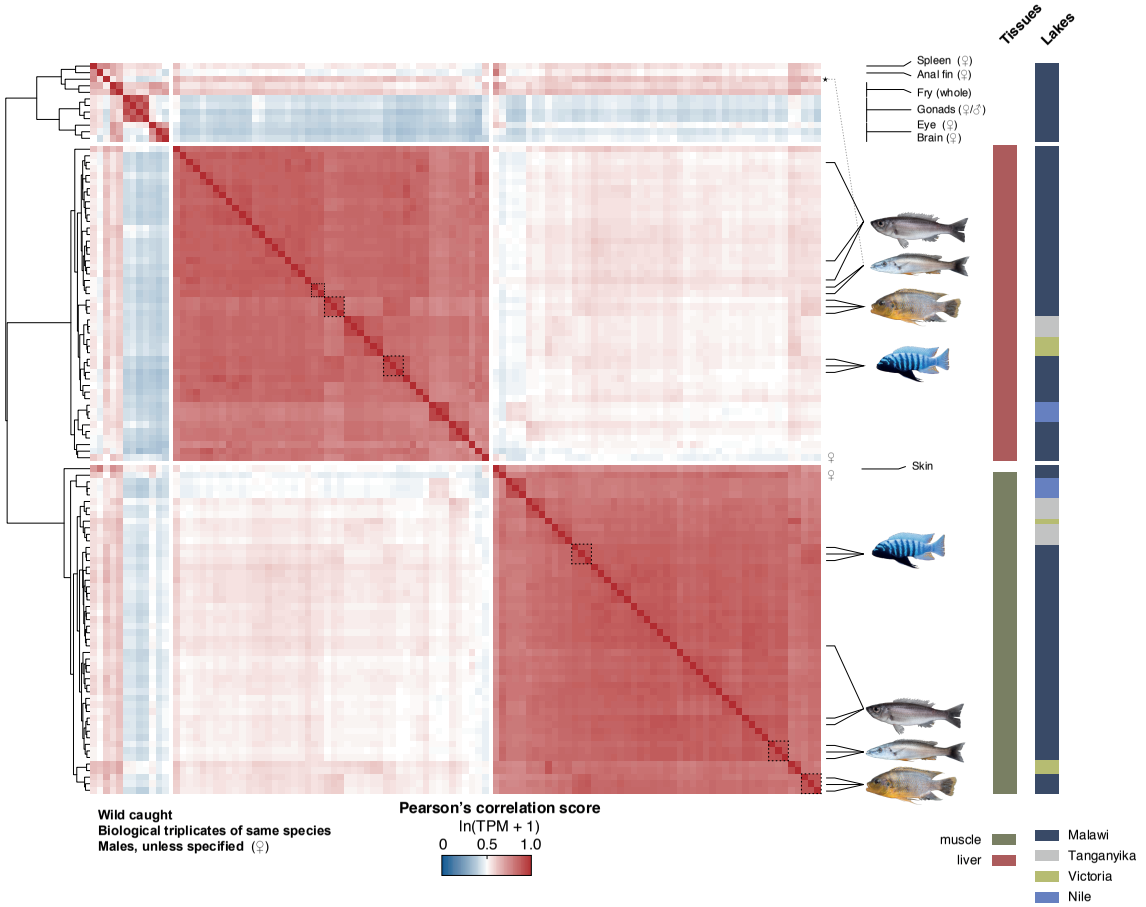


Fig. A.2 Transcriptomic variability in several tissues of East African cichlids. Heatmap of pairwise Pearson's correlations of transcriptomic sequencing data of several tissues of cichlids of Lakes Malawi, Tanganyika, Victoria and of the rivers part of the Lake Malawi and Lake Tanganyika catchments and included as well the riverine cichlid Nile Tilapia. All sequencing data (paired-end 75 bp-long reads) were mapped to Lake Malawi *M. zebra* reference genome (UMD2a). Gene expression values were estimated in ln(TPM + 1). Liver and muscle tissues were sequenced for all samples, apart from *A. calliptera* sp. Itupi (Riverine Malawi) for which anal fin, skin, spleen, eye, fry, brain and gonads were sequenced as well. All samples were wild-caught. Photographs indicate the Lake Malawi cichlid species analysed in this thesis.

Appendix B

Published work done in parallel to this thesis

In parallel to this thesis, I collaborated on two different projects, both are now published. These projects are not part of this thesis and therefore are not discussed here. Below, I briefly summarise my contribution and include the article references.

B.1 Epigenetic remodelling licences adult cholangiocytes for organoid formation and liver regeneration

The first collaborative project pertained to the investigation of the epigenetic basis of liver cell reprogramming and regeneration. In brief, we provided evidence that the gene *Tet1* was up-regulated upon liver damage both *in vitro* and *in vivo* in mouse, allowing for cell reprogramming and liver regeneration through global DNA demethylation (and production of 5hmC specifically at promoter regions) and transcriptional changes.

My contribution was to generate WGBS, RRBS and RNAseq and analyse sequencing data, as well as to participate in the writing process of the manuscript.

Aloia, L.*, McKie, M.*, **Vernaz, G.*** *et al.* Epigenetic remodelling licences adult cholangiocytes for organoid formation and liver regeneration. *Nat. Cell Biol.* 21, 1321–1333 (4th November 2019). doi: 10.1038/s41556-019-0402-6 – Ref. [70]

* These authors have contributed equally to this work.

B.2 Alterations in sperm long RNA contribute to the epigenetic inheritance of the effects of postnatal trauma

The second collaboration aimed at characterising the contribution of sperm long non-coding RNAs in the transgenerational epigenetic inheritance of early life trauma-associated behaviours. We provided evidence that both specific small and long non-coding RNAs were present in sperm of mouse and that were participating in the transmission of certain traits associated with parental early life expose to trauma to the progeny. We confirmed the functional implications of these sperm lncRNA by performing injection of these sperm lncRNA into the oocyte of unexposed, naive mouse females. Offspring (up to F2) born from these sperm lncRNA-injected oocytes consistently showed altered behaviours and metabolism, similar to the ones observed in the exposed parents, suggesting epigenetic inheritance of certain traits associated with early life trauma.

My contribution to this work was first to isolate and purify of long non-coding RNAs from sperm cells from both control and 'stressed' mouse mothers and then to perform differential gene expression analysis to identify lncRNA transcripts.

Gapp, K., ..., **Vernaz, G.** *et al.* Alterations in sperm long RNA contribute to the epigenetic inheritance of the effects of postnatal trauma. *Mol. Psychiatry* (30th October 2018). doi:10.1038/s41380-018-0271-6 – Ref. [130]