



Norwegian University of Life Sciences
Faculty of Biosciences

Philosophiae Doctor (PhD)
Thesis 2022:72

The contribution of repetitive elements to salmonid genome evolution

Bidraget frå gjentakande element til evolusjon av laksefiskers genom

Øystein Monsen

The contribution of repetitive elements to salmonid genome evolution

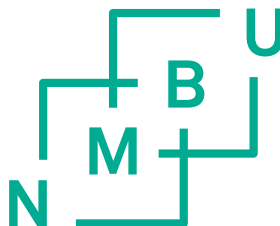
Bidraget frå gjentakande element til evolusjon av laksefiskers genom

Philosophiae Doctor (PhD) Thesis
Øystein Monsen

Norwegian University of Life Sciences
The PhD programme in Animal and Aquacultural Sciences
at the
Faculty of Biosciences

Ås 2022

Thesis number 2022:72
ISSN 1894-6402
ISBN 978-82-575-2025-0



*Det er den svarte prikken
midt i skiva du skal treffa,
nett den, der
skal pili stå og dirra!
Men nett der treffer du ikkje.
Du er nær, nærare,
nei, ikkje nær nok.
So lyt du gå og plukka upp pilene,
gå tilbake, prøva på nytt.
Til du forstår pili
som stend der og dirrar:
Her er òg eit midtpunkt.
-Olav H. Hauge*

Supervisors and Evaluation Committee

The candidate's supervisory group consists of:

Main supervisor: Prof. Simen Rød Sandve, NMBU

Co-supervisor: Prof. Rori Rohlf, San Francisco State University

Co-supervisor: Prof. Torgeir Rhoden Hvidsten, NMBU

Co-supervisor: Prof. Sigbjørn Lien, NMBU

Co-supervisor: Ass. Prof. Alexander Sand-Jae Suh

The candidate's evaluation committee consists of:

Dr. Josefa González, Spanish Research Council

Address: Passeig Marítim de la, Barceloneta 37-49, 08003 Barcelona (Spain)

Phone: +34638182935

E-mail: josefa.gonzalez@csic.es

Dr. Ole Kristian Tørresen, University of Oslo

Address: Postboks 1066 Blindern, 0316 Oslo

Phone: +47 98 81 34 15

E-mail: o.k.torresen@ibv.uio.no

Committee coordinator:

Ass. Prof. Matthew Petter Kent

Address: P.O. Box 5003 NMBU, NO-1432 Ås, NORWAY

Phone: +47 67232701

E-mail: matthew.peter.kent@nmbu.no

Acknowledgements

The work presented in this thesis has been carried out at the Department of Biosciences at the Norwegian University of Life Sciences (NMBU).

Firstly, my sincere thanks to my main supervisor, Simen R. Sandve. No collaboration of this nature can be entirely without friction and frustration, but at every juncture where it has mattered, you've had my back in an exemplary way for which I really am genuinely grateful. A special recognition must go to my co-supervisors, Lars Grønvold, Torgeir Hvidsten, Sigbjørn Lien and Alexander Suh. In various ways, this work would not be possible without their sage advice, input, analysis or just supplying basic competence and clarity of thought which I myself have lacked.

The work environment at Cigene has been excellent, with downright jovial colleagues. Special thanks go here to Kristina Stenløkk - working with you has been a pleasure; Gareth Gillard, whose presence at the office has been an unpredictable social boost and steady source of much-needed sugar at critical moments; Marie-Odile Baudement, who demonstrated to me the limits of my stamina in the lab; Olga Pawlowskaya who has consistently been an excellent friend in misery, and finally Lars Grønvold again for his refined sense of discipline and inexplicably stable humour. It is, naturally, impossible to list the names of everyone who's helped ameliorate my fundamental mediocrity. If you are not mentioned, I no doubt have some nefarious scheme to undermine your self-esteem. Please don't take it personally.

Finally, an enormous thanks to my family and friends. I would have gone completely mad without you. I owe you all an incalculable debt for keeping up with me and keeping me going through these years.

Bergen, September 2022

Øystein Monsen

TABLE OF CONTENTS

Summary	7
Samandrag	9
List of papers	11
1. Introduction	12
1.1: Introducing the study system	12
1.1.1: Our Hero	13
1.1.2: Genomic Redundancy	13
1.2.1: Repetitive DNA	16
1.2.1: TEs	17
1.2.2: Satellites etc	19
1.2.4: The origins of repetitive DNA	19
1.2.5: The salmonid repeat landscape	21
1.3: Consequences of whole-genome duplication on genome evolution	22
1.3.1: The vertebrate genomes and ancient genome duplication events	22
1.3.2: Double trouble - the immediate impact of WGD	22
1.3.3: Opportunities arise – long term consequences of WGD	23
1.3.4: Role of TEs following WGD	25
1.4: Gene regulation	27
1.4.1: Genes and regulatory concepts	27
1.4.2: Duplication, regulation and transcription factors	29
1.4.3: Chromatin and chromatin accessibility	30

1.5: Genome sequencing technology unleash a repetitive-DNA revolution	31
1.5.1. Where we are in gene sequencing technology	31
1.5.2. A new generation of sequencing	32
1.5.3. The return of the repeats	34
2: Summary of results	35
2.1: Paper 1 - Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication	35
2.2: Paper 2 - The role of transposable elements in the evolution of cis-regulatory element landscapes after whole genome duplication	36
2.3: Paper 3 - Structural variation landscape reflects telomeric tandem repeat expansions in Atlantic salmon	38
3. Discussion	40
3.1: Shifting grounds - technological advances in genome assembly and research on repeat DNA	40
3.2 Transposable elements; classification, annotation and a maturing field	41
3.3 Evolutionary dynamics of transposable elements	43
3.4 Satellite DNA; status of the field, proliferation and effects	46
3.5 Future perspectives	48

Summary

Eukaryotic genomes typically consist of a substantial proportion of repetitive DNA in the form of transposable elements (TEs) and satellite DNA. From studies of mammals, model species, and a few other well studied lineages it is clear that repetitive elements play roles in many important cellular processes and shape evolution of genomes and organisms. However, little is still known about the role of repetitive DNA in biology and genome evolution for most eukaryotic species. Here we use a suite of omics data and genomics analyses to ask the question: What is the role of repeat DNA in genome regulation and structural variation in the Atlantic salmon genome?

In papers 1 and 2 we studied the link between evolution of gene regulation and transposable elements in the context of the salmonid whole genome duplication. We found that gene duplicate copies that had evolved lower gene expression across most tissues had increased TE insertion rates in the promoters. In addition, we found that duplicate copies evolving liver specific increase in gene expression, had gained transcription factor binding sites (TFBS) for liver-specific transcription factors in the promoters, and some of these were found inside TEs. In depth analyses of cis-regulatory elements (CREs) in Paper 2 showed that 15-20% of CRE are within TEs (TE-CREs) and that there were fewer TE-CREs active in brain tissue compared to liver. Interestingly, a small heterogeneous group of TE subfamilies (11%) had contributed ~45% of all TE-CREs, but the 'superspreader' activity did not seem to peak in the time shortly following the WGD. CREs donated by 'superspreaders' were enriched for many different TFBSs, however, highly brain specific TFBSs were extremely rare in TEs, indicating that strong purifying selection shape TE-CRE evolution.

In Paper 3 we studied the role of repeat-DNA in the evolution of structural genomic variation (SVs) (>50bp). Leveraging seven new long read genome

assemblies we find a large number of so far unknown structural variants, and conclude that satellite DNA is highly associated with indel variants. TEs, on the other hand, had contributed comparatively much less to the SV landscape. We conclude that the enormous number of novel SV found in our study is mostly due to satellite expansion and -contraction processes. This thesis provides an advance in our understanding of the role of repetitive DNA in the evolution of salmonid genomes, and paves the way for future studies into the functional importance of this vast sea of repetitiveness.

Samandrag

Ekaryote genom består vanlegvis av ein vesentleg andel gjentakande DNA i form av transposable element (TE-ar) og satelitt-DNA. Frå studiar i pattedyr, modelartar og nokre andre velstuderte organismar er det klart at gjentakande element spelar rollar i mange viktige cellulære prosessar og formar evolusjon av genom og organismar. Mykje er imidlertid ukjend om rolla til gjentakande DNA i biologi og genomevolusjon i dei fleste eukaryote artar. Her nyttar vi ein rekke omikk-data og genomiske analysar for å stille spørsmålet: Kva er rolla til gjentakande DNA i genomregulering og strukturell variasjon i genomet til atlantisk laks?

I artikkel 1 og 2 studerte vi koplinga mellom evolusjon av genregulering og transposable element i lys av den salmonide heilgenomdupliseringa. Vi fann at dupliserte gener som hadde redusert uttrykk over dei fleste vev hadde større andeler TE i promoteren. I tillegg fann vi at dupliserte gener som evolverte leverspesifikk auke i genuttrykk hadde fått transkripsjonsfaktorbindingseter (TFBS) for lever-spesifikke transkripsjonsfaktorar, og at somme av desse var inne i TE-ar. Djupare analysar av cis-regulatoriske element (CRE-ar) i Artikkel 2 synta at 15-20% av CRE-ar er inne i TE-ar (TE-CRE-ar) og at det var færre TE-CRE-ar aktive i hjernevev enn i lever. Interessant nok bidro ei lita, heterogren gruppe (11%) med TE-ar med ca 45% av alle TE-CRE-ar, men denne “superspreiaraktiviteten” så ikkje ut til å nå høgda umiddelbart etter heilgenomdupliseringa. CRE-ar gitt av “superspreiarar” var anrika for mange forskjellige TFBS-ar, men svært hjernespesifikke TFBS-ar var svært sjeldne i TE-ar, som antyder at sterk seleksjon har forma TE-CRE-evolusjon.

I artikkel 3 studerte vi rolla til gjentakande DNA med tanke på strukturell genomisk variasjon (SV-ar) (>50bp). Vi utnytta sju genomsamansetjingar basert på «long-read-sekvensering» for å finne eit stort tal hittil ukjende strukturelle variantar, og sluttar at satelittDNA er svært nært kopla til indelvariantar. TE-ar hadde på den andre sida bidrege mykje mindre til SV-landskapet. Vi sluttar at den enorme auka i SV-ar vi fann i vår studie stort sett følgjer av satelittutviding og -innsnevring.

Denne avhandlinga gir eit framsteg i får kjennskap til gjentakande DNA si rolle i evolusjonen til salmonide genom, og reiar grunnen for vidare studier på den funksjonelle betydninga til dette gjentakande havet.

List of papers

Paper 1: Gillard, G. B., Grønvold, L., Røsæg, L. L., Holen, M. M., **Monsen, Ø.**, Koop, B. F., Rondeau, E. B., Gundappa, M. K., Mendoza, J., Macqueen, D. J., Rohlf, R. V., Sandve, S. R. & Hvidsten, T. R.

Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. Genome Biol. 22, 103 (2021).

My contribution to Paper 1: Analyses of TE-annotation and TE-landscape in promoters. Helped with interpretation of results.

Paper 2: **Monsen, Ø.**, Grønvold, L., Kijas, J., Suh, A., Hvidsten, T. R. & Sandve, S. R.
The role of transposable elements in the evolution of cis-regulatory element landscapes after whole genome duplication

My contribution to Paper 2: Helped plan the study. Performed data analyses and/or visualisations of results from these analyses related. Contributed to the interpretation of all findings. Wrote the first draft of the manuscript.

Paper 3: **Monsen, Ø.**, Stenløkk, K.R.S., Sandve, S.R. & Lien, S.

Structural variation landscape reflects telomeric tandem repeat expansions in Atlantic salmon

My contribution to Paper 3: Made repeat libraries and annotations of both TEs and non-TE repetitive DNA in the new long-read assembly of Atlantic salmon. Performed all analyses related to repeat DNA and their association (or not) with structural variation. Contributed to the interpretation of all findings. Co-drafted the manuscript with K.R.S. Stenløkk.

1 Introduction

1.1 Introducing the study system

1.1.1 Our Hero

Once upon a time, some 100 million years ago (Grimholt, 2018), there was a fish. The fish was a bony fish - a teleost - and likely lived in a period of evolutionary turmoil (Carretero-Paulet and Van de Peer, 2020), and one or a few eggs carried with them a spectacular mutation: A whole-genome duplication. The descendants of these first mutants survived, adapted successfully to new environments, climates, and habitats, and eventually became dozens of new species (Macqueen et al., 2017). These descendants of our fishy protagonist, the clade of salmonids (Fig. 1a), are the main characters of this PhD-story.

1.1.2 Genomic Redundancy

The history of the ancestral genome duplication in the salmonids is fascinating on its own terms. This sudden burst of genomic redundancy involves a doubling not only of gene numbers and regulatory elements, but also provides a tremendous amount of added raw material, non-coding RNA and assorted genetic debris. This creates an evolutionary playground which has probably had consequences for genome and trait evolution. And most certainly it has sparked speculations and hypotheses (Campbell et al., 2019; Gillard, 2019; Grimholt, 2018) about the genomic basis of the wide array of salmonid lifestyles and adaptations. For example the curious adaptation to an anadromous lifestyle (Alexandrou et al., 2013); that is, some salmonids live in both fresh and salt water at different points in their life cycles.

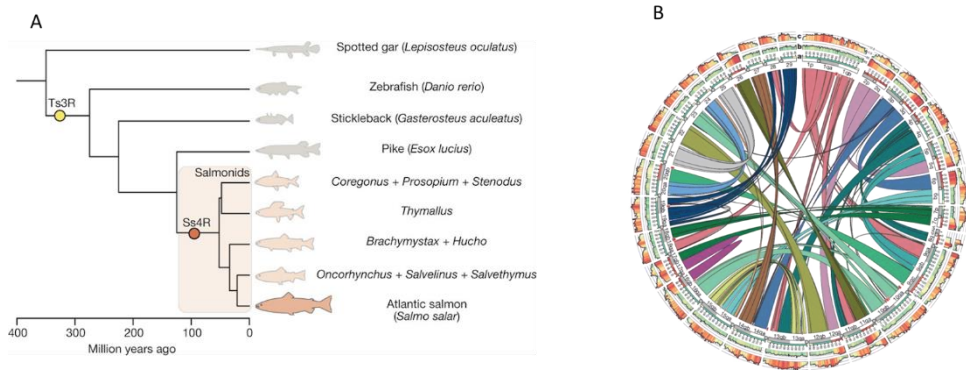


Fig. 1. The Atlantic Salmon genome A) A phylogeny showing representatives from the five major salmonid clades and some selected teleost outgroups. Yellow and red circles indicate the teleost-specific and salmonid-specific whole-genome duplications (*Ts3R* and *Ss4R*, respectively). B) The archetypal duplicated structure of a salmonid genome. Circos plot of the Atlantic salmon genome showing duplicated regions (bands) linking together duplicated chromosome regions originating from the salmonid whole genome duplication event. Figure adapted from (Lien et al., 2016) under CC B.Y. 4.0 licence.

However, the doubled genome is not the only source of genomic redundancy in salmonid genomes, nor the only thing that makes them interesting to both geneticists and evolutionary biologists. These genomes also hold huge amounts of repetitive DNA, most originating from genomic parasites called transposable elements (TEs) starting to multiply at the time of the genome duplication. They also contain a substantial amount of satellite DNA (satDNA), short DNA sequences repeated various number of times next to each other. In fact, salmonid genomes are proper outliers with respect to the amount of redundant (duplicated) genetic material they contain compared to most other vertebrates (Biscotti et al., 2015; Sotero-Caio et al., 2017). The consequences of this level of redundancy are many, and distinguishing between them will be a major part of this treatise and in working with this sort of issue in general, but for now it should suffice to say that the categories of repetitive DNA, more general gene duplications and whole-sale whole-genome duplications all fall under this rough rubric.

Why should we care about all these elements existing in several copies, either duplicated genes, or other DNA sequences with hundreds or thousands of genomic copies? Because accumulation and retention of these redundant genomic features can, as we will see in the following chapters, have a large impact on the evolution of genome function and thereby the organism's biology. (Andalis et al., 2004; Ohno et al., 1968)

1.2 Repetitive DNA

DNA, in essence being a recipe for life written in a four-letter alphabet, by necessity has a lot of repetition going on. Much of this is simply random, but purely non-functional entities do not tend to survive long in evolution. Thus, large arrays of repetition are not naïvely expected. However, practically all known genomes contain repetitive DNA - even in notoriously compact viral genomes, TEs have been detected (Loiseau et al., 2021). In some plants, the proportion of repetitive DNA is extreme - famously, maize has around 85% (Stitzer et al., 2021). Though that is exceptional, many other plants also have highly repetitive genomes often in the order of 80% TE-derived/repeat DNA (Garbus et al., 2015; Wicker et al., 2017). In the vertebrate group, where our main characters of this thesis belong, it is common that around half the genome is repetitive DNA (Sotero-Caio et al., 2017). While the proportion of repetitive DNA in the genomes of the Atlantic salmon and its fellow salmonid comrades, however, is around 60%.

As I have briefly mentioned already, repetitive DNA comes in two main flavours: Interspersed repeats, also known as Transposable Elements (TEs), and tandem repeats, in this treatise sometimes used interchangeably with “satellite DNA”. Repetitive DNA is a major driver of evolution, and, due to reasons which will be touched upon in this chapter, an underexplored one. In the following chapters, I will briefly review the classification and diversity of different genomic repeat elements, their origin, and the current state-of-knowledge when it comes to the repeat landscapes of salmonid genomes.

1.2.1 Transposable Elements

Transposable elements are bits of DNA which are capable of reproducing – transposing – themselves throughout the genome (Wicker et al., 2007). TEs come in a tremendous variety of shapes and sizes, from low double digits of nucleotides to tens or even hundreds of thousands-of-nucleotide long megaelements (Arkhipova and Yushenova, 2019). Some TEs maintain their own genes to facilitate this transposition autonomously, whereas others “hitchhike” off those which do (Naville et al., 2019). While TEs are not random in their insertion pattern – in particular, they need to insert into open chromatin – they are found all over the genome and show no strong preferences for particular genomic regions (recombination hotspots, certain telomere arms etc) (Quesneville, 2020; Wells and Feschotte, 2020). However, when breaking TEs into smaller sub-groups based on sequence homology and study accumulation of TEs over long term (millions of years), it is clear that some types of TEs are not equally likely to end up in all types of genomic regions (e.g. (Buckley et al., 2017), reviewed in (Bourque et al., 2018)).

TEs come in two main variants, according to their mechanism of transposition: Cut-and-paste and copy-paste (Wicker et al., 2007) (Fig 2). Cut-and-paste transposons (Fig 2a) are excised from the genome and reproduced using a DNA intermediate. Thus, these are sometimes called “DNA” transposons. The most common type of cut-and-paste transposon is the terminal inverted repeat (TIR), which forms a kind of circular structure along the “sticky” TIRs once excised, is reproduced by the ordinary apparatus of cell division, and then reintegrated by docking at certain target motifs. Thus, they “jump” around the genome, leaving behind them target site duplications.

Copy-paste transposons are not excised (Fig 2b): they are transcribed as though they were ordinary genes, and an RNA intermediate is generated before DNA is once more produced and the transposon reintegrated into the host genome. Consequently, they have more of a size limit due to the relative instability of RNA and are more heterogeneous in terms of gene content (Arkhipova and Yushenova, 2019; Gluck-Thaler et al., 2021) (Gluck-Thaler et al., 2021), though there are also giant copy-paste transposons (Arkhipova and Yushenova, 2019). Whereas most superfamilies of autonomous TIR transposons make do with a single transposase gene of its own, RNA transposons often need several genes encoding for the full repertoire of proteins needed for the transposition, including among other things reverse transcriptases and endonucleases.

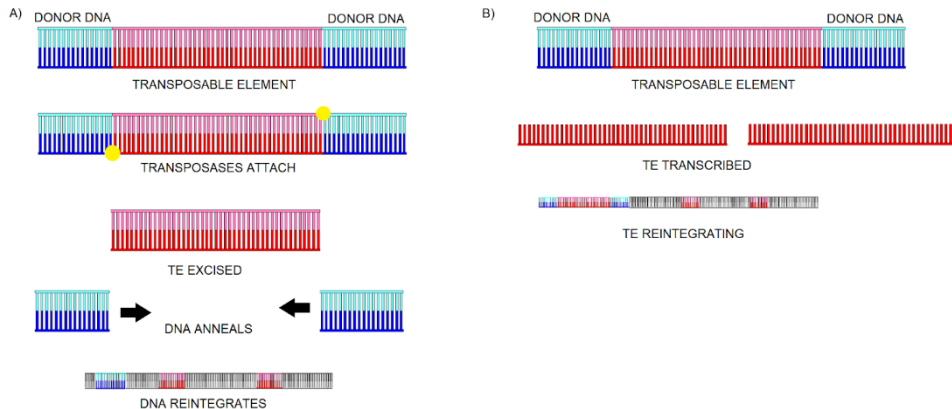


Fig. 2. The two main classes of transposable elements. A) a TIR's cut-and-paste transposition process, by which the entire element is excised from the genome leaving a double-strand break which must be repaired before reproducing and reintegrating. B) a retrotransposon's copy-paste mechanism, by which a single-stranded RNA intermediary is transcribed and the "donor" element left intact in its original context. There are also other, more exotic mechanisms such as the rolling-pin structures of helitrons, but for the purposes of this treatise, the relevant distinction is cut-and-paste and copy-paste replication.

1.2.2 Satellites

Another important kind of repetitive element is satellite DNA (Garrido-Ramos, 2017) (satDNA). SatDNA is made up of repeated DNA “words” (Fig 3), placed in tandem, i.e. directly adjacent to each other. SatDNA are the most difficult and enigmatic parts of the genome, making up much of what used to be considered “junk” (i.e. non-coding and thus wrongly assumed to be biologically irrelevant) DNA.



Fig. 3. A simple illustration of a short satDNA array. *In this case, the repeating “word”/monomer is the simple “TA” motif, which is surrounded by less-repeated DNA sequence. SatDNA arrays can have much longer monomers and will often span many thousands of bps in total.*

1.2.3 The origins of repetitive DNA

If repetitive DNA makes up such a large proportion of the genome, and its evolution is so intertwined with that of genes, regulatory mechanisms and chromatin structure, its genesis becomes a matter of great interest and importance. Intriguingly, the origins of repetitive DNA are both multifarious and occasionally somewhat unclear, but we know some things and can speculate as to others.

Transposable elements have origins which appear dependent on their subtype and class; irritatingly, the class is a functional category and does not necessarily denote any actual phylogenetic relation (Amselem et al., 2019; Kapitonov and Jurka, 2008; Wells and Feschotte, 2020; Wicker et al., 2007); two given DNA/TIR superfamilies may be genuinely unrelated to each other, for instance (McDonald, 1993), though attempts can be made to connect them through their signature elements (in the case of DNA/TIR transposons,

the mode of action of their transposases) (Wells and Feschotte, 2020).

Attempts to prove broader common ancestry quickly become very tenuous, however - consider figure 2 in the previous citation. To my knowledge, little recent work has been done in attempting to elucidate the precise origins of TE families, possibly for this very reason; findings are not obviously transferable, and our present system for studying them is focussed on function and then structure, rather than phylogenetic traits.

Some things may be said with reasonable confidence, however:

1: Transposable elements engage in a certain measure of horizontal transfer (Zhang et al., 2020), though the mechanisms vary greatly.

2: Many DNA transposons are likely evolved from genes (McDonald, 1993; Wells and Feschotte, 2020) with which they previously interacted

3: Some, but not all eukaryotic transposons have homologues active in prokaryotes, i.e. they have either committed a relatively recent horizontal transfer or they predate the eukaryote/prokaryote split (Gilbert and Cordaux, 2013) (Loiseau et al., 2021; Zhang et al., 2020).

4: Despite a quite bewildering diversity of TE sequences and families, the fundamental modes of action of known eukaryotic TEs are relatively well-defined (Wells and Feschotte, 2020) and seem to be related with other RNA- or DNA-binding motifs.

The origin of satellite DNA seems, if not simpler then at least more unitary. Generally, satDNA appears to expand from “accidents” in DNA replication, where the polymerase machinery for some reason “slips” (Richard et al., 2008) - this can be from hairpin formations or other basic enzymokinetic mishaps - and replicates the same bit of DNA several times before moving on. This implies that there are certain sequences which are more susceptible to such accidents, and where it is less subject to purifying selection under

certain circumstances. This can be seen as a motivation for the so-called library hypothesis, to which we will return later.

1.2.4 The salmonid repeat landscape

Salmonid species, following their whole genome duplication, experienced a significant expansion in repeat content (Berthelot et al., 2014; Christensen et al., 2018; Lien et al., 2016) with transposable elements usually making up north of 40% of the total assembled DNA. One peculiarity of teleosts in general is their relatively high abundance of DNA transposons; in many teleosts, hAT transposons predominate but in salmonids the tc1-Mariner superfamily is the largest (Brynhildsen, 2016). One recent paper on this, examining structural variation and transposable elements in whitefish, finds 60% of the genome made up out of TEs (Mérot et al., 2022). Interestingly, they have a greater preponderance of young retroelements than is seen in Atlantic Salmon.

In our work with this treatise, we have found some evidence suggesting that satellites are “seeded” throughout the genome before “re-expanding”, but this has been difficult to formalise. Many satellites seem quite “native” to certain chromosomes, more or less specifically associated with one chromosome - or, in our case, with the chromosome and its ohnologue, i.e. the duplicated chromosome region homologue derived from the whole genome duplication event. This seems to follow from our expectations of satDNA in general (Garrido-Ramos, 2017). Otherwise, it does not appear that the distribution of TEs or satDNA in salmonids differ notably from other teleosts - or, indeed, from most vertebrates. Satellites are predominant around centromeres and telomeres and are subject to frequent assembly collapse, TEs (and satDNA) are depleted in and immediately around genes,

and as we shall see there is a tendency of TEs to be depleted in satDNA-rich regions.

1.3 Consequences of whole-genome duplication on genome evolution

1.3.1 The vertebrate genomes and ancient genome duplication events

All vertebrates have two known whole-genome duplication events in their evolutionary history, referred to as 1R and 2R (Ref). All the teleosts (that is the bony fishes) share another ancient vertebrate WGD called Ts3R, dating back approximately 350 Mya (Gundappa et al., 2022; Jaillon et al., 2004; Robertson et al., 2017). As noted, in the previous section, the salmonids experienced their very own whole-genome duplication event, referred to as Ss4R, which dates back ~120-100 Mya (Gundappa et al., 2022). Each WGD has immediate consequence for the organism's cells (Baduel et al., 2019; Bomblies, 2020; Gemble et al., 2022; Hollister et al., 2012; Storchová et al., 2006), as well as potential wide-ranging long-term implications for evolution of the species, its genes, and its descendants (Berthelot et al., 2014; Gillard, 2019; Ohno et al., 1968; Van de Peer et al., 2017)

1.3.2 Double trouble - the immediate impact of WGD

The sheer scale of transformation following a whole-genome duplication is difficult to overstate. Billions of base-pairs and tens of thousands of additional gene copies are introduced into the genome. In the case of an autopolyploidisation event (doubling of its own genome), such as the salmonid ancestor went through, having four identical copies of each

chromosome will pose issues during meiosis (Pelé et al., 2018; Schmid et al., 2015). From this, we expect that WGDs will be followed by an immediate and strong selection pressure to stabilise meiosis and avoid aneuploidy. This prediction is supported by selection signatures on meiosis genes in autopolyploid *Arabidopsis* (Bohutínská et al., 2021; Wright et al., 2015) and selection on gene expression of meiosis related genes following WGD in salmonids (Gillard, 2019).

If new-born polyploids escape immediate death due to meiosis-related issues, polyploid genomes start their journey towards rediploidization, meaning the return to a diploid-like genome structure and stable bivalent pairing of chromosomes in meiosis. One effective way to break unwanted pairing of duplicated chromosomes is to evolve structural mutations that can block recombination (Dréau et al., 2019; Zhang et al., 2021).

Interestingly, TEs are known to be associated with the rise of structural variants through for example uneven recombination (Bourque et al., 2018), and it is therefore hypothesised that TE activity after the salmonid WGD (see Figure Y) has promoted rediploidization through generation of structural variants (SVs) (Lien et al., 2016).

1.3.3 Opportunities arise – long term consequences of WGD

As rediploidisation becomes the new norm, duplicated genes and genomic regions get protected from “gene flow” between duplicated chromosomes. Duplicates therefore starts evolving as independent chromosomes, independently accumulate new mutations, which drive sequence- and functional divergence of genes (Lynch and Conery, 2000; Naseeb et al., 2017; Sandve et al., 2018).

There are four main models of long term gene duplicate evolution; retention of ancestral function(s), neofunctionalisation, subfunctionalisation, or non-functionalisation (Force et al., 1999; Lynch and Force, 2000).

Neofunctionalisation means the evolutionary divergence of duplicates into separate biological functions; an example could be that a gene coding for a gene implicated in fatty acid synthesis turns into two different genes involved in fatty acid synthesis as in (Carmona-Antoñanzas et al., 2016). Subfunctionalisation happens when the function of the ancestral gene becomes divided between the duplicates, as has famously happened with haemoglobin (Hardison, 2012), where the original protein has divided up its subunits but where both are necessary to optimise fitness. Which one of these fates is realised appears to depend on many (and rather unclear) circumstances, but the most common is the gradual loss of one copy. Asymmetric evolution, often interpreted as neofunctionalisation, appears to be common both when using sequence and regulatory divergence as metrics, while symmetric duplicate divergence as expected under the second-most frequent - subfunctionalisation model as a final fate appears rare (He and Zhang, 2005; Lien et al., 2016; Sandve et al., 2018). Despite novel functionality (controversies regarding the precise definition of biological “function” notwithstanding) being relatively rare, gene duplication has also been implicated in speciation through reproductive isolation (Lynch and Conery, 2000).

In Atlantic salmon, the process of pseudogenisation appears to be in progress (Gillard et al., 2021); while many gene pairs have indeed seen the complete loss of one copy, many more have simply evolved strongly reduced gene expression in one or both copies. This asymmetric loss of expression in one duplicate copy also tends to affect the expression across most tissues be

constitutive across tissues - that is, a decrease in expression in liver tends to coincide with a decrease in expression in other tissues as well. Additionally, it appears that most gene pairs have evolved asymmetrically; that is, one copy decreases much more than the other one. Genes corresponding to certain cellular processes are seen to be relatively quickly evolving, namely fatty acid metabolism; this seems to correspond to the evolutionary context of the salmonid ancestor following WGD.

1.3.4 Role of TEs following WGD

The role of repetitive DNA following whole-genome duplications as such has to my knowledge been a somewhat underexamined subject. As far as I can tell, there have been no systematic studies of the role of either TEs or of satDNA in resolving duplicate divergence or even of their general activity following the whole-genome duplication event. This chapter will, therefore, attempt to bring together some threads from a rather scattered literature and hopefully arrive at some coherent expectations for the roles and behaviour of repetitive DNA following whole-genome duplication. Therefore, in the following sections I will allow myself to be more speculative than what I have for most other parts of this treatise. Genome duplications, as well as repetitive DNA, remain major drivers of genome expansion (Marburger et al., 2018; Naville et al., 2019). Such drivers help address known issues such as the C-value paradox, i.e. the observation that organisms' complexity does not seem clearly correlated with their genome size (Elliott and Gregory, 2015). Although the value of this observation is limited - for one thing, it relies on gene number as a measure of complexity, a deeply problematic assumption in the world of alternative splicing - it does imply an evolutionary role of repetitive DNA in the context of genome expansion, just as is the case with whole-genome duplications.

Furthermore, since a whole-genome duplication is almost by necessity an evolutionary bottleneck event (and one with a lot of built-in redundancy, a topic to which we will return later in this treatise), strong founder effects will emerge. In the salmonids, the duplication was followed by a major expansion of TE content with a time lag. This indicates that the TE expansion was not a classical founder effect, though a relatively small effective population size should not be ruled out as creating a permissive environment for TE expansion. Similar patterns (in the sense that we observe a WGD followed by a delayed TE expansion) has been observed in other organisms (Feng et al., 2021; Giraud et al., 2021), and though no general pattern can be established from these findings this does seem like a possible venue for fruitful theoretical or in silico investigations; the sudden expansion of “extraneous” genetic content could very well allow for TE invasions or reactivation of dormant elements. Indeed, evolutionary stress and TE mobilisation is an active and intriguing field of inquiry (Fouché et al., 2019; Roquis et al., 2021).

Additionally, the normally very strong selection against TE insertions in or around genes may be relaxed following a whole-genome duplication. Since the most common fate of duplicated genes is nonfunctionalisation (Birchler and Yang, 2022; Ohno, 1970; Ohno et al., 1968), and transposable elements do tend to mess things up when introduced into or near a gene (manuscripts 1 and 3), it is conceivable that TE insertions play a role in resolution of gene duplicates through nonfunctionalisation. This could apply both to dramatic “killer” insertions and to more gradual but dosage-preserving promoter erosion, and be especially relevant for compensating for potential fitness reducing effects of increased gene dosage. Finally, TEs also play a role in rewiring post-duplication regulatory networks through distribution of cis-regulatory elements (Paper 3).

Where the role of TEs following whole-genome duplication and duplicate divergence is unclear, at least there has been some work done on it and some basis for informed speculation. This is not, to my knowledge, the case with satDNA. Our findings in Paper 2 indicate that there is indeed some relationship, but it remains quite obscure, and speculation on the subject appears almost unseemly. It does not help that the evolutionary roles of satDNA remain so poorly understood.

Despite the paucity of published works on the subject, there are many tantalising hints to the many roles played by TEs in duplicate divergence and their co-evolution with WGD events, and they will certainly be more closely examined in the years to come as the number of high quality genome sequences from ancient and more recent polyploid lineages increases.

1.4 Gene regulation

The gene concept is a surprisingly thorny one. Traditionally, a gene has been seen as a sequence of nucleic material which is interpreted by the biological apparatus to create some product which has an effect (Meunier, 2022; Portin and Wilkins, 2017). However, a gene cannot function without an extensive semiotic landscape of promoters, enhancers, and various regulatory elements. In this chapter, I will briefly summarise the workings of gene regulatory landscapes.

1.4.1 Genes and regulatory concepts

A gene has two main states: It can be expressed (“turned on”) or it can not be expressed (“turned off”). Both states are necessary for most genes. Even constitutive, “household” genes need to mobilise a significant machinery in order to maintain their function in the cellular context - that is, they are

subject to regulation (Silver et al., 2006). For a gene to be “turned on”, a great number of things have to coincide. Firstly, the chromatin structure itself must be relaxed enough that the gene is accessible to the cellular machinery (Bell et al., 2011; Buccitelli and Selbach, 2020). Second, that machinery (itself proteins, i.e., products of various other genes) must be recruited to the specific place on the genome where the gene is located (Buccitelli and Selbach, 2020). Finally, extensive modifications are made after the point of gene expression itself (Buccitelli and Selbach, 2020). In this treatise, I mainly concern myself with factors that are involved in modulating chromatin accessibility and interaction between transcriptional regulatory molecules and the DNA.

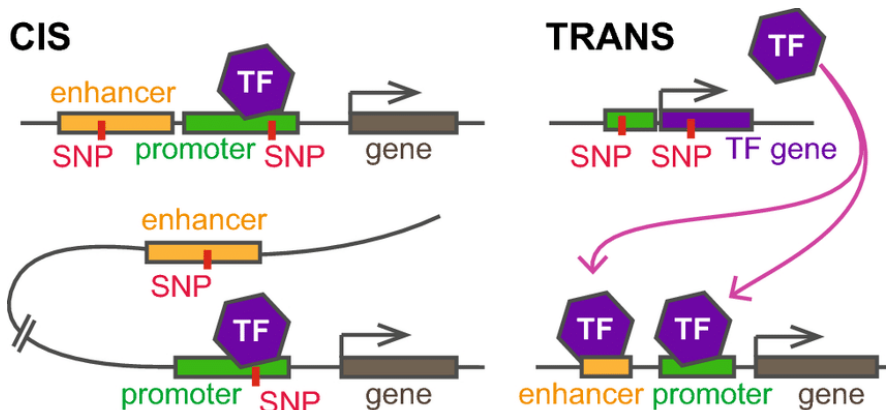


Fig. 4. Gene regulation, cis and trans. In cis-regulation, the regulatory element is physically proximate to the gene being regulated. In trans-regulation, the regulatory element can be arbitrarily far away, even on a different chromosome. This figure is not exhaustive, but included simply to get the general point across. Figure from (Ohnmacht et al., 2020), used under terms of <http://creativecommons.org/licenses/by/4.0/> licence.

Gene regulation by other gene products, called “transcription factors”, can be carried out in *trans* or in *cis*, meaning either regulating a nearby gene via

a product or directly (see Fig. 4). A cis-regulatory element is a genetic element which is physically close to the regulated gene, typically somewhat “upstream” of the transcription start site (Wittkopp et al., 2004). This distinction is complicated somewhat by the existence of distal enhancers (Agrawal et al., 2018) and the three-dimensional organisation of chromatin meaning that the distance between two genomic regions on the same chromosome within the cell nucleus is not always easy to predict (Mozziconacci et al., 2020). So, a cis-regulatory element does not actually need to be in or very near the core promoter in our flat representation of the genome in order to be acting in cis, and a trans-acting element can also be located quite close to the gene it impacts, or even on another chromosome altogether.

1.4.2 Duplication, regulation and transcription factors

Transcription factors bind to roughly conserved sequence fragments, “motifs”, on the chromosome. Typically, a motif is made up of a relatively short sequence (6-26bps) and has a spectrum of variation (Fig. 5). These motifs make up the core of one of the main forms of cis-regulatory element.

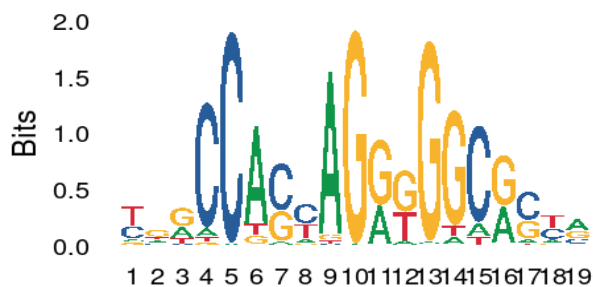


Fig. 5. An example of a logo representation of a motif, here the human CTCF motif. From the JASPAR database (Castro-Mondragon et al., 2022). Cited under terms of the <https://creativecommons.org/licenses/by-nc/4.0/> licence.

Because the number of transcription factors in the genome is fundamentally limited and the number of genes is likewise limited, building network graphs of regulatory activity is often a fruitful endeavour: If a single transcription factor is involved in regulating several genes, those genes will tend towards being co-expressed. When a whole-genome duplication event occurs, these networks double in nodes, but the number of connections multiplies by an order of magnitude, and so the complexity of the regulatory network immediately following whole-genome duplication is dramatically increased.

1.4.3 Chromatin and chromatin accessibility

Chromatin is the substance of the chromosome. It consists of coils-of-coils of DNA wrapped around large protein complexes known as nucleosomes, which consists of histone protein tetrameres. If the chromatin has a tight structure, the DNA is inaccessible to the cellular machinery and cannot be expressed; if it is loose, however, CREs can be accessed by gene regulatory molecules and transcription can happen.

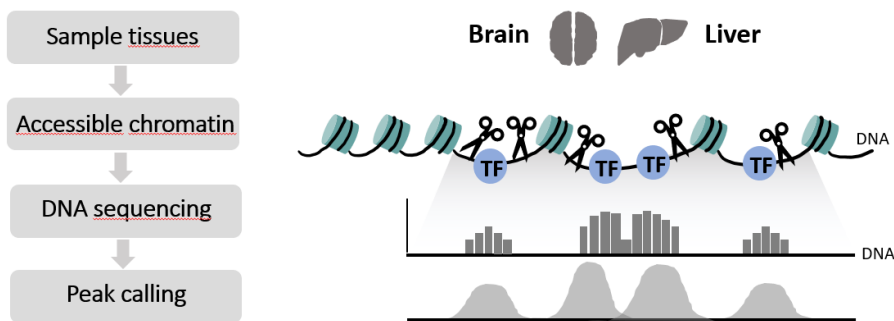


Fig. 6. A diagram of the ATAC-seq experimental process from tissue to usable data. *First, you take a tissue, then you extract and expose chromatin to a cutting transposase, you sequence the fragments and finally you call peaks based on some statistically based algorithm.*

The relative density of chromatin is itself regulated in trans by proteins attaching chemical modifications to histone “tails” protruding out of the nucleosomes; these chemical modifications also function as signals for the recruitment of transcription factors and other parts of the gene expression apparatus. Chromatin accessibility can thus be used as an indirect measure of the expression of genes in a given region, for instance using technologies such as ATAC-seq (Buenrostro et al., 2015) (Fig. 6). ATAC-seq is a method by which one effectively cuts all DNA not wrapped around a nucleosome to ribbons, determines the sequence of those ribbons and figures out whence they came to obtain a picture of which genes are transcriptionally available – and which regions are regulatorily relevant.

1.5 Genome sequencing technology unleash a repetitive-DNA revolution

1.5.1 Where we are in gene sequencing technology

Until recently, the relevant paradigm of genome sequencing was a massively parallel system (Buermans and den Dunnen, 2014; Heather and Chain, 2016). This involves splitting the DNA into smaller, more manageable fragments (size varies, but these days it is common to exceed 100bps in length), ligating an additional sequence called an “adapter” and then sequencing many millions of them simultaneously, at which point they are called “reads”. NGS sequencing has allowed for the assembly of hundreds of genomes and a whole host of ancillary technologies (RNA-seq (Emrich et al., 2007), ATAC-seq (Buenrostro et al., 2015), ChIP-seq (Park, 2009), etc. - it’s a very long list for a multitude of uses) which remain in heavy use throughout

the “-omics” sciences. However, it has a number of problems: For one thing, sequencing is an inherently error-prone process, and individual base-pairs can easily be mis-represented. In addition, the relatively short length of individual reads means that reads from different genomic regions that contain repetitive DNA have very similar or identical sequences giving rise to assembly collapse and the “multi-mapping” problem. Finally, the parallel nature of the technology means that reads have no known “anchor” point and must be assembled or interpreted without knowing anything about their position. Advances such as paired-end reads, longer read lengths, higher quality and various clever bioinformatics methods have ameliorated these issues, but they remain fundamental and to some extent insurmountable attributes of the technological paradigm.

1.5.2 A new generation of sequencing

A more recent development has been third-generation long-read sequencing methods, a set of technologies which have already been taken into wide usage for genome assembly. The two main contenders in this field (Oxford Nanopore Technologies (Jain et al., 2016) and Pacific Biosciences (Rhoads and Au, 2015)) use somewhat different technologies, but they both amount to removing the need for deliberately fragmenting reads down to predetermined lengths by doing sequencing in a very different way.

ONP sequencing is done by squeezing DNA through a very small hole.

Effectively what happens is that a specially-designed pore is inserted into an electrically resistant biological membrane with a current differential, and a special motor protein is used to guide a DNA fragment through the pore in a controlled manner. This current is measured by a sensor, and when a certain nucleotide passes through the pore, it causes a disruption to the current which can be interpreted to deduce what passed through. ONP reads can therefore in principle be as long as the full chromosome being sequenced,

but in practice does not usually exceed a few hundred thousand bases due to pores being clogged, motor proteins detaching, DNA breaking etc. (Wang et al., 2021; Zhang et al., 2022)

Pacific Biosciences' HiFi sequencing works by extracting double-stranded DNA and ligating adapters onto the ends which circularises it. It then sequences individual circularised DNA molecules in so-called "zero-mode waveguides" using a polymerase binding to the adapter and labelled nucleotides. As these labelled nucleotides are incorporated, they emit light of different wavelengths, which is detected and interpreted by a sensor apparatus connected to the zero-mode waveguide. As this goes on, the DNA is sequenced over several iterations, yielding a read which has the same sequence repeated several times - HiFi reads. These reads are highly accurate, since they can accommodate any issues in base calling by majority rules, and are comparably long to ONT reads (PacBio, 2022).

While the two technologies are quite distinct, they are very comparable in that they are producing incomparably longer reads than previous technology, greatly ameliorating the issues discussed in 1.5.1. They are both reaching wide usage, and many of their shortcomings are being addressed, e.g. for ONP reads with higher error rates, "polishing" with now-old-fashioned Illumina reads can be done (Weirather et al., 2017), especially when the sequencing coverage is not high enough (need >80x) to correct these single read errors in the consensus genome assembly (Wang et al., 2021; Zhang et al., 2022).

1.5.3 Return of the repeats

These new technologies have had a tremendous impact on genome assembly and especially on the assembly of repeat-rich parts of the genome. As we find in Paper 3, and as has been noted elsewhere e.g. in (Mérot et al., 2022), the issue of genome assembly from reads is much more easily handled when one has very long, reliable reads. We can simply sequence our way through regions which would previously have caused our assembly to collapse, and this is reflected in a dramatically reduced number of contigs in long-read based assemblies (Amarasinghe et al., 2020).

Those regions are disproportionately repeat-heavy due to the mode by which assemblies are generated from reads by assembly software. This means that we see marked improvements everywhere from the most studied species (e.g. (Nurk et al., 2022)) to systems which are very much non-model species (Mérot et al., 2022). The difference between our new Atlantic Salmon assembly discussed in Paper 3 and the old one (Lien et al., 2016) is stark despite being only a few years apart and the latter having incomparable amounts of resources dedicated to it. The difference must be attributed to this advance in technology.

2 Summary of results

2.1 Paper 1 - *Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication*

In Paper 1 we investigate gene expression evolution following whole genome duplication. We found that many ohnologues had asymmetrical expression divergence, most often with one copy evolving lower expression over time. Typically, the lower-expressed ohnologue would also have lower expression levels across many tissues.

We found evidence for TEs playing a dual role in gene regulatory evolution. Firstly, TE accumulation was clearly associated with asymmetric loss of expression level in one ohnologue copy (Figure 1h). This finding supports the idea of TEs being agents of destruction, i.e. that TEs drive nonfunctionalisation (i.e. slow pseudogenization process). This is most likely a neutral process for most genes, with functional redundancy being high in a duplicated genome. However, it is also possible that some TE accumulation in promoters could have been adaptive, as we identify signatures of adaptive evolution of reduced gene expression among duplicated genes, likely due to selection on gene dosage. Conversely, in a few cases where one ohnologue copy had evolved liver specific increase of gene expression, this was associated with gains in TF binding sites for liver-biassed TFs in the promoter, and some of these TF-binding sites were found within TEs. This gives credence to the idea that TEs play a role as CRE donors and also drive ohnolog divergence by gain of tissue specific CREs.

Our findings support a model of pervasive gene expression evolution following WGD to overcome immediate negative fitness consequences from having a double genome. We propose that TE-activity that disrupted existing CRE-landscapes could have been one mechanism to evolve new 'optimal' gene expression phenotypes following the WGD.

2.2 Paper 2 - *The role of transposable elements in the evolution of cis-regulatory element landscapes after whole genome duplication*

In paper 2, we ask what role TEs have had on CRE evolution, and in particular at the time following the salmonid WGD. We do this by first quantifying chromatin accessibility (from ATAC-seq) and use regions of elevated chromatin accessibility as a proxy for cis-regulatory elements (enhancer/promoter activity). Next, we integrate this tissue specific knowledge of CREs from liver and brain tissue with TE annotations to better understand the contribution of these genomic parasites to genome regulatory evolution.

Firstly, we find that TEs are underrepresented in regions of accessible chromatin and immediately downstream of transcription start sites, suggesting that our method can indeed detect signals of selection (Figure 3). TEs are significantly underrepresented in CREs in both tissues, but there are more TE-CREs in liver than in brain (Figure 2). A minority (~18%) of TE-CREs are shared between the tissues, and these are more often located in promoter regions.

Furthermore we find that a few TE subfamilies (about 11%) are enriched in genomic regions with accessible chromatin, and can be characterised as “superspreaders” of CREs. These superspreaders together have contributed to just under half of all CREs in TEs (Figure 3). Our estimates of sequence similarity to the subfamily consensus sequences suggests that the “superspreader” TEs are older than the TEs not enriched in open chromatin, and that tissue specific TE-CREs are younger than tissue shared TE-CREs (Figure 4C). Although “superspreaders” are relatively older, many appear to have been active in the time following whole-genome duplication (Figure 4B), hence their role in rewiring of regulatory networks post-WGD cannot be ruled out. But, the lack of a steep increase in super-spreader activity post-WGD certainly casts doubt about the hypothesis that TE-activity played a major role in expression evolution through deposition of novel CREs shortly after the WGD.

Many of the CRE “superspreader” TEs had uneven genomic representation of consensus sequence bases (Figure 5A). One explanation could be purifying selection on particular TE-CREs from particular TE subfamilies, resulting in only parts of the TEs being conserved through evolution. These long-lasting “remnants” could be parts of the TE with importance for the TE-activity (selfish gene selection) or specific CRE that are co-opted by the host genome. Our results also shed light on selection at the tissue level. We find substantially more TE-CREs in the liver compared to the brain (Figure 2) and TEs are depleted for TF binding sites for brain specific TFs (Figure 6). Both these findings support stronger purifying selection (i.e. to avoid TE-derived CREs) in the brain compared to the liver.

In conclusion, there is little evidence for WGD resulting in increased activity of TE-CRE evolution, but we find evidence that selection has shaped evolution of TE-CREs in several ways.

2.3 Paper 3 - *Structural variation landscape reflects telomeric tandem repeat expansions in Atlantic salmon*

This manuscript is part of a larger project on structural variation and comparative genomics in Atlantic salmon. Here we used new Nanopore long-read genome assembly technology to investigate the interplay between TEs and other repeat DNA, and structural variation.

The long-read-based assembly represents an increase in assembly quality by every measure. The number of contigs decreased by two orders of magnitude, among which was a BUSCO scores increased (92.3% to 95.7%), around 420Mb of additional sequence was included in the chromosome assemblies.

We annotated satellite DNA in two ways; one liberal approach and one more conservative, which required repeat motifs to be present in large scale or in large arrays to filter out local duplications and other small structures.

We found that the overwhelming majority of detected structural variation was made up of small SVs (less than 200bp, Figure 1 E-F). These small SV variants were especially strongly correlated with the presence of satDNA under the liberal annotation ($\rho=0.88$).

In particular, we found substantial overlap between the locations of the smaller structural variants and satDNA arrays (see Table 1; note that the bp-overlap is much higher than the proportion of numeric insertions), indicating that the variability is mostly due to repeat expansion and contraction. Curiously, our more liberal annotation of non-TE repeat DNA shows a much stronger affinity for these variants, possibly implying that relatively small tandem repeat structures are quite active. Finally, we found a strong tendency for SVs and non-TE repeats to adhere to telomeres (Figure

2D), including non-functional ancestral telomeres (Figure 3B; R^2 0.53 for both categories of telomere).

TEs had little overlap with structural variation with the exception of a specific PiggyBac family (spotted in a small insertion/deletion peak at around 1400bp in Figure 1E), which implies that there is little recent and/or current transposition in the Atlantic Salmon genome. Attempts at running extracted SV sequences through a TE classification pipeline yielded little result, further validating the notion that most repeat-associated variation in the Atlantic Salmon genome is not TE-based, and that the degree of this variation has hitherto been underestimated due to the technical limitations of short-read-based assemblies.

In conclusion, repeat DNA is important in the evolution of the complex SV landscape in Atlantic salmon genome, however it is the non-TE repeats that are the main source of novel structural variants.

3 Discussion

3.1 Shifting grounds - technological advances in genome assembly and research on repeat DNA

The scale of improvement in assembly quality following the introduction of third-generation sequencing technology is difficult to overstate. In Manuscript 3, we find that the Atlantic Salmon genome improves by every measure, going from 965,912 contigs to 4222 and with almost half a billion base pairs of additional DNA assembled around chromosomes. A large proportion of this is made up of repetitive DNA; as we discuss in that manuscript, it's mostly highly variable tandem repeats and satellite DNA. Our Nanopore reads, supplemented with Illumina polishing, demonstrated that we're in a new era of genome assemblies. Lower-quality nanopore base-calling may be an issue in discussing telomeric repeats in Paper 3 (as in (Tan et al., 2022)), but one for which the assembly has tried to compensate through Illumina polishing and various tuning. Our pipeline to call SVs has enough inbuilt cross-validation, and our SV definition is long enough (at least 50bp) that our observations do not seem fully attributable to base-calling errors, however. Nanopore-based assemblies also have fairly high precision in at least one benchmark study on structural variation (Dierckxsens et al., 2021).

As sequencing technologies continuously improve and sequencing costs have plummeted, we now see an explosion in the number of high quality genomes with more accurate representation of the genomic repeat-landscapes. Although these are exciting times for anyone with interests in repeat-DNA, it has also exacerbated the long-standing challenge associated

with repeat-DNA annotation. For example, with the enormous increase in assembled satellite repeats, it becomes clear that present methods for annotation and analysis of satellites of various kinds are insufficient. The best software available appears to be RepeatExplorer2 (Novák et al., 2020), which while a powerful and very useful tool is not capable of finding repeat motifs that make up less than a significant proportion of the DNA (0.01% under default settings - over 250kbs in our Atlantic Salmon assembly). Apart from this, the tools available are either very old and not made for the purpose (such as Tandem Repeat Finder (Benson, 1999)) or highly specialised (e.g. Straglr (Chiu et al., 2021)).

Another bottleneck is the annotation and classification of transposable elements. In our study system, even with a reasonably well curated TE library, the largest class of TEs are “unknown”. This is a recurrent issue in the literature, nicely highlighted by a recent study of ~600 insect genomes. This large scale comparative analyses of repeat DNA demonstrated very effectively how comparative TE landscape analyses becomes extremely biased when comparing well studied groups such as *Drosophila* with new genomes from non-model insect groups (Sproul et al., 2022) and serve to underline what remains perhaps the largest issue of TE annotation, namely the need for time-consuming and high-skilled manual curation.

3.2 Transposable elements; classification, annotation and a maturing field

The study of transposable elements is in some ways a more straightforward field than that of satDNA. There is a generally accepted system for classification of TEs based on mode of reproduction, structural features, common ancestry and finally sequence identity (Wicker et al., 2007) in that

order. This system has not been adopted without objection (Kapitonov and Jurka, 2008), and the way in which it is adopted is somewhat inconsistent. For instance, the most widely used database, RepBase (Bao et al., 2015), does not fully integrate the Wicker system of classification. We have chosen to attempt to implement the three-letter code classification system introduced by Wicker et al. (2007) (i.e. classifying along superfamily ID by designating first class, then order and finally superfamily with one letter each) and to adhere to it as closely as possible. However, since there is to my knowledge no central rule for the three-letter codes other than the original Wicker et al. (2007) study, we have sometimes had to improvise the third letter based on the name of the superfamily (e.g. Nimb elements become RIN). The division between superfamilies can also be somewhat ambiguous, and it is not always trivial to tell if a certain RepBase identifier is a superfamily of its own or a named family under another superfamily.

As can be surmised from the above, the TE field is somewhat fragmented and methods involved in e.g. manual curation and TE annotation have been heavily dependent on the individual researcher's preferences. This has obvious ramifications for our ability to rigorously compare findings between papers, even more so than the sort of definitional ambiguity which is relatively common in biology (e.g. (Keeling et al., 2019; Meunier, 2022)).

There are encouraging signs, however. A recent set of guidelines and handy tools for the manual curation of TEs (Goubert, 2021; Goubert et al., 2022) allows for standardisation of the production of TE libraries for annotation, and there have been initiatives made to try and dissociate from the RepBase database structure into something less temperamental - and even more importantly, open source and free for all (Amselem et al., 2019; TE Hub Consortium et al., 2021). As the TE-field matures, one may hope that these tendencies for co-ordination and standardisation bear further fruit. This

would make investigations such as what we do in Paper 2 much more straightforward, since there would be a more direct link to a wider literature. As it stands, our intuitions must be formulated and tested from a field in which there is a great deal of uncertainty. Improved classification tools, for example, might cast more light on the properties of specific superfamilies - or even, as we propose, specific subfamilies.

3.3 Evolutionary dynamics of transposable elements

That transposable elements spread through host genomes is very well known, and the mechanisms of their spread have also been subject of extensive study ((Feschotte and Pritham, 2007; Goodier, 2016; Wicker et al., 2007); issues pertaining to more specific such mechanisms are extensively discussed in (Craig et al., 2015)). Less well characterised, at least in non-model species, is how selective pressures shape the long term survival of TEs and to what extent they contribute to genome regulatory evolution. The distribution of TEs in any genome is the outcome of a co-evolutionary process involving the TEs and the host (Sultana et al., 2017). A successful TE must be able to multiply, but at the same time avoid large negative fitness effects for the host. Since most TE-insertions are not random (they have preference for different genomic contexts/sequences, e.g. see (Bourque et al., 2018; Sultana et al., 2017)) negative effects of TE-insertions on host fitness will drive evolution of less harmful TE insertion site preferences. In addition to this, very harmful insertions will be selected against and disappear from the population and species. Hence, a snapshot picture of the present day TE-landscape, as we have studied in this thesis, does provide information about selective pressure on TEs. For example, similar to what is found in certain well-studied eukaryotes (Miao et al., 2020; Wells and

Feschotte, 2020), we find strong evidence of significant depletion of TEs around promoter regions and exons, and that pattern is fairly consistent across superfamilies. This clearly demonstrates selection acting on the TE landscape. However, it is difficult to disentangle the effects of purifying selection following harmful insertions from insertion site biases which ultimately have evolved through a co-evolutionary arms race.

Genome duplications result in functional redundancy which potentially allows for liberal accumulation of regulatory mutations in one gene copy, if the other copy retains the ancestral function (Sandve et al., 2018). In Paper 1, we note that TE tends to accumulate in the promoter of the less expressed duplicate copy; this appears to be quite agnostic to TE superfamily (Paper 1, Figure S8). There has been some evidence to suggest that a duplicated genome is a generally more permissive environment for TE mobilisation (Ayala-Usma et al., 2021; Marburger et al., 2018), and our findings seem congruent with that notion. In addition, this tendency towards gradual nonfunctionalisation may offer a biological reason for such tolerance.

In mammals TE proliferation has resulted in the spread of functional CREs, which in turn is co-opted into the host genome regulatory network (Bourque et al., 2018; Sundaram and Wysocka, 2020; Sundaram et al., 2014). In Manuscript 2, we first identify tens of thousands of putative TE-CREs and show how many of these originate from specific parts of TEs. Although some of these cases could be accounted for by technical artefacts (e.g. consensus sequence errors and microsatellites contained within the TE sequence), most of them seem to be real (Manuscript 2, Figure 5). One interpretation of this observation is that many TE-CREs evolve through a “spread-and-select” model. By this, we mean that there are two main “phases” of TE-CRE evolution - first TEs disperse through the genome followed by heterogenous

negative selection pressure which erodes away the TE sequences with most deleterious effects on the host physiology. The conserved TE-CREs could then finally be co-opted into host regulatory networks. Under this model, there is an initial host-TE conflict during the proliferation phase, but once the TE has been rendered inactive this conflict gradually resolves in favour of the host. In the same manuscript we also note that several of these non-uniform shapes are disproportionately present in open chromatin. This is not trivial to interpret, but it may indicate that they are involved in regulation and constitute another signal of selection. Such interpretations rhyme somewhat with the implication of TEs as agents of rapid evolution and signals of evolvability in certain periods (Fablet and Vieira, 2011; Niu et al., 2019).

Our study into how TEs contribute to gene regulatory evolution is a first step to close the knowledge gap about the role of TEs in Atlantic salmon genome evolution, and in particular the links between WGD, TE activity, and evolution of genome regulation. One path to pursue this further is for example to integrate CRE-activity information from high-throughput reporter assays, combined with in-depth investigations into how different TEs spread specific transcription factor binding sites in more detail. We touch on this in Manuscript 2, but our approach is rather primitive, and some more creativity could almost certainly go a long way in this work; further work should therefore include analyses that link transcription factor binding motifs to a specific TE family and a specific region in the TE consensus with a functionally validated CRE. One could then start to chip at the hypothesis of the impact of TE proliferation on gene regulatory networks in a less indirect way.

3.4 Satellite DNA; status of the field, proliferation and effects

In contrast to the quickly maturing field of transposable elements, satellite DNA methodology remains somewhat haphazard. The most common way of annotating satDNA is by looking at the reference genome and simply detecting repeating tandem motifs with a certain tolerance for errors (Benson, 1999); this approach tends to be overly generous, and will capture local duplications and even certain genes (e.g. zinc finger genes with native repeating motifs). More recently, some progress has been made using graph-based approaches (Novák et al., 2017, 2020; Singchat et al., 2022), but the techniques involved are not fully mature and to my knowledge no approach has been developed which is not dependent on relative enrichment in the genome: As I touch upon in section 3.1, the present status of these methods involves taking overrepresented motifs from reads making up a certain proportion of the total data set. Consequently, any procedure for the general annotation of satDNA in a whole large genome is going to involve some compromises, (e.g. if we wish to include local, smaller repeats at the risk of including zinc finger proteins), which are mostly up to the discretion of the individual researcher. Many TE annotation software have rudimentary satellite detection tools implemented, but for looking specifically at satDNA these are insufficient (Shah et al., 2016); one approach is therefore to take a TE-like library approach where one finds a catalogue of satellite motifs and annotates those. We performed a version of this in Manuscript 3, which offered an interesting contrast to the simpler TRF-based analysis. While the TRF-based analysis finds more smaller repeats, the tendencies we observe with regard to telomericity and overlap remain the same. This indicates that, while we have likely captured a general property of satDNA, there is biological variance between the two annotation methods.

Satellite DNA tends to be relatively localised - that is, individual repeat motifs mostly stick to their own chromosome, and sometimes to their own part of the chromosome (Garrido-Ramos, 2017; Thakur et al., 2021). The clear exception to this is in duplicated regions, where the density of repeat landscapes correlates with sequence similarity, and where individual motifs can be seen to be shared among duplicated chromosome regions. This is interesting in light of the library hypothesis (Salser et al., 1976) of satellite distributions. As pre-duplication karyotypes are broken up and reshuffled (REF), forming new karyotypes, satellite arrays are located in new regions, where they can expand or contract based on their new circumstances. In this way, rediploidisation could allow for “re-seeding” of certain satellite motifs in novel locations; though we were not properly systematic about it, there were some indications of non-proportional expansion in ohnologous regions. We found no obvious evidence of de-novo emergence of repeats, but once again our investigation into this aspect was not structured enough to rule anything out.

As we show in Manuscript 3, (Manuscript 3, Figure 4 and Table 2), satDNA appears to be highly variable between populations and individuals; this could have interesting applications in population biology in the hands of someone interested and competent. The general distribution of variable satDNA is heavily dependent on closeness to telomeres (Manuscript 3, Figure 3b), notably including non-active ancestral telomeres; repeats in these are similarly variable to those in active telomeres. The most reasonable interpretation thus seems to be that the variability is a property of the repeats themselves, not of the specific telomeres.

One theoretically possible vector of satellite seeding is through being carried by transposable elements, in a manner similar to CREs. Recently, such TE-based seeding was indeed found to explain large variation in genome size in

squamates, even when the TE-content itself was mostly similar (Pasquesi et al., 2018). In our work with Manuscript 2, we found that some transposable elements do indeed contain satellite sequences, but we find very little evidence to suggest that this have led to large scale tandem repeat DNA increase. Indeed, as previously noted, we find a strong negative correlation between TE density and satDNA enrichment (Manuscript 3, Figure 3a). A tantalising hint we find for future inquiry is that highly repetitive regions are often also fairly similar between duplicates, implying that these regions may have continued recombining through rediploidisation. If this is the case, then satDNA may be much more central to the process of rediploidisation and chromosome rearrangements following WGD than previously assumed.

3.5 Future perspectives

The future of studies related to all forms of repetitive DNA is looking bright. As we realise the gains promised by our recent improvements in quality and cost reduction in sequencing technology, entire new vistas of scientific endeavour may open. In this chapter, I will discuss these in light of the work in this treatise.

In the TE field we can see some tendencies toward standardisation of methods and definitions, mentioned towards the end of section **3.2**. These should open up the possibility for more credible and less arduous comparative studies of individual superfamilies and -families of TEs and their impact on various kinds of evolution. One especially tantalising prospect is that of the direct examination of TE impact on comparative biology, which was originally a part of the work package for this Ph.D project, but which proved very difficult to implement. In Paper 2, we examine the impact of TEs to cis-regulatory evolution specifically in Atlantic

salmon - it is easy to imagine that tracing our “superspreader” elements through the salmonid clade and investigating their role in species specific insertion- and selection patterns could be of great interest. Likewise, it would be deeply illuminating to see studies on these elements performed in other organisms, to be able to properly evaluate whether the organism (or taxa) in question is a special case. Under present circumstances, such comparative work is generally the domain of large studies, such as (Sproul et al., 2022). Likewise, examining whether the same TEs are involved in the gradual down-regulation that we observe in Paper 1 in different species could offer insights into the mobilisation of TEs in gene regulation. One possible interpretation is that regulatorily rather uninvolved TEs such as Tc1/Mariner elements are so prevalent precisely because they’re rather inoffensive on their own (in Paper 2, we find them to be relatively depleted in CREs; per table S1, no automatically classified mariner sequences were found to be valid superspreaders) and so are not subject to strong negative selection. Thus, situations where the simple mass of TEs is a facet of regulatory evolution - such as promoter erosion - could tend to involve such elements. To find whether this is the case more generally, a clearer idea of the variation and prevalence of the elements in question would be extremely useful.

With satDNA the field is also very fragmented. It is facing a complete change in premises with the new assembly methods. Enormous amounts of previously very difficult genetic material can now be assembled in a relatively cost-efficient way, which combined with advances in graph-based genome analysis and our increasing appreciation of the importance of genome geometry poses a number of fascinating questions.

One thing which we somewhat touch upon in Paper 3 but do not properly answer is the role of satDNA in guiding ectopic recombination, and the role

of ectopic recombination in reintegrating genomes following whole-genome duplication. Another intriguing but not fully explored issue in Paper 3 is the variability of telomere-adjacent regions, including non-active telomeres. It seems reasonable to interpret this as these regions simply being especially repeat-heavy and thus having a lot of repeat-related variation, but it cannot be ruled out that there is something biological going on there; we tested various other hypotheses including density of repeats in a particular region and array size, and did not find anything which came close to the explanatory power of telomere adjacency. This may, once again, be because we did not look closely enough at the question, but it at least does bear investigating. For the most part, however, our work here only serves to underline that new questions may be posed, with the sheer increase in quality being our main finding; the implications of this increase seem somewhat out of the scope of this treatise.

Finally, the work presented here really does only scratch the surface of the developments in repDNA studies; fascinating work is being done in many fields including looking into the direct effect of satDNA on gene expression, TE mobilisation as a response to evolutionary stress, the life cycle of both satDNA and TEs, and much more. One cannot hope to investigate every particular in one work, but in the case of this field, almost every particular is interesting and awaits our attention. We now have the technology; we await the co-ordination of human will to exploit it.

Bibliography

Agrawal, P., Heimbruch, K.E., and Rao, S. (2018). Genome-Wide Maps of Transcription Regulatory Elements and Transcription Enhancers in Development and Disease. *Compr. Physiol.* 9, 439–455.

<https://doi.org/10.1002/cphy.c180028>.

Alexandrou, M.A., Swartz, B.A., Matzke, N.J., and Oakley, T.H. (2013). Genome duplication and multiple evolutionary origins of complex migratory behavior in Salmonidae. *Mol. Phylogenet. Evol.* *69*, 514–523. <https://doi.org/10.1016/j.ympev.2013.07.026>.

Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* *21*, 30. <https://doi.org/10.1186/s13059-020-1935-5>.

Amselem, J., Cornut, G., Choisne, N., Alaux, M., Alfama-Depauw, F., Jamilloux, V., Maumus, F., Letellier, T., Luyten, I., Pommier, C., et al. (2019). RepetDB: a unified resource for transposable element references. *Mob. DNA* *10*, 6. <https://doi.org/10.1186/s13100-019-0150-y>.

Andalis, A.A., Storchova, Z., Styles, C., Galitski, T., Pellman, D., and Fink, G.R. (2004). Defects arising from whole-genome duplications in *Saccharomyces cerevisiae*. *Genetics* *167*, 1109–1121. <https://doi.org/10.1534/genetics.104.029256>.

Arkhipova, I.R., and Yushenova, I.A. (2019). Giant transposons in eukaryotes: is bigger better? *Genome Biol. Evol.* *11*, 906–918. <https://doi.org/10.1093/gbe/evz041>.

Ayala-Usma, D.A., Cárdenas, M., Guyot, R., Mares, M.C.D., Bernal, A., Muñoz, A.R., and Restrepo, S. (2021). A whole genome duplication drives the genome evolution of *Phytophthora betacei*, a closely related species to *Phytophthora infestans*. *BMC Genomics* *22*, 795.

<https://doi.org/10.1186/s12864-021-08079-y>.

Baduel, P., Quadrana, L., Hunter, B., Bomblies, K., and Colot, V. (2019). Relaxed purifying selection in autopolyploids drives transposable element over-accumulation which provides variants for local adaptation. *Nat. Commun.* *10*, 5818. <https://doi.org/10.1038/s41467-019-13730-0>.

Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* *6*, 11. <https://doi.org/10.1186/s13100-015-0041-9>.

Bell, O., Tiwari, V.K., Thomä, N.H., and Schübeler, D. (2011). Determinants and dynamics of genome accessibility. *Nat. Rev. Genet.* *12*, 554–564. <https://doi.org/10.1038/nrg3017>.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* *27*, 573–580. <https://doi.org/10.1093/nar/27.2.573>.

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., Bento, P., Da Silva, C., Labadie, K., Alberti, A., et al. (2014). The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat. Commun.* *5*, 3657. <https://doi.org/10.1038/ncomms4657>.

Birchler, J.A., and Yang, H. (2022). The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. *Plant Cell* *34*, 2466–2474. <https://doi.org/10.1093/plcell/koac076>.

Biscotti, M.A., Olmo, E., and Heslop-Harrison, J.S.P. (2015). Repetitive

DNA in eukaryotic genomes. *Chromosome Res.* **23**, 415–420.

<https://doi.org/10.1007/s10577-015-9499-z>.

Bohutínská, M., Handrick, V., Yant, L., Schmickl, R., Kolář, F., Bomblies, K., and Paajanen, P. (2021). De Novo Mutation and Rapid Protein (Co-)evolution during Meiotic Adaptation in *Arabidopsis arenosa*. *Mol. Biol. Evol.* **38**, 1980–1994. <https://doi.org/10.1093/molbev/msab001>.

Bomblies, K. (2020). When everything changes at once: finding a new normal after genome duplication. *Proc. Biol. Sci.* **287**, 20202154. <https://doi.org/10.1098/rspb.2020.2154>.

Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., et al. (2018). Ten things you should know about transposable elements. *Genome Biol.* **19**, 199. <https://doi.org/10.1186/s13059-018-1577-z>.

Brynhildsen, W. (2016). In silico explorations of TE activity, diversity and abundance across 74 teleost fish species. Master thesis. University of Oslo.

Buccitelli, C., and Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. *Nat. Rev. Genet.* **21**, 630–644. <https://doi.org/10.1038/s41576-020-0258-4>.

Buckley, R.M., Kortschak, R.D., Raison, J.M., and Adelson, D.L. (2017). Similar evolutionary trajectories for retrotransposon accumulation in mammals. *Genome Biol. Evol.* **9**, 2336–2353. <https://doi.org/10.1093/gbe/evx179>.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-

seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* *109*, 21.29.1-21.29.9.

<https://doi.org/10.1002/0471142727.mb2129s109>.

Buermans, H.P.J., and den Dunnen, J.T. (2014). Next generation sequencing technology: Advances and applications. *Biochim. Biophys. Acta* *1842*, 1932–1941. <https://doi.org/10.1016/j.bbadis.2014.06.015>.

Campbell, M.A., Hale, M.C., McKinney, G.J., Nichols, K.M., and Pearse, D.E. (2019). Long-Term Conservation of Ohnologs Through Partial Tetrasomy Following Whole-Genome Duplication in Salmonidae. *G3 (Bethesda)* *9*, 2017–2028. <https://doi.org/10.1534/g3.119.400070>.

Carmona-Antoñanzas, G., Zheng, X., Tocher, D.R., and Leaver, M.J. (2016). Regulatory divergence of homeologous Atlantic salmon *elovl5* genes following the salmonid-specific whole-genome duplication. *Gene* *591*, 34–42. <https://doi.org/10.1016/j.gene.2016.06.056>.

Carretero-Paulet, L., and Van de Peer, Y. (2020). The evolutionary conundrum of whole-genome duplication. *Am. J. Bot.* *107*, 1101–1105. <https://doi.org/10.1002/ajb2.1520>.

Castro-Mondragon, J.A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R.B., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., et al. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* *50*, D165–D173. <https://doi.org/10.1093/nar/gkab1113>.

Chiu, R., Rajan-Babu, I.-S., Friedman, J.M., and Birol, I. (2021). Straglr: discovering and genotyping tandem repeat expansions using whole

genome long-read sequences. *Genome Biol.* 22, 224.

<https://doi.org/10.1186/s13059-021-02447-3>.

Christensen, K.A., Leong, J.S., Sakhrani, D., Biagi, C.A., Minkley, D.R., Withler, R.E., Rondeau, E.B., Koop, B.F., and Devlin, R.H. (2018). Chinook salmon (*Oncorhynchus tshawytscha*) genome and transcriptome. *PLoS ONE* 13, e0195461. <https://doi.org/10.1371/journal.pone.0195461>.

Craig, Chandler, Gellert, Lambowitz, Rice, and Sandmeyer, eds. (2015). *Mobile DNA III* (American Society of Microbiology) <https://doi.org/10.1128/9781555819217>.

Dierckxsens, N., Li, T., Vermeesch, J.R., and Xie, Z. (2021). A benchmark of structural variation detection by long reads through a realistic simulated model. *Genome Biol.* 22, 342. <https://doi.org/10.1186/s13059-021-02551-4>.

Dréau, A., Venu, V., Avdievich, E., Gaspar, L., and Jones, F.C. (2019). Genome-wide recombination map construction from single individuals using linked-read sequencing. *Nat. Commun.* 10, 4309. <https://doi.org/10.1038/s41467-019-12210-9>.

Elliott, T.A., and Gregory, T.R. (2015). What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370, 20140331. <https://doi.org/10.1098/rstb.2014.0331>.

Emrich, S.J., Barbazuk, W.B., Li, L., and Schnable, P.S. (2007). Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res.* 17, 69–73. <https://doi.org/10.1101/gr.5145806>.

Fablet, M., and Vieira, C. (2011). Evolvability, epigenetics and transposable elements. *Biomol. Concepts* 2, 333–341.

<https://doi.org/10.1515/BMC.2011.035>.

Feng, S., Liu, Z., Cheng, J., Li, Z., Tian, L., Liu, M., Yang, T., Liu, Y., Liu, Y., Dai, H., et al. (2021). Zanthoxylum-specific whole genome duplication and recent activity of transposable elements in the highly repetitive paleotetraploid *Z. bungeanum* genome. *Hortic. Res.* 8, 205.

<https://doi.org/10.1038/s41438-021-00665-1>.

Feschotte, C., and Pritham, E.J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.* 41, 331–368.

<https://doi.org/10.1146/annurev.genet.40.110405.090448>.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151, 1531–1545.

<https://doi.org/10.1093/genetics/151.4.1531>.

Fouché, S., Badet, T., Oggenfuss, U., Plissonneau, C., Francisco, C.S., and Croll, D. (2019). Stress-driven transposable element de-repression dynamics in a fungal pathogen. *BioRxiv* <https://doi.org/10.1101/633693>.

Garbus, I., Romero, J.R., Valarik, M., Vanžurová, H., Karafiátová, M., Cáccamo, M., Doležel, J., Tranquilli, G., Helguera, M., and Echenique, V. (2015). Characterization of repetitive DNA landscape in wheat homeologous group 4 chromosomes. *BMC Genomics* 16, 375.

<https://doi.org/10.1186/s12864-015-1579-0>.

Garrido-Ramos, M.A. (2017). Satellite DNA: an evolving topic. *Genes*

(Basel) 8. <https://doi.org/10.3390/genes8090230>.

Gemble, S., Wardenaar, R., Keuper, K., Srivastava, N., Nano, M., Macé, A.-S., Tijhuis, A.E., Bernhard, S.V., Spierings, D.C.J., Simon, A., et al. (2022). Genetic instability from a single S phase after whole-genome duplication. *Nature* 604, 146–151. <https://doi.org/10.1038/s41586-022-04578-4>.

Gilbert, C., and Cordaux, R. (2013). Horizontal transfer and evolution of prokaryote transposable elements in eukaryotes. *Genome Biol. Evol.* 5, 822–832. <https://doi.org/10.1093/gbe/evt057>.

Gillard, G.B. (2019). Evolution of gene expression following the whole genome duplication in salmonid fish (Norwegian University of Life Sciences, Ås).

Gillard, G.B., Grønvold, L., Røsæg, L.L., Holen, M.M., Monsen, Ø., Koop, B.F., Rondeau, E.B., Gundappa, M.K., Mendoza, J., Macqueen, D.J., et al. (2021). Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. *Genome Biol.* 22, 103. <https://doi.org/10.1186/s13059-021-02323-0>.

Giraud, D., Lima, O., Rousseau-Gueutin, M., Salmon, A., and Ainouche, M. (2021). Gene and transposable element expression evolution following recent and past polyploidy events in spartina (poaceae). *Front. Genet.* 12, 589160. <https://doi.org/10.3389/fgene.2021.589160>.

Gluck-Thaler, E., Ralston, T., Konkell, Z., Grabowski Ocampos, C., Devi Ganeshan, V., Dorrance, A.E., Niblack, T.L., Wood, C.W., Slot, J.C., Lopez-Nicora, H.D., et al. (2021). Giant *Starship* elements mobilize accessory genes in fungal genomes. *BioRxiv*

<https://doi.org/10.1101/2021.12.13.472469>.

Goodier, J.L. (2016). Restricting retrotransposons: a review. *Mob. DNA* 7, 16. <https://doi.org/10.1186/s13100-016-0070-z>.

Goubert, C. (2021). TE-Aid: Annotation helper tool for the manual curation of transposable element consensus sequences (Github: Clement Goubert).

Goubert, C., Craig, R.J., Bilat, A.F., Peona, V., Vogan, A.A., and Protasio, A.V. (2022). A beginner's guide to manual curation of transposable elements. *Mob. DNA* 13, 7. <https://doi.org/10.1186/s13100-021-00259-7>.

Grimholt, U. (2018). Whole genome duplications have provided teleosts with many roads to peptide loaded MHC class I molecules. *BMC Evol. Biol.* 18, 25. <https://doi.org/10.1186/s12862-018-1138-9>.

Gundappa, M.K., To, T.-H., Grønvold, L., Martin, S.A.M., Lien, S., Geist, J., Hazlerigg, D., Sandve, S.R., and Macqueen, D.J. (2022). Genome-Wide Reconstruction of Rediploidization Following Autopolyploidization across One Hundred Million Years of Salmonid Evolution. *Mol. Biol. Evol.* 39. <https://doi.org/10.1093/molbev/msab310>.

Hardison, R.C. (2012). Evolution of hemoglobin and its genes. *Cold Spring Harb. Perspect. Med.* 2, a011627.

<https://doi.org/10.1101/cshperspect.a011627>.

Heather, J.M., and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics* 107, 1–8.

<https://doi.org/10.1016/j.ygeno.2015.11.003>.

He, X., and Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169, 1157–1164.

<https://doi.org/10.1534/genetics.104.037051>.

Hollister, J.D., Arnold, B.J., Svedin, E., Xue, K.S., Dilkes, B.P., and Bomblies, K. (2012). Genetic adaptation associated with genome-doubling in autotetraploid *Arabidopsis arenosa*. *PLoS Genet.* 8, e1003093.

<https://doi.org/10.1371/journal.pgen.1003093>.

Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., et al. (2004). Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431, 946–957.

<https://doi.org/10.1038/nature03025>.

Jain, M., Olsen, H.E., Paten, B., and Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* 17, 239. <https://doi.org/10.1186/s13059-016-1103-0>.

Kapitonov, V.V., and Jurka, J. (2008). A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet.* 9, 411–412; author reply 414. <https://doi.org/10.1038/nrg2165-c1>.

Keeling, D.M., Garza, P., Nartey, C.M., and Carvunis, A.-R. (2019). The meanings of “function” in biology and the problematic case of de novo gene emergence. *ELife* 8. <https://doi.org/10.7554/eLife.47014>.

Lien, S., Koop, B.F., Sandve, S.R., Miller, J.R., Kent, M.P., Nome, T., Hvidsten, T.R., Leong, J.S., Minkley, D.R., Zimin, A., et al. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature* 533, 200–205. <https://doi.org/10.1038/nature17164>.

Loiseau, V., Peccoud, J., Bouzar, C., Guillier, S., Fan, J., Gueli Alletti, G., Meignin, C., Herniou, E.A., Federici, B.A., Wennmann, J.T., et al. (2021). Monitoring Insect Transposable Elements in Large Double-Stranded DNA Viruses Reveals Host-to-Virus and Virus-to-Virus Transposition. *Mol. Biol. Evol.* 38, 3512–3530. <https://doi.org/10.1093/molbev/msab198>.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155. <https://doi.org/10.1126/science.290.5494.1151>.

Lynch, M., and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154, 459–473. <https://doi.org/10.1093/genetics/154.1.459>.

Macqueen, D.J., Primmer, C.R., Houston, R.D., Nowak, B.F., Bernatchez, L., Bergseth, S., Davidson, W.S., Gallardo-Escárate, C., Goldammer, T., Guiguen, Y., et al. (2017). Functional Annotation of All Salmonid Genomes (FAASG): an international initiative supporting future salmonid research, conservation and aquaculture. *BMC Genomics* 18, 484. <https://doi.org/10.1186/s12864-017-3862-8>.

Marburger, S., Alexandrou, M.A., Taggart, J.B., Creer, S., Carvalho, G., Oliveira, C., and Taylor, M.I. (2018). Whole genome duplication and transposable element proliferation drive genome expansion in

Corydoradinae catfishes. *Proc. Biol. Sci.* 285.

<https://doi.org/10.1098/rspb.2017.2732>.

McDonald, J.F. (1993). Evolution and consequences of transposable elements. *Curr. Opin. Genet. Dev.* 3, 855–864.

[https://doi.org/10.1016/0959-437X\(93\)90005-A](https://doi.org/10.1016/0959-437X(93)90005-A).

Mérot, C., Stenløkk, K.S.R., Venney, C., Laporte, M., Moser, M., Normandeau, E., Árnýasi, M., Kent, M., Rougeux, C., Flynn, J.M., et al. (2022). Genome assembly, structural variants, and genetic differentiation between lake whitefish young species pairs (*Coregonus* sp.) with long and short reads. *Mol. Ecol.* <https://doi.org/10.1111/mec.16468>.

Meunier, R. (2022). Gene (Stanford Encyclopedia of Philosophy). .

Miao, B., Fu, S., Lyu, C., Gontarz, P., Wang, T., and Zhang, B. (2020). Tissue-specific usage of transposable element-derived promoters in mouse development. *Genome Biol.* 21, 255.

<https://doi.org/10.1186/s13059-020-02164-3>.

Mozziconacci, J., Merle, M., and Lesne, A. (2020). The 3D genome shapes the regulatory code of developmental genes. *J. Mol. Biol.* 432, 712–723.

<https://doi.org/10.1016/j.jmb.2019.10.017>.

Naseeb, S., Ames, R.M., Delneri, D., and Lovell, S.C. (2017). Rapid functional and evolutionary changes follow gene duplication in yeast.

Proc. Biol. Sci. 284. <https://doi.org/10.1098/rspb.2017.1393>.

Naville, M., Henriët, S., Warren, I., Sumic, S., Reeve, M., Volff, J.-N., and Chourrout, D. (2019). Massive Changes of Genome Size Driven by Expansions of Non-autonomous Transposable Elements. *Curr. Biol.* 29,

1161-1168.e6. <https://doi.org/10.1016/j.cub.2019.01.080>.

Niu, X.-M., Xu, Y.-C., Li, Z.-W., Bian, Y.-T., Hou, X.-H., Chen, J.-F., Zou, Y.-P., Jiang, J., Wu, Q., Ge, S., et al. (2019). Transposable elements drive rapid phenotypic variation in *Capsella rubella*. *Proc Natl Acad Sci USA* *116*, 6908–6913. <https://doi.org/10.1073/pnas.1811498116>.

Novák, P., Ávila Robledillo, L., Koblížková, A., Vrbová, I., Neumann, P., and Macas, J. (2017). TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Res.* *45*, e111. <https://doi.org/10.1093/nar/gkx257>.

Novák, P., Neumann, P., and Macas, J. (2020). Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. *Nat. Protoc.* *15*, 3745–3776. <https://doi.org/10.1038/s41596-020-0400-y>.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. *Science* *376*, 44–53. <https://doi.org/10.1126/science.abj6987>.

Ohnmacht, J., May, P., Sinkkonen, L., and Krüger, R. (2020). Missing heritability in Parkinson's disease: the emerging role of non-coding genetic variation. *J. Neural Transm.* *127*, 729–748. <https://doi.org/10.1007/s00702-020-02184-0>.

Ohno, S. (1970). *Evolution by Gene Duplication*.

Ohno, S., Wolf, U., and Atkin, N.B. (1968). Evolution from fish to mammals by gene duplication. *Hereditas* *59*, 169–187.

<https://doi.org/10.1111/j.1601-5223.1968.tb02169.x>.

PacBio (2022). How HiFi sequencing works <https://www.pacb.com/technology/hifi-sequencing/how-it-works/>.

Park, P.J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* *10*, 669–680.
<https://doi.org/10.1038/nrg2641>.

Pasquesi, G.I.M., Adams, R.H., Card, D.C., Schield, D.R., Corbin, A.B., Perry, B.W., Reyes-Velasco, J., Ruggiero, R.P., Vandewege, M.W., Shortt, J.A., et al. (2018). Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. *Nat. Commun.* *9*, 2774. <https://doi.org/10.1038/s41467-018-05279-1>.

Pelé, A., Rousseau-Gueutin, M., and Chèvre, A.-M. (2018). Speciation success of polyploid plants closely relates to the regulation of meiotic recombination. *Front. Plant Sci.* *9*, 907.
<https://doi.org/10.3389/fpls.2018.00907>.

Pezer, Z., Brajković, J., Feliciello, I., and Ugarkovć, D. (2012). Satellite DNA-mediated effects on genome regulation. *Genome Dyn.* *7*, 153–169.
<https://doi.org/10.1159/000337116>.

Portin, P., and Wilkins, A. (2017). The evolving definition of the term “gene”. *Genetics* *205*, 1353–1364.
<https://doi.org/10.1534/genetics.116.196956>.

Quesneville, H. (2020). Twenty years of transposable element analysis in the *Arabidopsis thaliana* genome. *Mob. DNA* *11*, 28.

<https://doi.org/10.1186/s13100-020-00223-x>.

Rhoads, A., and Au, K.F. (2015). Pacbio sequencing and its applications. *Genomics Proteomics Bioinformatics* 13, 278–289.

<https://doi.org/10.1016/j.gpb.2015.08.002>.

Richard, G.-F., Kerrest, A., and Dujon, B. (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol. Mol. Biol. Rev.* 72, 686–727. <https://doi.org/10.1128/MMBR.00011-08>.

Robertson, F.M., Gundappa, M.K., Grammes, F., Hvidsten, T.R., Redmond, A.K., Lien, S., Martin, S.A.M., Holland, P.W.H., Sandve, S.R., and Macqueen, D.J. (2017). Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. *Genome Biol.* 18, 111. <https://doi.org/10.1186/s13059-017-1241-z>.

Roquis, D., Robertson, M., Yu, L., Thieme, M., Julkowska, M., and Bucher, E. (2021). Genomic impact of stress-induced transposable element mobility in *Arabidopsis*. *Nucleic Acids Res.* 49, 10431–10447.

<https://doi.org/10.1093/nar/gkab828>.

Ruiz-Ruano, F.J., López-León, M.D., Cabrero, J., and Camacho, J.P.M. (2016). High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Sci. Rep.* 6, 28333. <https://doi.org/10.1038/srep28333>.

Salser, W., Bowen, S., Browne, D., el-Adli, F., Fedoroff, N., Fry, K., Heindell, H., Paddock, G., Poon, R., Wallace, B., et al. (1976). Investigation of the organization of mammalian chromosomes at the DNA sequence level. *Fed. Proc.* 35, 23–35.

Sandve, S.R., Rohlfs, R.V., and Hvidsten, T.R. (2018). Subfunctionalization versus neofunctionalization after whole-genome duplication. *Nat. Genet.* *50*, 908–909. <https://doi.org/10.1038/s41588-018-0162-4>.

Schmid, M., Evans, B.J., and Bogart, J.P. (2015). Polyploidy in Amphibia. *Cytogenet. Genome Res.* *145*, 315–330. <https://doi.org/10.1159/000431388>.

Shah, A.B., Schielzeth, H., Albersmeier, A., Kalinowski, J., and Hoffman, J.I. (2016). High-throughput sequencing and graph-based cluster analysis facilitate microsatellite development from a highly complex genome. *Ecol. Evol.* *6*, 5718–5727. <https://doi.org/10.1002/ece3.2305>.

Shatskikh, A.S., Kotov, A.A., Adashev, V.E., Bazylev, S.S., and Olenina, L.V. (2020). Functional significance of satellite dnas: insights from drosophila. *Front. Cell Dev. Biol.* *8*, 312. <https://doi.org/10.3389/fcell.2020.00312>.

Silver, N., Best, S., Jiang, J., and Thein, S.L. (2006). Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. *BMC Mol. Biol.* *7*, 33. <https://doi.org/10.1186/1471-2199-7-33>.

Singchat, W., Ahmad, S.F., Jaisamut, K., Panthum, T., Ariyaraphong, N., Kraichak, E., Muangmai, N., Duengkae, P., Payungporn, S., Malaivijitnond, S., et al. (2022). Population Scale Analysis of Centromeric Satellite DNA Reveals Highly Dynamic Evolutionary Patterns and Genomic Organization in Long-Tailed and Rhesus Macaques. *Cells* *11*. <https://doi.org/10.3390/cells11121953>.

Sotero-Caio, C.G., Platt, R.N., Suh, A., and Ray, D.A. (2017). Evolution and

diversity of transposable elements in vertebrate genomes. *Genome Biol. Evol.* *9*, 161–177. <https://doi.org/10.1093/gbe/evw264>.

Sproul, J.S., Hotaling, S., Heckenhauer, J., Powell, A., Larracuente, A.M., Kelley, J.L., Pauls, S.U., and Frandsen, P.B. (2022). Repetitive elements in the era of biodiversity genomics: insights from 600+ insect genomes. *BioRxiv* <https://doi.org/10.1101/2022.06.02.494618>.

Stitzer, M.C., Anderson, S.N., Springer, N.M., and Ross-Ibarra, J. (2021). The genomic ecosystem of transposable elements in maize. *PLoS Genet.* *17*, e1009768. <https://doi.org/10.1371/journal.pgen.1009768>.

Storchová, Z., Breneman, A., Cande, J., Dunn, J., Burbank, K., O'Toole, E., and Pellman, D. (2006). Genome-wide genetic analysis of polyploidy in yeast. *Nature* *443*, 541–547. <https://doi.org/10.1038/nature05178>.

Sultana, T., Zamborlini, A., Cristofari, G., and Lesage, P. (2017). Integration site selection by retroviruses and transposable elements in eukaryotes. *Nat. Rev. Genet.* *18*, 292–308. <https://doi.org/10.1038/nrg.2017.7>.

Sundaram, V., and Wysocka, J. (2020). Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* *375*, 20190347. <https://doi.org/10.1098/rstb.2019.0347>.

Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., Snyder, M.P., and Wang, T. (2014). Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res.* *24*, 1963–1976. <https://doi.org/10.1101/gr.168872.113>.

Tan, K.-T., Slevin, M.K., Meyerson, M., and Li, H. (2022). Identifying and correcting repeat-calling errors in nanopore sequencing of telomeres. *Genome Biol.* *23*, 180. <https://doi.org/10.1186/s13059-022-02751-6>.

TE Hub Consortium, Elliott, T.A., Heitkam, T., Hubley, R., Quesneville, H., Suh, A., and Wheeler, T.J. (2021). TE Hub: A community-oriented space for sharing and connecting tools, data, resources, and methods for transposable element annotation. *Mob. DNA* *12*, 16. <https://doi.org/10.1186/s13100-021-00244-0>.

Thakur, J., Packiaraj, J., and Henikoff, S. (2021). Sequence, chromatin and evolution of satellite DNA. *Int. J. Mol. Sci.* *22*. <https://doi.org/10.3390/ijms22094309>.

Van de Peer, Y., Mizrachi, E., and Marchal, K. (2017). The evolutionary significance of polyploidy. *Nat. Rev. Genet.* *18*, 411–424. <https://doi.org/10.1038/nrg.2017.26>.

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., and Au, K.F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* *39*, 1348–1365. <https://doi.org/10.1038/s41587-021-01108-x>.

Weirather, J.L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., Buck, D., and Au, K.F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. [version 2; peer review: 2 approved]. *F1000Res.* *6*, 100. <https://doi.org/10.12688/f1000research.10571.2>.

Wells, J.N., and Feschotte, C. (2020). A field guide to eukaryotic transposable elements. *Annu. Rev. Genet.* *54*, 539–561.

<https://doi.org/10.1146/annurev-genet-040620-022145>.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 8, 973–982. <https://doi.org/10.1038/nrg2165>.

Wicker, T., Schulman, A.H., Tanskanen, J., Spannagl, M., Twardziok, S., Mascher, M., Springer, N.M., Li, Q., Waugh, R., Li, C., et al. (2017). The repetitive landscape of the 5100 Mbp barley genome. *Mob. DNA* 8, 22. <https://doi.org/10.1186/s13100-017-0102-3>.

Wittkopp, P.J., Haerum, B.K., and Clark, A.G. (2004). Evolutionary changes in cis and trans gene regulation. *Nature* 430, 85–88. <https://doi.org/10.1038/nature02698>.

Wright, K.M., Arnold, B., Xue, K., Šurinová, M., O’Connell, J., and Bomblies, K. (2015). Selection on meiosis genes in diploid and tetraploid *Arabidopsis arenosa*. *Mol. Biol. Evol.* 32, 944–955. <https://doi.org/10.1093/molbev/msu398>.

Zhang, H.-H., Peccoud, J., Xu, M.-R.-X., Zhang, X.-G., and Gilbert, C. (2020). Horizontal transfer and evolution of transposable elements in vertebrates. *Nat. Commun.* 11, 1362. <https://doi.org/10.1038/s41467-020-15149-4>.

Zhang, L., Reifová, R., Halenková, Z., and Gompert, Z. (2021). How important are structural variants for speciation? *Genes (Basel)* 12. <https://doi.org/10.3390/genes12071084>.

Zhang, X., Liu, C.-G., Yang, S.-H., Wang, X., Bai, F.-W., and Wang, Z. (2022).

Benchmarking of long-read sequencing, assemblers and polishers for yeast genome. *Brief. Bioinformatics* 23.
<https://doi.org/10.1093/bib/bbac146>.


PAPER 1

RESEARCH

Open Access

Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication



Gareth B. Gillard^{1†}, Lars Grønvold^{2†}, Line L. Røsæg², Matilde Mengkrog Holen², Øystein Monsen², Ben F. Koop³, Eric B. Rondeau³, Manu Kumar Gundappa⁴, John Mendoza⁵, Daniel J. Macqueen⁴, Rori V. Rohlf⁶, Simen R. Sandve^{2*} and Torgeir R. Hvidsten^{1*} 

* Correspondence: simen.sandve@nmbu.no; torgeir.r.hvidsten@nmbu.no

[†]Gareth B. Gillard and Lars Grønvold contributed equally to this work.

²Center for Integrative Genetics, Department of Animal and Aquacultural Sciences, Faculty of Biosciences, Norwegian University of Life Sciences, Ås, Norway

¹Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås, Norway
Full list of author information is available at the end of the article

Abstract

Background: Whole genome duplication (WGD) events have played a major role in eukaryotic genome evolution, but the consequence of these extreme events in adaptive genome evolution is still not well understood. To address this knowledge gap, we used a comparative phylogenetic model and transcriptomic data from seven species to infer selection on gene expression in duplicated genes (ohnologs) following the salmonid WGD 80–100 million years ago.

Results: We find rare cases of tissue-specific expression evolution but pervasive expression evolution affecting many tissues, reflecting strong selection on maintenance of genome stability following genome doubling. Ohnolog expression levels have evolved mostly asymmetrically, by diverting one ohnolog copy down a path towards lower expression and possible pseudogenization. Loss of expression in one ohnolog is significantly associated with transposable element insertions in promoters and likely driven by selection on gene dosage including selection on stoichiometric balance. We also find symmetric expression shifts, and these are associated with genes under strong evolutionary constraints such as ribosome subunit genes. This possibly reflects selection operating to achieve a gene dose reduction while avoiding accumulation of “toxic mutations”. Mechanistically, ohnolog regulatory divergence is dictated by the number of bound transcription factors in promoters, with transposable elements being one likely source of novel binding sites driving tissue-specific gains in expression.

Conclusions: Our results imply pervasive adaptive expression evolution following WGD to overcome the immediate challenges posed by genome doubling and to exploit the long-term genetic opportunities for novel phenotype evolution.



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Whole genome duplication (WGD) events have played a major role in eukaryotic evolution by increasing genomic complexity and functional redundancy [1]. This can allow gene duplicates (referred to as ohnologs) to escape selective constraints and thereby accumulate previously forbidden mutations that may become adaptive [2]. In agreement with this idea, WGD has been associated with the evolution of adaptive traits in yeast [3], plants [4, 5], and vertebrates [6–8]. At the same time, it is also evident that most polyploids go extinct shortly after formation [9] and that becoming a successful new polyploid likely requires new adaptations to overcome fitness costs stemming from having a doubled genome [10, 11]. Yet, the importance of selection in shaping polyploid genome evolution in the aftermath of WGDs is still not well understood [1, 12].

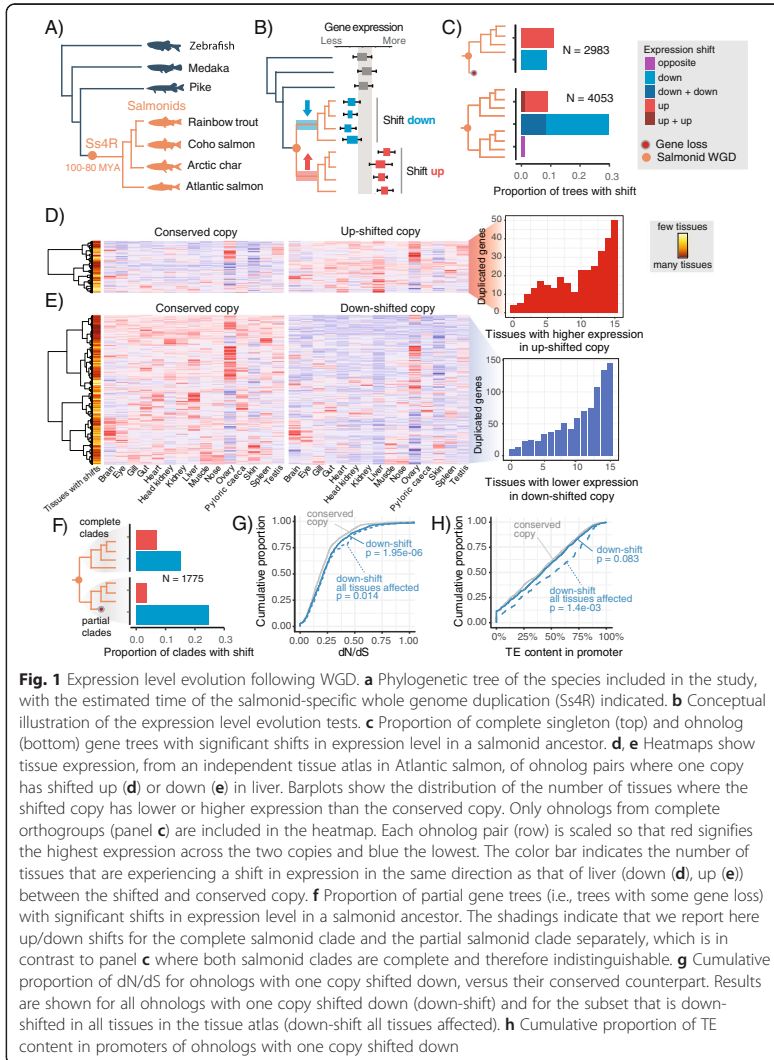
Gene expression levels are relatively easy to measure and compare, and represent a major source of complex trait variation [13] and novel adaptive phenotypes [14, 15]. Hence, there has been substantial interest in understanding consequences of WGDs on gene regulatory evolution. Comparative transcriptomics has both revealed immediate plastic responses to adjust gene dosages [16], as well as widespread regulatory divergence at evolutionary timescales (e.g., [17–20]). Ohnolog regulatory evolution is also mostly asymmetric, with one copy retaining an ancestral-like regulation, and the other copy losing and/or gaining expression in one or more tissue [12]. Although this observation can be reconciled with adaptive evolution of gene regulatory phenotypes following WGD, methodological limitations have made it difficult to distinguish between the outcomes of selection and neutral drift [12, 21].

Here we take a novel approach to improve our understanding of how selection shapes novel gene regulatory phenotypes following WGD. We first developed a flexible and user friendly version of a phylogenetic Ornstein-Uhlenbeck (OU) model of gene expression evolution [22, 23] in R (<https://gitlab.com/sandve-lab/evemodel>). The crux of this model is that it allows us to evaluate if changes in expression evolution deviate from the null hypothesis of stabilizing selection, and thereby identify putative adaptive shifts in expression regulation. We then used this model to analyze the liver transcriptome of four salmonids and three non-salmonid fish species to assess the impact of the 80–100-million-year-old salmonid-specific WGD (Ss4R) [24, 25]. We find that this WGD led to a burst of gene expression evolution, leading to rare tissue-specific gains in expression and pervasive tissue non-specific dosage selection, reflecting both adaptive possibilities afforded by genome doubling and immediate challenges that must be overcome to succeed as a polyploid lineage.

Results

Adaptive shifts in expression levels following WGD

To study expression level evolution following WGD, we generated RNA-seq datasets from livers (four biological replicates) of four salmonids and three non-salmonid outgroup species (Fig. 1a). We then computed gene trees to identify retained ohnologs from the salmonid WGD. In total, we included 10,154 gene trees in our analyses (Additional file 1: Figure S1), of which 65% (6689 trees) contained ohnologs derived from the salmonid WGD. For each gene tree, we then applied a phylogenetic Ornstein-Uhlenbeck (OU) process model to test for adaptive shifts in expression evolution



(referred to simply as ‘shifts’) in the ancestor of the salmonids included in this study (Fig. 1b, Additional file 1: Figures S2, S3 and S4).

Two major observations arise from this analysis. First, it is evident that the rate of adaptive gene expression evolution is increased for salmonid ohnologs. Forty percent of trees (1649) with retained ohnologs display evolution of novel expression levels in at least one ohnolog compared to only 20 % of trees with a single copy gene (Fig. 1c). Secondly, there is a clear difference in the nature of the expression evolution between ohnologs and singleton genes. Ohnologs are strongly biased towards evolving decreased expression levels following WGD (Fig. 1c), with 75% (1234/1649) of the ohnolog pairs displaying a shift down in either one or both copies. Conversely, singletons show a

small bias towards evolving increased expression (Fig. 1c). This difference could not be explained by differences in statistical power related to systematic differences in gene expression levels between singletons and ohnologs (Additional file 1: Figure S5).

To test if the identified expression level shifts following WGD were tissue-specific, we analyzed RNA-seq data from 15 Atlantic salmon tissues (Additional file 1: Figure S6A). We find that most cases of expression evolution are not liver-specific (Fig. 1d, e), and that this is true both for genes evolving increased and decreased expression following WGD. When one ohnolog copy had evolved a shift in liver expression level, this copy also displayed similar trends in the majority of the other 14 tissues compared to its conserved ohnolog partner (shift down 77% (682/885), shift up 70% (221/317)). Hence, evolution of liver-specific changes in ohnolog expression following WGD is rare, irrespective of the directionality of change.

Upon reaching a new optimal ohnolog gene dosage, the expectation is that the copy with the highest expression level contributes the most to the proteome and cell function, which will result in reduced purifying selection pressure on the more lowly expressed copy [26]. Several lines of evidence support this expectation. Firstly, species-specific gene loss events (expected for genes evolving under relaxed selection) are associated with increased probability of evolving lower liver expression in one copy (Fig. 1f) and with increased probability of the down-shifted copy to have reduced expression levels across all the other 14 tissues (Fisher's exact test, $p = 3.1e-07$, Additional file 1: Figure S6B). Secondly, we find that the down-shifted copy shows increased signatures of relaxed purifying selection on coding sequences in the form of elevated dN/dS rates (Fig. 1g, $p = 2.1e-6$, $N = 732$, one-sided paired Wilcoxon test, Additional file 1: Figure S7). Lastly, we also observe that down-shifted ohnolog copies have a significantly higher load of potentially destructive transposable element (TE) insertions in promoters compared to the conserved partner (Fig. 1h, one-sided paired Wilcoxon test, $p = 6.5e-4$, Additional file 1: Figure S8). Importantly, the effect size of increased dN/dS and TE-load were larger when only considering ohnologs with signatures of down-shift across all tissues (Fig. 1g, h).

Pervasive differences in purifying selection pressure within individual ohnolog pairs raise the question of whether these ohnologs might belong to duplicated genome blocks experiencing large-scale differences in selective constraints. This could lead to uneven ohnolog loss rates, a process that is referred to as biased fractionation [27]. In line with previous studies on teleosts [28, 29], we found significant biases in local gene loss, albeit only in 9 of 47 syntenic duplicate blocks. However, we did not find equivalent large-scale biases in expression loss (Additional file 1: Figure S9), thus rendering regional differences in selection constraints an unlikely explanation for the large number of ohnologs experiencing loss of expression in one copy.

In conclusion, we find widespread signatures of adaptive regulatory evolution in retained ohnologs following WGD; however, most adaptive events were associated with ohnolog gene dose reduction across many tissues. Thus, ohnolog copies that evolve lower expression levels compared to their partner continue to evolve under relaxed purifying selection pressure, following a likely path towards pseudogenization.

Strong selection on housekeeping gene dose after WGD

To test if selection on gene regulation following WGD was linked to particular cellular functions or pathways, we performed KEGG enrichment analyses for two ohnolog gene sets that had evolved either increased (up) or decreased (down) expression levels. Genes with increased expression level were enriched (Fisher's exact test, $p < 0.05$) in three pathways: "fatty acid elongation," "fatty acid metabolism," and the "cell cycle" (Additional file 1: Table S1). Detailed analysis identified 29 up-shifted genes encoding proteins with essential cell division functions. These genes were highly enriched in protein-protein interactions conserved in both unicellular and multicellular eukaryotes (Additional file 1: Table S2, Additional file 1: Figure S10) and suggest compensatory regulatory adaptation to maintain a functional cell division and ensure genome stability.

Down-shifted genes had comparatively stronger functional signatures (Additional file 1: Table S1) with nine enriched pathways (Fisher's exact test, $p < 0.05$). The three pathways with the strongest enrichment were "oxidative phosphorylation" ($p = 0.003$) involved in mitochondrial-associated cellular energy production, "ribosome biogenesis in eukaryotes" ($p = 0.008$) which consists of genes involved in assembly of the ribosome, and "ribosome" ($p = 5.6e-9$) which consists of ribosomal subunit genes (Supplementary figures 11, 12 and 13). These results support strong selection on gene dosage for many housekeeping functions following WGD, which aligns well with our observation (Fig. 1d, e) that most expression level shifts occurred across most tissues.

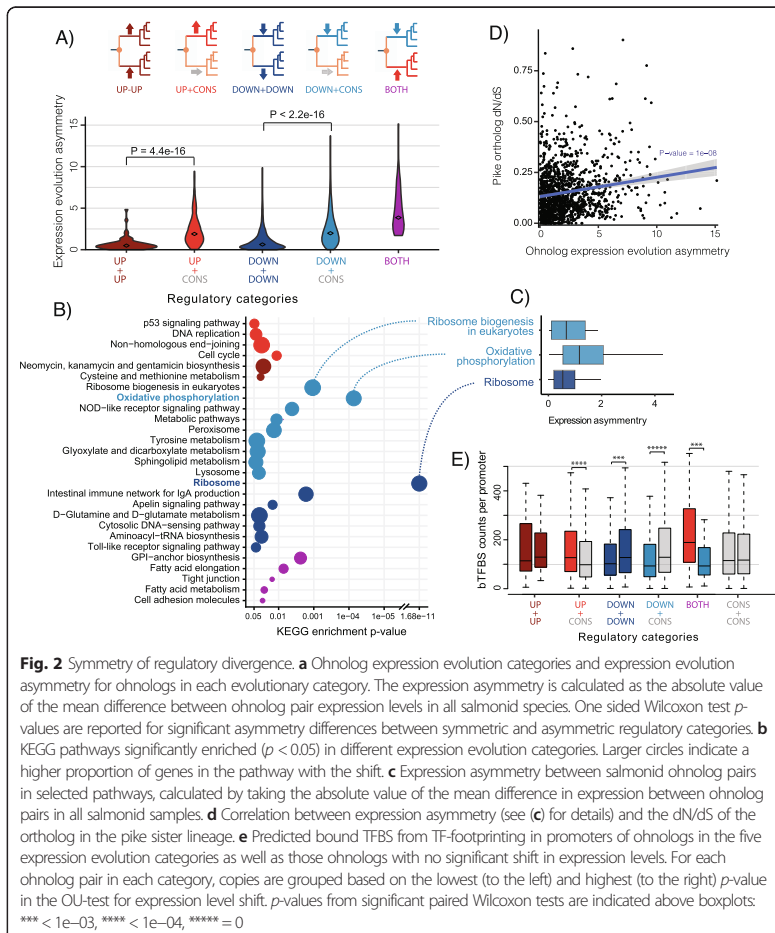
The gene balance hypothesis predicts that selection operates to maintain stoichiometry of interacting gene products [30], and this is believed to result in long-term retention of ohnologs. Using the human orthologs of salmonid genes, we queried the CORUM database of protein complexes and found that the proportion of ohnologs in protein complexes was slightly higher (22%) than the proportion of singletons (18%) (Fisher's exact test, $p = 1.04e-5$, Additional file 1: Figure S14A). We also found that complexes tended to contain only singletons ($p = 0.03$) or only duplicates ($p = 1E-3$, Additional file 1: Figure S14B) more often than expected by chance. It is also plausible that stoichiometric imbalances could be rescued through evolution of novel gene dosage. Under this model, we predict that singletons in protein complexes that contain ohnologs should be enriched for shifts up in expression, while shifts down are predicted for ohnologs in complexes with singletons. These predictions are not well supported for singletons (Fisher's exact test, $p = 0.07$) nor ohnologs (Fisher's exact test, $p > 0.48$) (Additional file 1: Table S3).

Taken together, we find strong evidence for dosage selection following WGD on genes involved in basic cellular maintenance and cell division. In addition, we find evidence for selection to retain stoichiometric balance both at the sequence and expression level.

Mechanism driving ohnolog regulatory divergence is associated with functional constraints

Our analysis allows us to assign ohnolog pairs to different regulatory categories (Fig. 2a) that potentially represent distinct evolutionary routes to new gene dosage optimums after WGD. Indeed our results show that ohnolog pairs with expression evolution shifts in the same direction evolve more symmetrically (down+down and up+up) while

ohnologs where expression shifts occur in only one copy or in opposite directions display stronger asymmetric divergence (e.g., up/down+conserved) (Fig. 2a). To explore the links between these modes of regulatory divergence and gene function, we performed KEGG enrichment on each expression evolution category. Twenty-seven pathways were found enriched across these categories (Fig. 2b, Additional file 1: Table S4), which is more than twice as many as when grouping ohnologs into up- or down-shifted genes (Additional file 1: Table S1). This supports that different pathways are biased towards either symmetric or asymmetric regulatory evolution. The three most enriched pathways were the same as when testing up- and down-shifted genes only, but our stratification on regulatory categories of ohnologs reveals that ribosomal subunit ohnologs (“Ribosome”) evolved lower gene dosage through highly symmetrical down-shifts, while “oxidative phosphorylation” and “ribosome biogenesis in eukaryotes” are biased towards asymmetric divergence (Fig. 2c).



As ribosome subunit genes are known to be extremely slowly evolving genes (i.e., high sequence evolution constraints), we tested whether there is a broader correlation between sequence constraints and regulatory symmetry. Indeed, we find that ohnolog expression level symmetry is significantly correlated with the level of purifying selection on coding sequences (Spearman correlation, $p = 1e-8$, Fig. 2d).

To further dissect regulatory mechanisms driving ohnolog expression level evolution, we generated high coverage ATAC-seq data from the liver of Atlantic salmon and identified bound transcription factor binding sites (bTFBSs) using a footprinting approach (Additional file 1: Figure S15). We hypothesized that ohnolog regulatory evolution symmetry is shaped by the relative importance of selection on cis- versus trans-mutations. One simple prediction from this is that ohnolog pairs where one copy has evolved novel expression would have higher promoter divergence than ohnolog pairs with symmetric evolution. The divergence of bTFBSs in promoters ($-3000/+200$ bps from transcription start site) largely matched this prediction (Fig. 2e) with ohnologs having more asymmetric expression shifts (up+cons and down+cons) differing more with respect to the number of bTFBSs in their promoters compared to symmetrically evolving ohnologs (up+up, down+down, and cons+cons) (Fig. 2e). This offers a simple explanation of expression divergence after WGD, where genes with decreased expression level have lost TFBSs, and genes with increased expression have gained TFBSs, compared to the ancestral promoter structure. Comparing the overall similarity of promoters, computed as the correlation of bTFBS between symmetrically evolving (down+down) and asymmetrically evolving (down+cons) ohnolog pairs, did not reveal a similar trend (Wilcoxon test, $p = 0.234$, Additional file 1: Figure S16), which is consistent with high turnover of bTFBS even for highly conserved genes [31].

Together these results support that evolutionary constraints at the coding sequence divert ohnologs down different evolutionary routes towards novel gene dosage—either in an asymmetric or symmetric fashion.

Adaptive gain in liver expression through acquisition of tissue-specific cis-regulatory elements

Although the vast majority of adaptive expression evolution was associated with selection on lower gene dosage, our OU-analyses did reveal 30 ohnolog pairs where one copy had evolved liver-specific adaptive gains in expression following WGD. These genes are predicted to be involved in a variety of functions such as developmental processes, cell fate specificity, and more liver-centric functions such as endocrine signaling and lipid- and fatty-acid metabolism (Additional file 1: Table S5). To better understand the regulatory mechanisms involved in the evolution of these potential novel liver functions, we used our TF-footprinting data to test the hypothesis that adaptive gains in liver expression are linked to the acquisition of binding sites for TFs controlling liver-specific regulatory networks. Indeed, we found that promoters of up-shifted copies were occupied by many more liver-specific TFs than their non-shifted partners (Fig. 3a, Wilcoxon paired test, $p = 7.7e-05$). These liver-specific TFs are thus candidates for being involved in regulatory rewiring of up-shifted ohnologs (Fig. 3b). Interestingly, many TFs with the strongest bias towards occupying the promoters of up-shifted ohnolog copies have known general liver functions (i.e., hepatocyte nuclear factors; FOX1A,

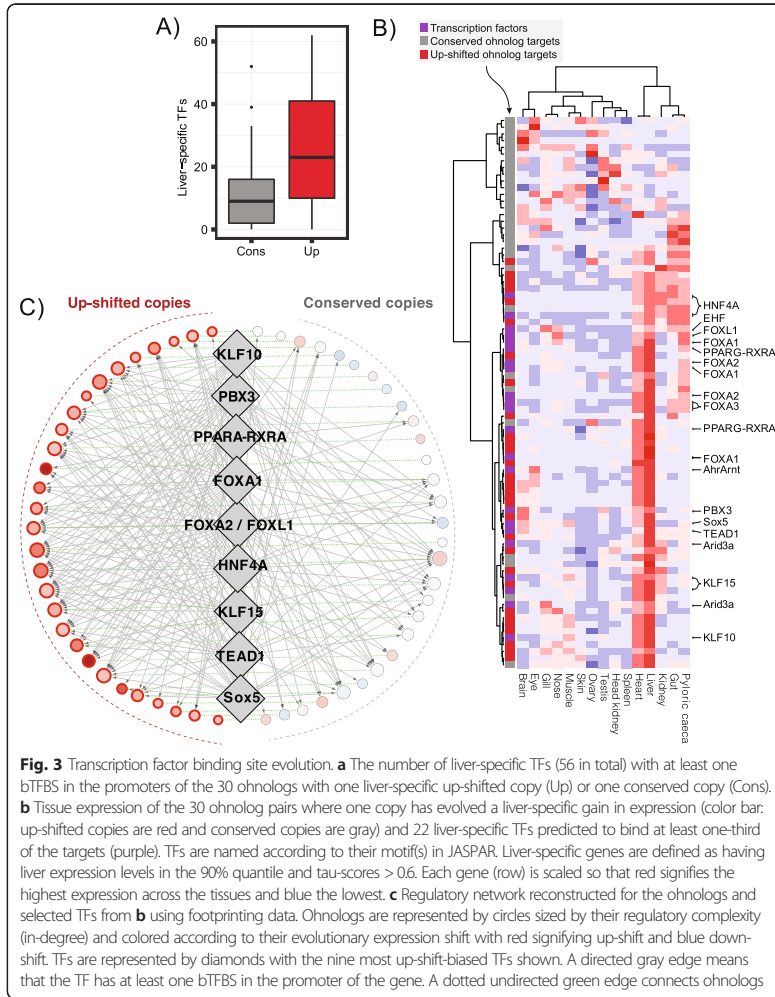


Fig. 3 Transcription factor binding site evolution. **a** The number of liver-specific TFs (56 in total) with at least one bTFBS in the promoters of the 30 ohnologs with one liver-specific up-shifted copy (Up) or one conserved copy (Cons). **b** Tissue expression of the 30 ohnolog pairs where one copy has evolved a liver-specific gain in expression (color bar: up-shifted copies are red and conserved copies are gray) and 22 liver-specific TFs predicted to bind at least one-third of the targets (purple). TFs are named according to their motif(s) in JASPAR. Liver-specific genes are defined as having liver expression levels in the 90% quantile and tau-scores > 0.6. Each gene (row) is scaled so that red signifies the highest expression across the tissues and blue the lowest. **c** Regulatory network reconstructed for the ohnologs and selected TFs from **b** using footprinting data. Ohnologs are represented by circles sized by their regulatory complexity (in-degree) and colored according to their evolutionary expression shift with red signifying up-shift and blue down-shift. TFs are represented by diamonds with the nine most up-shift-biased TFs shown. A directed gray edge means that the TF has at least one bTFBS in the promoter of the gene. A dotted undirected green edge connects ohnologs.

HNF4A) [32] and roles in lipid metabolism (RXR, PPARG, KLF15) [33, 34] (Fig. 3c, see the “Methods” section for details).

Next, we hypothesized that liver-specific increases in expression are driven by gains in new TFBSs. One way promoters can gain novel TFBSs is through insertions of TEs that either contain a functional TFBS or subsequently accumulate mutations that give rise to new TFBSs [35]. Indeed, we did find that TFBSs predicted to be bound by liver-specific TFs overlapped TEs more often in up-shifted copies than in conserved copies (Wilcoxon paired test, $p = 0.037$, Additional file 1: Figure S17A). Furthermore, at the level of TE superfamilies we found that the TIR TC1-Mariner TE superfamily were associated with gain in liver-specific bTFBS in up-shifted copies ($p = 0.018$, Additional file 1: Figure S17B), which included known liver and lipid metabolism transcription factors such as HNF4A, KLF15, and RXRA (Additional file 1: Table S6).

In conclusion, we find that adaptive gain in liver-specific expression is strongly associated with gain in liver-specific bound TFBSs, some of which have been facilitated by transposable element insertions.

Discussion

The consequence of WGDs for evolution of novel adaptations, including gene expression levels, has been an actively debated topic within evolutionary biology [1]. A key challenge has been to distinguish neutral from adaptive evolution in systems where experimental evolution is not possible [12]. Here, we generated a large comparative transcriptomics dataset and for the first time applied a formal phylogenetic model to infer selection on gene expression in the aftermath of a vertebrate WGD that occurred 80–100 million years ago.

Selection on gene dosage ameliorates immediate polyploid fitness costs

Newly formed polyploids often display augmented rates of abnormal mitosis, chromosome loss, and gross chromosomal rearrangements [36, 37]. Hence, a primary challenge for the evolutionary success of polyploids is to maintain genomic stability. In line with this, we find that adaptive evolution of gene expression was highly biased towards cellular functions not specific to the liver (Figs. 1e, f and 3b) and with a clear potential impact on genome stability. Firstly, we find genes directly involved in the cell cycle to be enriched for adaptive evolution (higher dosage). Related genes have experienced selective sweeps following WGD in plants [38, 39]. Furthermore, we find strong evidence for selection on genes involved in oxidative phosphorylation (lower dosage). Polyploidization in plants, fungi, and mammalian cells have been shown to increase levels of reactive oxygen species, which is causally linked to increased cellular stress, cell cycle failure, and increased genome instability [40–42]. Lastly, we find adaptive expression evolution (lower dosage) for genes involved in translation (ribosome subunits and ribosome assembly) after WGD. Regulation of translation also interacts with cell cycle regulation, with potential implications for genome stability [43]. However, selection for decreased expression of translation-related genes could also be linked to direct fitness costs of wasteful protein translation or harmful effects linked to the over-production of particular proteins. Overall, our study provides evidence for a scenario where a critical first step in becoming a successful polyploid lineage is pervasive adaptive evolution on gene dosage to ameliorate fitness costs linked to genome stability.

Long-term ohnolog retention and selection on gene dosage

Following an early phase of selection on gene dosage, the long-term fates of ohnologs can be shaped by various adaptive processes [21, 44], including adaptive regulatory evolution. One potential outcome is adaptive divergence between ohnologs, resulting in two functionally non-redundant ohnologs under purifying selection. Our results demonstrate that tissue-specific shifts (up+cons) in expression are rare (Fig. 1c), and interpret this to mean that adaptive evolution of novel tissue-specific regulation likely has very little impact on genome wide ohnolog retention.

Selection on molecular stoichiometry has been proposed to play a major role in genome evolution after WGDs [30]. This narrative is supported by our finding that

molecular complexes are both enriched for retained ohnologs and biased to include only singletons or only ohnologs (Additional file 1: Figure S14). However, selection on molecular stoichiometry could also drive evolution of gene expression. Indeed, we do find some (but weak) support for selection on stoichiometric balance also operating through selection for higher expression levels of singleton genes that are in complexes with ohnologs (Additional file 1: Table S3, $p = 0.07$). Moreover, it is plausible that the strong bias of “oxidative phosphorylation”-ohnologs towards highly asymmetric expression regulation also is linked to selection on stoichiometry (Fig. 2b, c). These genes are nuclear encoded genes involved in energy-related functions in mitochondria. As WGDs do not double the plastid numbers it has been proposed that in plants stoichiometric imbalances between nuclear and plastid genomes act as selection pressure to reduce the ratio of nuclear to plastid gene dosage following WGD [45]. In line with this reasoning, the driver behind the strong asymmetric down shift of “oxidative phosphorylation”-ohnologs could be the reinstatement of stoichiometric balance between the nuclear and mitochondrial genes.

At the other end of the ohnolog expression evolution symmetry spectrum, we find ohnologs belonging to the “ribosomal protein” pathway evolving lower expression in a highly symmetric fashion (Fig. 2b, c). We also demonstrate a significant correlation between constraints at the coding sequence level and symmetry of ohnolog regulatory evolution (Fig. 2d). This is in line with findings from plants that ribosomal proteins are retained over long evolutionary times and evolve slowly at both sequence and expression levels following WGD [46]. One potential explanation for this pattern could be the “toxic effects model” where long-term conservation of ohnologs is intrinsically linked to the “danger” of accumulating highly toxic coding sequence mutations [47, 48]. We therefore hypothesize that in situations where lowering the total gene dosage increases fitness, and the tolerance for accumulation of deleterious mutations is low (i.e., the toxic effect), symmetric ohnolog evolution towards lower gene dosage could be favored over pseudogenization of one copy. Eventually, mutations can arise that create completely non-functional pseudogenes without toxic-effects, and these can then be fixed in the population. This would result in an enrichment of singletons among genes that are likely to produce toxic effects, as observed in plants [45].

Divergence of chromatin landscapes and ohnolog expression

Regulatory divergence after gene duplication is hypothesized to be linked to evolution of local chromatin landscapes [18, 49]. Using ATAC-seq data we show that signals of adaptive expression level shifts are associated with the numbers of bound TFBSs (Fig. 2e), consistent with a billboard-like model of gene regulation [50]. Furthermore, we find that both loss of expression (Fig. 1h) and tissue-specific gains in expression level (Additional file 1: Figure S17) is linked to TE activity, highlighting the dual role of TEs in regulatory evolution following WGD.

Conclusion

Our study supports pervasive selection on gene dosage across millions of years following WGD, in particular for genes involved in basic cellular maintenance and genome stability. Interestingly, many of the homologous genes and pathways also show similar

responses in gene dosage adjustments immediately after polyploidization in plants [16]. Reconciling these immediate effects of polyploidization with our findings strongly supports the following model: Plastic genome regulatory response to polyploidization alleviate immediate fitness costs following genome doubling. Since gene loss is absent in early generations polyploids, and all genes are duplicated and in stoichiometric balance, early plastic changes in gene regulatory phenotypes is likely a result of deleterious fitness effects due to suboptimal absolute gene dosages. Over evolutionary time-scales however, selection will favor and fix regulatory mutations that can “hard code” novel transcriptional phenotypes to optimize gene dosages (as seen following the salmonid WGD). Together, this paper points to critical genome regulatory adjustments for becoming a successful polyploid lineage.

Methods

Ortholog inference

For ortholog inference, we used thirteen species including six salmonids (*Thymallus thymallus*, *Hucho hucho*, *Salmo salar*, *Salvelinus* sp., *Oncorhynchus mykiss*, and *Oncorhynchus kisutch*), four teleosts as outgroups to the salmonids (*Danio rerio*, *Oryzias latipes*, *Gasterosteus aculeatus*, and *Esox lucius*), one non-teleost fish (*Lepisosteus oculatus*) and two mammals as outgroups to the teleosts (*Homo sapiens* and *Mus musculus*). We only report the genus name for the char (*Salvelinus* sp.) because it was recently discovered that the material used for sequencing *Salvelinus alpinus* could have been a very closely related sister species (*Salvelinus malma*) or a hybrid between the two [51]. Protein sequences were obtained from ENSEMBL (release 92) for *H. sapiens*, *M. musculus*, *L. oculatus*, *D. rerio*, *O. latipes*, and *G. aculeatus*, from NCBI RefSeq assemblies for *S. salar* (GCF_000233375.1), *Salvelinus* sp. (GCF_002910315.2), *O. mykiss* (GCF_002163495.1), *O. kisutch* (GCF_002021735.1), and *E. lucius* (GCF_000721915.3), from the genome paper for *T. thymallus* [52] and from an in-house annotation using Transdecoder (<https://github.com/TransDecoder/TransDecoder/wiki>) for *H. hucho* (GCA_003317085). The single longest protein per gene was assigned to gene ortholog groups (orthogroups) using OrthoFinder (v2.3.1) [53]. For each orthogroup, the corresponding CDS sequences were aligned using MACSE (v2.03) before gene trees were generated and reconciled against the species tree using TreeBest (v1.9.2). The gene trees were then split at the level of monophyletic teleost clades, defining what we refer to as trees in this article, and again at the level of the salmonid clade (excluding *T. thymallus* and *H. hucho*), defining the Ss4R duplicate clades. Trees were then selected based on their topology (Additional file 1: Figure S1). Specifically, this filtered any trees that showed more than two salmonid clades or that contained additional paralogs inside the salmonid clades or in the outgroup species. Trees with all orthologs retained in the salmonid clade(s) were designated as complete, and otherwise as partial. In addition, trees were excluded from further analysis if (1) one or both salmonid clades had no expressed genes (zero mapped reads, RNA-seq data described below), (2) the *E. lucius* ortholog was missing or not expressed, and (3) both the *D. rerio* and *O. latipes* orthologs were missing or not expressed.

RNA-sequencing data

Liver tissue samples were collected from adult individuals of *D. rerio* (zebrafish), *O. latipes* (medaka), *E. lucius* (pike), *O. mykiss* (rainbow trout), *S. alpinus* (Arctic char), and *O. kisutch* (coho salmon) (Fig. 1a). Samples were taken in replicates of four, or three in the case of rainbow trout. All fish were raised in fresh water under standard rearing conditions in aquaculture facilities (salmonids), animal laboratory facilities (zebrafish and medaka), or restocking hatcheries (pike). Total RNA was extracted from the liver samples using the RNeasy Plus Universal Kit (QIAGEN). Quality was determined on a 2100 Bioanalyzer using the RNA 6000 Nano Kit (Agilent). Concentration was determined using a Nanodrop 8000 spectrophotometer (Thermo Scientific). cDNA libraries were prepared using the TruSeq Stranded mRNA HT Sample Prep Kit (Illumina). Library mean length was determined by running on a 2100 Bioanalyzer using the DNA 1000 Kit (Agilent) and library concentration was determined with the Qbit BR Kit (Thermo Scientific). Paired-end sequencing of sample libraries was completed on an Illumina HiSeq 2500 with 125-bp reads. Raw RNA-seq and processed count data have been deposited into ArrayExpress under the projects E-MTAB-8959 and E-MTAB-8962. For *S. salar* (Atlantic salmon), RNA-seq data was obtained from a feeding trial using four samples from individuals in freshwater fed a marine based diet [54], available in the European Nucleotide Archive (ENA) under project PRJEB24480 (samples: ERS2101563, ERS2101567, ERS2101568, ERS2101569).

To generate gene expression data, RNA-seq reads were mapped to the annotated reference genomes using the STAR aligner with default settings [55]. RSEM [56] was used to estimate read counts and Transcripts Per Million reads (TPM)-expression values that are normalized for average transcript lengths and the total number of reads from each sample.

The trimmed mean of M values (TMM), from the R package edgeR [57], was used to compute normalization factors for the gene expression data. The replicates were first normalized within each species and then between species (Additional file 1: Figure S2). Between-species normalization was accomplished by first computing species-specific normalization factors using genes from singleton orthogroups (i.e., groups containing only one gene from each species) and their mean expression values (i.e., mean of the replicates within each species), and then by normalizing the individual replicates from each species using these normalization factors. All expression values were log transformed ($\log_2(\text{TPM}+0.01)$) prior to testing for expression shifts.

Evolutionary shifts in gene expression

The EVE model [22] was used to test for shifts in gene expression levels in the salmonid clade(s) within each gene tree. For this paper, we developed and implemented a user friendly version of the EVE algorithm in R (<https://gitlab.com/sandve-lab/evemodel>). This method models an OU process, i.e., random drift in expression level that is constrained around an optimal level. The test compares a model with two optimal expression levels, one for the salmonid branch and another for the outgroup species, against the null-model which has the same optimal expression level across the entire tree (Additional file 1: Figure S3C). For ohnolog gene trees which contain two

duplicate salmonid clades, each clade was tested separately by removing the other salmonid clade.

EVE was given the expression data for each species (four samples/replicates per species) and the species tree produced by OrthoFinder. For every ortholog, a likelihood ratio test (LRT) score is calculated, representing the likelihood of the alternative hypothesis over the null hypothesis. LRT scores were compared to a chi-squared distribution with one degree of freedom and scores above the 95% quantile were considered to be significant. EVE reports estimates of the expression optimum for the salmonid branch and the rest of the tree (i.e., outgroup species), and the difference between salmonid estimates and outgroup estimates provided the direction of the expression shift.

Tissue atlas

Gene expression data from an Atlantic salmon tissue atlas [17] was clustered using Pearson correlation and the R function `hclust` with `method = "ward.D"`. Heatmaps were drawn using the R function `pheatmap` with `scale = "row"`.

Coding sequence selection pressure

We estimated branch-specific selection pressure on coding sequences in ohnolog gene trees by calculating dN/dS measured at the branch from the WGD node to the root of each duplicate clade using the aBSREL (adaptive Branch-Site Random Effects Likelihood) method [58] in HyPhy (Hypothesis Testing using Phylogenies) [59]. A one-sided paired Wilcoxon test was then performed to test if there is a difference in selection pressure between ohnolog pairs classified as asymmetrically shifted at the expression level.

Transposable elements

Transposable element (TE) annotations were taken from [17]. For Atlantic salmon genes, we calculated the proportion of gene promoter sequence (+2 kb/-200b from TSS) that was overlapped with TEs using `bedtools intersect` of promoter and TE annotations. We used a one-sided paired Wilcoxon test to test the hypothesis that, for ohnologs with an asymmetric shift down in expression, the shifted copy had a higher proportion of TE overlap than the conserved copy.

Gene function enrichment

We assigned KEGG pathway annotations to the orthogroups based on the Northern pike ortholog and its KEGG annotations. We then tested each set of ohnologs within an expression shift category for the enrichment of KEGG pathways using the `kegga` function from the R package `limma`, with all tested ohnologs as the background.

Protein complexes

We assigned orthogroups as being in a protein complex or not based on the human ortholog and its protein complex annotations from the CORUM database [60]. We used the Fisher's exact test, for singleton and ohnolog genes, to test whether more genes within an expression shift category were in a protein complex than expected by chance. To test if complexes were biased towards only containing singletons or only

ohnologs, we randomized the singleton/ohnolog label 10,000 times and reported empirical *p*-values.

ATAC-seq generation and TF footprinting

Four Atlantic salmon (freshwater stage, 26–28 g) were euthanized using a Schedule 1 method following the Animals (Scientific Procedures) Act 1986. Around 50-mg homogenized brain and liver tissue was processed to extract nuclei using the Omni-ATAC protocol for frozen tissues [61]. Nuclei were counted on an automated cell counter (TC20 BioRad, range 4–6 μ m) and further confirmed intact under microscope. A total of 50,000 nuclei were used in the transposition reaction including 2.5 μ L Tn5 enzyme (Illumina Nextera DNA Flex Library Prep Kit), incubated for 30 min at 37 °C in a shaker at 200 rpm. The samples were purified with the MinElute PCR purification kit (Qiagen) and eluted in 12 μ L elution buffer. qPCR was used to determine the optimal number of PCR cycles for library preparation [62] (8–10 cycles used). Sequencing libraries were prepared with short fragments and fragments > 1000 bp removed using AMPure XP beads (Beckman Coulter, Inc.). Fragment length distributions and confirmation of nucleosome banding patterns were determined on a 2100 Bioanalyzer (Agilent) and the library concentration estimated using a Qubit system (Thermo Scientific). Libraries were sent to the Norwegian Sequencing Centre, where paired-end 2 \times 75 bp sequencing was done on an Illumina HiSeq 4000. The raw sequencing data for brain and liver is available through ArrayExpress (Accession: E-MTAB-9001).

Reads were mapped using BWA-MEM [63]. Duplicate reads and reads mapping to mitochondrial or unplaced scaffolds were removed. Peaks were called using MACS2 [64]. TF footprinting was performed with TOBIAS [65] based on the aligned reads, peaks, and TF motifs from JASPAR (JASPAR 2020 non-redundant vertebrate CORE PFMs) [66]. TOBIAS performs Tn5 bias correction, generates footprint scores for each base within the peaks, scans for TFBSs using the given TF motifs, and finally classifies each TFBS as bound or unbound based on the footprint scores.

For the analysis of ohnolog pairs with evolved liver-specific expression increases in one copy, we identified 30 up+cons pairs (60 target genes) where the liver expression of the up-copy was at least 90% of the maximum expression in the tissue atlas and the up-copy had a tissue specificity score (τ) > 0.6 [17]. To identify regulators of these genes, we BLASTed UniProt TF sequences with a motif in JASPAR to the Atlantic salmon proteome, and retained the top four hits with E-value < 1E-10 and alignment length > 100. We then filtered these TFs for having bTFBS in the promoter of at least 20 of the target genes and for having liver-specific expression (same criteria as for up-targets). This resulted in 22 liver-specific TFs predicted to bind 17 different JASPAR motifs in 52 target promoters (Fig. 3b, c). Finally, to draw the network in Fig. 3c, we (1) selected, for each JASPAR motif, the single TF with the strongest evolutionary shift in expression; (2) removed JASPAR motifs with highly similar binding profiles (> 80% overlap in target genes, retaining the TF with the strongest evolutionary shift); and (3) merged TFs associated with more than one JASPAR motif into one node and selected the nine TFs with the strongest bias towards up-shifted targets.

Reproducibility

The scripts developed to implement analyses described in this study are available here: <https://gitlab.com/sandve-lab/gillard-groenvold> [67] and <https://doi.org/10.5281/zenodo.4478402> [68].

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-021-02323-0>.

Additional file 1: Supplementary figures S1-S17 and Supplementary Tables S1-S6.

Additional file 2. Review history.

Acknowledgements

We thank Zuzana Storchová, Galal Metwalli, Marc Robinson-Rechavi, Gavin Conant, Camille Berthelot, and Jeremy Coate for valuable discussions.

Review history

The review history is available as Additional file 2.

Peer review information

Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

SRS, TRH, and RR conceived the study. GG, LG, LR, MMH, ØM, MKG, DJM, JM, BK, and EBR were involved in formal generation, analysis, and/or curation of data. LG, GG, SRS, and TRH wrote the paper. All authors contributed intellectually to data analyses and interpretation. The authors read and approved the final manuscript.

Funding

The research was conducted as part of the NRC funded project Rewired (NRC project number 274669) and the NRC and NMBU funded project Transpose (NRC project number 275310). GG was funded by NMBU.

Availability of data and materials

Raw RNA-seq and processed count data have been deposited into ArrayExpress under the projects E-MTAB-8959 [69] and E-MTAB-8962 [70]. Raw ATAC-seq data is also available through ArrayExpress under project E-MTAB-9001 [71]. The scripts developed to implement analyses described in this study are available on GitLab [67] and Zenodo [68].

Declarations

Ethics approval and consent to participate

The fish used in this study were treated according to the Norwegian Animal Research Authority (NARA) in accordance with the Norwegian Animal Welfare Act of 19th of June 2009.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås, Norway. ²Center for Integrative Genetics, Department of Animal and Aquacultural Sciences, Faculty of Biosciences, Norwegian University of Life Sciences, Ås, Norway. ³Department of Biology, University of Victoria, Victoria, Canada. ⁴The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Midlothian, UK. ⁵Department of Computer Science, San Francisco State University, San Francisco, USA. ⁶Department of Biology, San Francisco State University, San Francisco, USA.

Received: 27 August 2020 Accepted: 23 March 2021

Published online: 13 April 2021

References

1. Van de Peer Y, Mizrahi E, Marchal K. The evolutionary significance of polyploidy. *Nat Rev Genet.* 2017;18(7):411–24. <https://doi.org/10.1038/nrg.2017.26>.
2. Ohno S. Evolution by gene duplication. Berlin, Heidelberg: Springer Berlin Heidelberg; 1970. <https://doi.org/10.1007/978-3-642-86659-3>.
3. Merico A, Sulo P, Piskur J, Compagno C. Fermentative lifestyle in yeasts belonging to the *Saccharomyces* complex. *FEBS J.* 2007;274(4):976–89. <https://doi.org/10.1111/j.1742-4658.2007.05645.x>.

4. Soltis PS, Soltis DE. Ancient WGD events as drivers of key innovations in angiosperms. *Curr Opin Plant Biol.* 2016;30:159–65. <https://doi.org/10.1016/j.pbi.2016.03.015>.
5. Lohaus R, Van de Peer Y. Of dups and dinos: evolution at the K/Pg boundary. *Curr Opin Plant Biol.* 2016;30:62–9. <https://doi.org/10.1016/j.pbi.2016.01.006>.
6. Holland PW, Garcia-Fernández J, Williams NA, Sidow A. Gene duplications and the origins of vertebrate development. *Development.* 1994;Supplement:125–33.
7. Meyer A, Van De Peer Y. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays.* 2005;27(9):937–45. <https://doi.org/10.1002/bies.20293>.
8. Volff JN. Genome evolution and biodiversity in teleost fish. *Heredity.* 2005;94(3):280–94. <https://doi.org/10.1038/sj.hdy.6800635>.
9. Soltis DE, Segovia-Salcedo MC, Jordon-Thaden I, Majure L, Miles NM, Mavrodiev EV, Mei W, Cortez MB, Soltis PS, Gitzendanner MA. Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose et al. (2011). *New Phytol.* 2014;202(4):1105–17. <https://doi.org/10.1111/nph.12756>.
10. Andalis AA, Storchová Z, Styles C, Galitski T, Pellman D, Fink GR. Defects arising from whole-genome duplications in *Saccharomyces cerevisiae*. *Genetics.* 2004;167(3):1109–21. <https://doi.org/10.1534/genetics.104.029256>.
11. Kuznetsova AY, Seget K, Moeller GK, de Pagter MS, de Roos JADM, Dürbaum M, Kuffer C, Müller S, Zaman GJR, Kloosterman WP, Storchová Z. Chromosomal instability, tolerance of mitotic errors and multidrug resistance are promoted by tetraploidization in human cells. *Cell Cycle.* 2015;14(17):2810–20. <https://doi.org/10.1080/15384101.2015.1068482>.
12. Sandve SR, Rohlfs RV, Hvidsten TR. Subfunctionalization versus neofunctionalization after whole-genome duplication. *Nat Genet.* 2018;50(7):908–9. <https://doi.org/10.1038/s41588-018-0162-4>.
13. Boyle EA, Li YL, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. *Cell.* 2017;169(7):1177–86. <https://doi.org/10.1016/j.cell.2017.05.038>.
14. Verta J-P, Jones FC. Predominance of cis-regulatory changes in parallel expression divergence of sticklebacks. *elife.* 2019; 8. <https://doi.org/10.7554/eLife.43785>.
15. Ishikawa A, Kabeya N, Ikeya K, Kakioka R, Cech JN, Osada N, Leal MC, Inoue J, Kume M, Toyoda A, Tezuka A, Nagano AJ, Yamasaki YY, Suzuki Y, Kokita T, Takahashi H, Lucek K, Marques D, Takehana Y, Naruse K, Mori S, Monroig O, Ladd N, Schubert CJ, Matthews B, Peichel CL, Seehausen O, Yoshizaki G, Kitano J. A key metabolic gene for recurrent freshwater colonization and radiation in fishes. *Science.* 2019;364(6443):886–9. <https://doi.org/10.1126/science.aau5656>.
16. Song MJ, Potter B, Doyle JJ, Coate JE. Gene balance predicts transcriptional responses immediately following ploidy change in *Arabidopsis thaliana*. *Plant Cell.* 2020;32(5):1434–48. <https://doi.org/10.1105/tpc.19.00832>.
17. Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, Hvidsten TR, Leong JS, Minkley DR, Zimin A, Grammes F, Grove H, Gjuvsland A, Walenz B, Hermansen RA, von Schalburg K, Rondeau EB, di Genova A, Samy JKA, Olav Vik J, Vigeland MD, Caler L, Grimholt U, Jentoft S, Inge Våge D, de Jong P, Moen T, Baranski M, Palti Y, Smith DR, Yorke JA, Nederbragt AJ, Tooming-Klunderud A, Jakobsen KS, Jiang X, Fan D, Hu Y, Liberles DA, Vidal R, Iturra P, Jones SJM, Jonassen I, Maass A, Omholt SW, Davidson WS. The Atlantic salmon genome provides insights into rediploidization. *Nature.* 2016; 533(7602):200–5. <https://doi.org/10.1038/nature17164>.
18. Marlétaz F, Firas PN, Maeso I, Tena JJ, Bogdanovic O, Perry M, Wyatt CDR, de la Calle-Mustienes E, Bertrand S, Burguera D, Acemel RD, van Heeringen SJ, Naranjo S, Herrera-Ubeda C, Skvortsova K, Jimenez-Gancedo S, Aldea D, Marquez Y, Buono L, Kozmikova I, Permyner J, Louis A, Albuixech-Crespo B, le Petillon Y, Leon A, Subirana L, Balwierz PJ, Duckett PE, Farahani E, Aury JM, Mangenot S, Wincker P, Albalat R, Benito-Gutiérrez E, Cañestro C, Castro F, D'Aniello S, Ferrer DEK, Huang S, Laudet V, Marais GAB, Pontarotti P, Schubert M, Seitz H, Somorjai I, Takahashi T, Mirabeau O, Xu A, Yu JK, Carninci P, Martínez-Morales JR, Crollius HR, Kozmik Z, Weirauch MT, Garcia-Fernández J, Lister R, Lenhard R, Holland PWH, Escriva H, Gómez-Skarmeta JL, Irimia M. Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature.* 2018;564(7734):64–70. <https://doi.org/10.1038/s41586-018-0734-6>.
19. De Smet R, Sabaghian E, Li Z, Saey Y, Van de Peer Y. Coordinated functional divergence of genes after genome duplication in *Arabidopsis thaliana*. *Plant Cell.* 2017;29(11):2786–800. <https://doi.org/10.1105/tpc.17.00531>.
20. Rodgers-Melnick E, Mane SP, Dharmawardhana P, Slavov GT, Crasta OR, Strauss SH, Brunner AM, DiFazio SP. Contrasting patterns of evolution following whole genome versus tandem duplication events in *Populus*. *Genome Res.* 2012;22(1):95–105. <https://doi.org/10.1101/gr.125146.111>.
21. Hallin J, Landry CR. Regulation plays a multifaceted role in the retention of gene duplicates. *PLoS Biol.* 2019;17(11): e3000519. <https://doi.org/10.1371/journal.pbio.3000519>.
22. Rohlfs RV, Nielsen R. Phylogenetic ANOVA: the expression variance and evolution model for quantitative trait evolution. *Syst Biol.* 2015;64(5):695–708. <https://doi.org/10.1093/sysbio/syv042>.
23. Rohlfs RV, Harrigan P, Nielsen R. Modeling gene expression evolution with an extended Ornstein-Uhlenbeck process accounting for within-species variation. *Mol Biol Evol.* 2014;31(1):201–11. <https://doi.org/10.1093/molbev/mst190>.
24. Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, Bento P, da Silva C, Labadie K, Alberti A, Aury JM, Louis A, Dehais P, Bardou P, Montfort J, Klopp C, Cabau C, Gaspin C, Thorgaard GH, Bousaha M, Quillet E, Guyomard R, Galiana D, Bobe J, Volff JN, Genêt C, Wincker P, Jaillon O, Crollius HR, Guiguen Y. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun.* 2014;5(1):3657. <https://doi.org/10.1038/ncomms4657>.
25. Macqueen DJ, Johnston IA. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proc Biol Sci.* 2014;281(1778):20132881. <https://doi.org/10.1098/rspb.2013.2881>.
26. Gout J-F, Lynch M. Maintenance and loss of duplicated genes by dosage subfunctionalization. *Mol Biol Evol.* 2015;32(8):2141–8. <https://doi.org/10.1093/molbev/msv095>.
27. Langham RJ, Walsh J, Dunn M, Ko C, Goff SA, Freeling M. Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics.* 2004;166(2):935–45. <https://doi.org/10.1534/genetics.166.2.935>.
28. Conant GC. The continuing impact of an ancient polyploidy on the genomes of teleosts. *BioRxiv.* 2019. <https://doi.org/10.1101/619205>.

29. Xu P, Xu J, Liu G, Chen L, Zhou Z, Peng W, Jiang Y, Zhao Z, Jia Z, Sun Y, Wu Y, Chen B, Pu F, Feng J, Luo J, Chai J, Zhang H, Wang H, Dong C, Jiang W, Sun X. The allotetraploid origin and asymmetrical genome evolution of the common carp *Cyprinus carpio*. *Nat Commun.* 2019;10(1):4625. <https://doi.org/10.1038/s41467-019-12644-1>.
30. Birchler JA, Veitia RA. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc Natl Acad Sci U S A.* 2012;109(37):14746–53. <https://doi.org/10.1073/pnas.1207726109>.
31. Otto W, Stadler PF, López-Giraldez F, Townsend JP, Lynch VJ, Wagner GP. Measuring transcription factor-binding site turnover: a maximum likelihood approach using phylogenies. *Genome Biol Evol.* 2009;1:85–98. <https://doi.org/10.1093/gbe/evp010>.
32. Lau HH, Ng NHJ, Loo LSW, Jasmen JB, Teo AKK. The molecular functions of hepatocyte nuclear factors - in and beyond the liver. *J Hepatol.* 2018;68(5):1033–48. <https://doi.org/10.1016/j.jhep.2017.11.026>.
33. Prosdocimo DA, Anand P, Liao X, Zhu H, Shelkay S, Artero-Calderon P, Zhang L, Kirsh J, Moore DV, Rosca MG, Vazquez E, Kerner J, Akat KM, Williams Z, Zhao J, Fujioka H, Tuschl T, Bai X, Schulze PC, Hoppel CL, Jain MK, Haldar SM. Kruppel-like factor 15 is a critical regulator of cardiac lipid metabolism. *J Biol Chem.* 2014;289(9):5914–24. <https://doi.org/10.1074/jbc.M113.531384>.
34. Carmona-Antoñanzas G, Tocher DR, Martínez-Rubio L, Leaver MJ. Conservation of lipid metabolic gene transcriptional regulatory networks in fish and mammals. *Gene.* 2014;534(1):1–9. <https://doi.org/10.1016/j.gene.2013.10.040>.
35. Feschotte C. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet.* 2008;9(5):397–405. <https://doi.org/10.1038/nrg2337>.
36. Storchova Z, Pellman D. From polyploidy to aneuploidy, genome instability and cancer. *Nat Rev Mol Cell Biol.* 2004;5:45–54. <https://doi.org/10.1038/nrm1276>.
37. Storchová Z, Breneman A, Cande J, Dunn J, Burbank K, O'Toole E, Pellman D. Genome-wide genetic analysis of polyploidy in yeast. *Nature.* 2006;443:541–7. <https://doi.org/10.1038/nature05178>.
38. Marburger S, Monahan P, Seear PJ, Martin SH, Koch J, Paaanen P, Bohutínská M, Higgins JD, Schmickl R, Yant L. Interspecific introgression mediates adaptation to whole genome duplication. *Nat Commun.* 2019;10(1):5218. <https://doi.org/10.1038/s41467-019-13159-5>.
39. Hollister JD, Arnold BJ, Svedin E, Xue KS, Dilkes BP, Bombles K. Genetic adaptation associated with genome-doubling in autotetraploid *Arabidopsis arenosa*. *PLoS Genet.* 2012;8(12):e1003093. <https://doi.org/10.1371/journal.pgen.1003093>.
40. Roh M, van der Meer R, Abdulkadir SA. Tumorigenic polyploid cells contain elevated ROS and ARE selectively targeted by antioxidant treatment. *J Cell Physiol.* 2012;227(2):801–12. <https://doi.org/10.1002/jcp.22793>.
41. Thomson GJ, Hernon C, Austricco N, Shapiro RS, Belenky P, Bennett RJ. Metabolism-induced oxidative stress and DNA damage selectively trigger genome instability in polyploid fungal cells. *EMBO J.* 2019;38:e101597. <https://doi.org/10.15252/emboj.2019101597>.
42. del Pozo JC, Ramírez-Parra E. Deciphering the molecular bases for drought tolerance in *Arabidopsis* autotetraploids. *Plant Cell Environ.* 2014;37(12):2722–37. <https://doi.org/10.1111/pce.12344>.
43. Zhou X, Liao W-J, Liao J-M, Liao P, Lu H. Ribosomal proteins: functions beyond the ribosome. *J Mol Cell Biol.* 2015;7(2):92–104. <https://doi.org/10.1093/jmcb/mjv014>.
44. Conant GC, Wolfe KH. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet.* 2008;9(12):938–50. <https://doi.org/10.1038/nrg2482>.
45. De Smet R, Adams KL, Vandepoele K, Van Montagu MCE, Maere S, Van de Peer Y. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc Natl Acad Sci U S A.* 2013;110(8):2898–903. <https://doi.org/10.1073/pnas.1300127110>.
46. Blanc G, Wolfe KH. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell.* 2004;16(7):1679–91. <https://doi.org/10.1105/tpc.021410>.
47. Roux J, Liu J, Robinson-Rechavi M. Selective constraints on coding sequences of nervous system genes are a major determinant of duplicate gene retention in vertebrates. *Mol Biol Evol.* 2017;34(11):2773–91. <https://doi.org/10.1093/molbev/msx199>.
48. Singh PP, Affeldt S, Cascone I, Selimoglu R, Camonis J, Isambert H. On the expansion of “dangerous” gene repertoires by whole-genome duplications in early vertebrates. *Cell Rep.* 2012;2(5):1387–98. <https://doi.org/10.1016/j.celrep.2012.09.034>.
49. Lan X, Pritchard JK. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science.* 2016;352(6288):1009–13. <https://doi.org/10.1126/science.1248411>.
50. Kulkarni MM, Arnosti DN. Information display by transcriptional enhancers. *Development.* 2003;130(26):6569–75. <https://doi.org/10.1242/dev.00890>.
51. Christensen KA, Rondeau EB, Minkley DR, Leong JS, Nugent CM, Danzmann RG, Ferguson MM, Stadnik A, Devlin RH, Muzzerall R, Edwards M, Davidson WS, Koop BF. Retraction: the Arctic charr (*Salvelinus alpinus*) genome and transcriptome assembly. *PLoS One.* 2021;16(2):e0247083. <https://doi.org/10.1371/journal.pone.0247083>.
52. Varadarajan S, Sandve SR, Gillard GB, Tørresen OK, Mulugeta TD, Hvidsten TR, Lien S, Asbjørn Vøllestad L, Jentoft S, Nederbragt AJ, Jakobsen KS. The grayling genome reveals selection on gene expression regulation after whole-genome duplication. *Genome Biol Evol.* 2018;10(10):2785–800. <https://doi.org/10.1093/gbe/evy201>.
53. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 2019;20(1):238. <https://doi.org/10.1186/s13059-019-1832-y>.
54. Gillard G, Harvey TN, Gjuvsland A, Jin Y, Thomassen M, Lien S, Leaver M, Torgersen JS, Hvidsten TR, Vik JO, Sandve SR. Life-stage-associated remodelling of lipid metabolism regulation in Atlantic salmon. *Mol Ecol.* 2018;27(5):1200–13. <https://doi.org/10.1111/mec.14533>.
55. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29(1):15–21. <https://doi.org/10.1093/bioinformatics/bts635>.
56. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics.* 2011;12(1):323. <https://doi.org/10.1186/1471-2105-12-323>.
57. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2010;26(1):139–40. <https://doi.org/10.1093/bioinformatics/btp616>.

58. Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol Biol Evol.* 2015;32(5):1342–53. <https://doi.org/10.1093/molbev/msv022>.
59. Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, Shank SD, Magalis BR, Bouvier D, Nekrutenko A, Wisotsky S, Spielman SJ, Frost SDW, Muse SV. HyPhy 2.5-a customizable platform for evolutionary hypothesis testing using phylogenies. *Mol Biol Evol.* 2020;37(1):295–9. <https://doi.org/10.1093/molbev/msz197>.
60. Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Ruepp A. CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res.* 2019;47(D1):D559–63. <https://doi.org/10.1093/nar/gky973>.
61. Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, Satpathy AT, Rubin AJ, Montine KS, Wu B, Kathiria A, Cho SW, Mumbach MR, Carter AC, Kasowski M, Orloff LA, Risca VI, Kundaje A, Khavari PA, Montine TJ, Greenleaf WJ, Chang HY. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods.* 2017;14(10):959–62. <https://doi.org/10.1038/nmeth.4396>.
62. Buenostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol.* 2015;109(1):21.29.1–9. <https://doi.org/10.1002/0471142727.mb2129s109>.
63. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013;arXiv:1303.3997.
64. Gaspar JM. Improved peak-calling with MACS2. *BioRxiv.* 2018. <https://doi.org/10.1101/496521>.
65. Bentsen M, Goymann P, Schultheis H, Klee K, Petrova A, Wiegandt R, Fust A, Preussner J, Kuenne C, Braun T, Kim J, Looso M. ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat Commun.* 2020;11(1):4267. <https://doi.org/10.1038/s41467-020-18035-1>.
66. Fomes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, George M, Baranašić D, Santana-Garcia W, Tan G, Chèneby J, Ballester B, Parcy F, Sandelin A, Lenhard B, Wasserman WW, Mathelier A. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 2020;48:D87–92. <https://doi.org/10.1093/nar/gkz1001>.
67. Gillard GB, Grønvold L, Sandve SR, Hvidsten TR. Gillard Grønvold GitLab repository. *GitLab.* 2020; <https://gitlab.com/sandve-lab/gillard-grønvold>.
68. Gillard GB, Grønvold L, Hvidsten TR, Sandve SR. Gillard Grønvold source code. *Zenodo.* 2021. <https://doi.org/10.5281/zenodo.4478402>.
69. Gillard GB, Sandve SR. RNA-seq of tissue panel samples from zebrafish (*Danio rerio*), medaka (*Oryzias latipes*), and rainbow trout (*Oncorhynchus mykiss*). E-MTAB-8959. *ArrayExpress.* 2020; <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-8959/>.
70. Gillard GB, Sandve SR, Rondeau EB, Koop BF. RNA-Seq of liver tissue samples from northern pike (*Esox lucius*), coho salmon (*Oncorhynchus kisutch*) and Arctic charr (*Salvelinus alpinus*). E-MTAB-8962. *ArrayExpress.* 2020; <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-8962/>.
71. Røssæg LL, Holen MM, Sandve SR, Kent MP, Lien S. Fresh versus slow-frozen ATACseq samples for salmon tissues. E-MTAB-9001. *ArrayExpress.* 2020; <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-9001/>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



PAPER 2

The role of transposable elements in the evolution of cis-regulatory element landscapes after whole genome duplication

Øystein Monsen¹, Lars Grønvold¹, James Kijas², Alexander Suh^{3,4}, Torgeir R. Hvidsten⁵, Simen Rød Sandve¹

1: Department of Animal and Aquacultural Sciences, Faculty of Bioscience, Norwegian University of Life Sciences

2: Aquaculture Programme, Commonwealth Scientific and Industrial Research Organisation

3: School of Biological Sciences – Organisms and the Environment, University of East Anglia, Norwich Research Park, NR4 7TU, Norwich, UK

4: Department of Organismal Biology – Systematic Biology (EBC), Uppsala University, Norbyvägen 18D, SE-752 36 Uppsala, Sweden

5: Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås, Norway

Abstract

Background: Armed with sequence motifs that can function as, or evolve into, cis-regulatory elements (CRE), transposable elements (TEs) hold great potential for impacting the evolution of genome regulation. Salmonid genomes have a high TE content (~50%), and interestingly, a burst of salmonid-specific TE activity is linked to an ancestral whole genome duplication (WGD) event ~100 Mya. Here we use open chromatin as detected through ATAC-seq as an indicator of regulatory activity to investigate the role of TEs in CRE evolution in Atlantic salmon and in the context of the WGD event.

Results: In general, open chromatin regions were >2 fold depleted in TE sequences compared to the overall TE content in the genome. We identified 61,309 regions of open chromatin overlapping TEs (TE-CREs), of which 82% were specific to liver (43%) or brain (39%). Tissue-shared TE-CREs originated from older TE insertions compared to tissue-specific TE-CREs and were four times more likely to overlap with promoters. The relative contribution to the TE-CRE evolution differed among TEs. At the superfamily-level, Tc1-Mariners were highly depleted in TE-CREs relative to their genomic copy numbers, while one DNA transposon (hAT) and two retrotransposon superfamilies (Nimb, Gypsy) were enriched in TE-CREs relative to their copy numbers. At the subfamily-level, 174 TEs (16% of all subfamilies) were classified as TE-CRE 'superspreaders', and these accounted for 46% of all TE-CREs. Most of these (110 subfamilies) were likely active after the WGD, however we did not observe a general burst of superspreader transposition activity coinciding with the WGD. Enrichment analyses of bound transcription factor binding sites (TFBS) showed that 43% and 58% of superspreader TE subfamilies had at least one enriched TFBS in brain and liver, respectively (mean of 15 enriched TFBS in brain and 21 in liver). Finally, TFBS motifs with strong genome-wide occupancy in brain were rarely found within TE sequences.

Conclusion: Most TE-CREs have evolved within old copies of TEs, with little support for the hypothesis that the WGD-associated bursts of TE activity contributed substantially to the evolution of the cis-regulatory landscape. However, we do find compelling evidence for specific TE subfamilies acting as 'superspreaders' of CREs, and that tissue-specific selective pressures have been important in shaping evolution of the TE-associated CRE landscape.

Introduction

Transposable elements (TEs) are a diverse group of mobile genetic elements capable of self-replication or replication through co-optation of host molecular machinery. They can be subdivided into two classes based on their replication mechanism (copy-paste or cut-and-paste) ([Wicker et al. 2007](#)). TEs can be found in almost all known genomes ([Aziz et al. 2010](#); [Bourque et al. 2018](#)), and typically make up a significant proportion of eukaryotic genomes ([Feschotte and Pritham 2007](#)). Their common ability to replicate and move within the host genome, makes TEs potent actors in genome and organismal evolution. TEs contribute to evolution in many ways by; generating novel protein coding genes ([Elisaphenko et al. 2008](#)) or small RNAs ([Qin et al. 2015](#)), modulating chromatin structure ([Diehl et al. 2020](#)), rearranging genome structure ([Bourque et al. 2018](#)), as well as supplying “raw material” for gene regulatory evolution in the form of cis-regulatory elements (CREs) ([Bourque et al. 2018](#); [Cosby et al. 2019](#); [Feschotte 2008](#); [Chuong et al. 2017](#); [Sundaram and Wysocka 2020](#); [Diehl et al. 2020](#)).

A CRE is usually defined as a genomic region containing one or several specific DNA sequence motifs - i.e. short stretches of roughly similar DNA sequence - that modify the regulation of genes. These genes can be in close proximity to the CRE but sometimes also several megabases apart, as in the case of long-range enhancers-promoter interactions ([Visel et al. 2009](#)). CREs impact gene regulation when bound by transcription factors (TFs). TF recognise DNA motifs which are often conserved across species. These proteins can regulate gene expression through various mechanism ([Spitz and Furlong 2012](#)), e.g. by recruiting or hindering molecules that increase transcription (e.g. RNA-polymerases), or by modulating the chromatin structure ([Morimoto 1992](#)).

Studies in mammalian systems have provided deep insights into the role of TEs in CRE-evolution and the potency of TE derived CREs (TE-CREs) to regulate gene expression (reviewed in ([Fueyo et al. 2022](#))). For example, as much as 40% of the genomic binding sites of TFs in mouse and human have been shown to be within TEs ([Sundaram et al. 2014](#)), and as many as 19% of TF binding sites (TFBS) for pluripotency factors are located in TEs ([Kuniarso et al. 2010](#); [Sundaram et al. 2017](#)). Interestingly, in mammals distinct TEs have been associated with gene regulation during development (usually younger TEs) compared to in adult somatic tissues (usually older TEs, reviewed in ([Fueyo et al. 2022](#))), suggesting different evolutionary pressures on TEs with

distinct regulatory roles. Nevertheless, we still have limited understanding of TE-CRE evolution in most non-mammalian vertebrate systems.

The genomes of the Atlantic salmon (*salmo salar*) and other salmonid fish is an interesting study system for TE evolution. Salmonid genomes have a relatively high repeat content (est. 50-60%) ([Lien et al. 2016](#)), and they share a whole genome duplication (WGD) event in a common ancestor around 80-100mya ([Lien et al. 2016](#)). Intriguingly, this WGD coincided with a major invasion of Tc1-mariner elements and this observation has led to the hypothesis that post-WGD TE activity was a major driver of regulatory divergence between gene duplicates. Indeed, in a recent study, we find support for TE accumulation in promoters being associated with gene duplicate regulatory divergence in several ways, both downregulation and liver-specific increase of gene expression levels ([Gillard et al. 2021](#)). However, a systematic interrogation of the TE-CRE landscape evolution, and in particular the coupling to the WGD, is still lacking.

In this study we therefore use ATAC-seq data to define CREs and study TE-CRE evolution in two tissues (brain and liver) with very different rates of gene regulatory evolution ([Wang et al. 2020](#)). Our results do not support a model of TE-driven gene regulatory rewiring shortly after the WGD but find that 10% of the TE-families have TE-CRE 'superspreading' abilities. Furthermore, our study sheds light on selective pressures that shape tissue-specific selective constraints in TE-CRE evolution.

Results

The TE-CRE landscape of Atlantic salmon

In order to investigate the contributions of different TEs to CRE evolution, we first characterised the TE landscape of the salmon genome using an updated version of the existing TE annotation from ([Lien et al. 2016](#)). The total transposable element annotation covered 51.92% of the genome. Consistent with previous findings ([Goodier and Davidson 1994](#); [Lien et al. 2016](#)), the dominating TE group was DNA transposons from the Tc1-Mariner superfamily with >655,000 copies, covering 327 million base pairs, just shy of 10% of the genome (Figure 1A-C). The genomic context of TE insertions was heavily biased towards intergenic and intronic regions, with only a minor fraction of TEs located in promoters (+1000/-200 bp from transcription start sites) (Figure 1D). An exception to this pattern was the Nimb retrotransposon superfamily, for which 18% of the copies were found in promoter regions (Figure 1D).

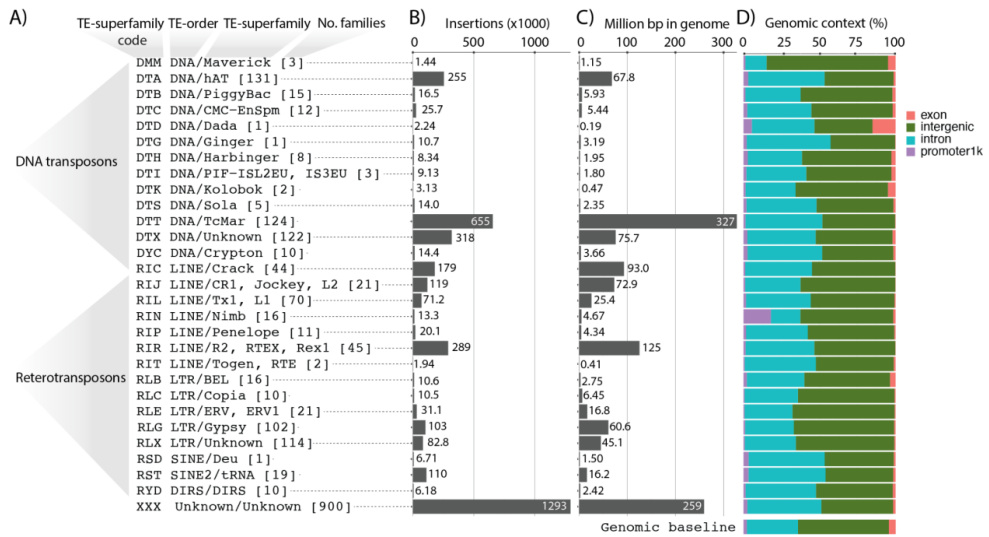


Figure 1. Overview of the genomic TE landscape. A) Superfamily level overview of TE annotations in the Atlantic salmon genome. Number of TE subfamilies per superfamily in square brackets. B) TE insertions per superfamily. C) Annotated base pairs at the TE superfamily level. D) TE annotations (bp proportions) overlapping different genomic contexts. Genomic baseline is the proportion of the entire genomic sequence that is assigned to the four genomic contexts.

Active CREs in tissues and cells are associated with elevated chromatin accessibility [\[McGhee et al. 1981; Keene et al. 1981; Buenrostro et al. 2013\]](#). To study the contribution of TEs to the salmon CRE landscape, we therefore integrated our TE annotation with annotations of accessible chromatin regions that we identified using ATAC-seq data of tissue samples from liver and brain (ENA accession number E-MTAB-9001). This revealed a large depletion of TEs in accessible chromatin. If TEs were randomly distributed in the genome with respect to accessible chromatin, we would expect ~52% of accessible chromatin to overlap with TEs. However, ATAC-seq from brain and liver showed that only <20% of the TE insertions overlapped with regions of accessible chromatin (Figure 2A), with liver having the largest proportion of annotated TEs in accessible chromatin.

less (RIC and DTT) or more (RIN, RLG, DTA) to TE-CREs compared to expectations where TE-CRE numbers are a simple linear function of the number of genomic copies (Figure 2G). For example, although Tc1-Mariners (DTT) is the dominating TE superfamily (representing ~10% of the TE copies in the genome), this superfamily only represented 4% of the TE-CREs. Yet, other superfamilies such as hAT DNA transposons (DTA), as well as Nimb (RIN) and Gypsy (RLG) retrotransposons, are contributing disproportionately to the total TE-CRE landscape (Figure 2G). The most extreme being the Nimb superfamily, which despite being among the smallest superfamilies in terms of total copy numbers ranks seventh (out of 28) on the list of TE-CRE contributors.

Some TE subfamilies are CRE “superspreaders”

TEs within a superfamily share certain structural and biological characteristics, yet there can be large differences among TE subfamilies in when they are active and their potential for contributing to CRE evolution ([Fueyo et al. 2022](#)). Therefore, to further characterise the subfamily-level TE-CRE evolution, we identified TE subfamilies enriched in open chromatin: so-called TE-CRE superspreaders. Among the 1119 TE subfamilies with >500 genomic copies, only 178 TE subfamilies (16%) were identified as being situated in accessible chromatin more often than expected by chance (Figure 3A). About half of these TE subfamilies were enriched in open chromatin in both tissues (88 subfamilies), and 76% (69 subfamilies) of the tissue-specific enrichments were specific to liver (Figure 3A). As the proportion of unclassified repeats (XXX) among the CRE-superspreading TE-families was high (101 subfamilies), manual curation of all 178 TE subfamilies were carried out. This effort resulted in four unclassified subfamilies being discarded, and a reduction of unclassified subfamilies to 34, though many could only be classified confidently to the level of TE order (Supplementary table 1). In total, the 174 curated superspreaders, representing only 11% of the large TE subfamilies (>500 copies), contributed to 46% of all TE-CREs (27,960/61,309) in the genome (Figure 3B).

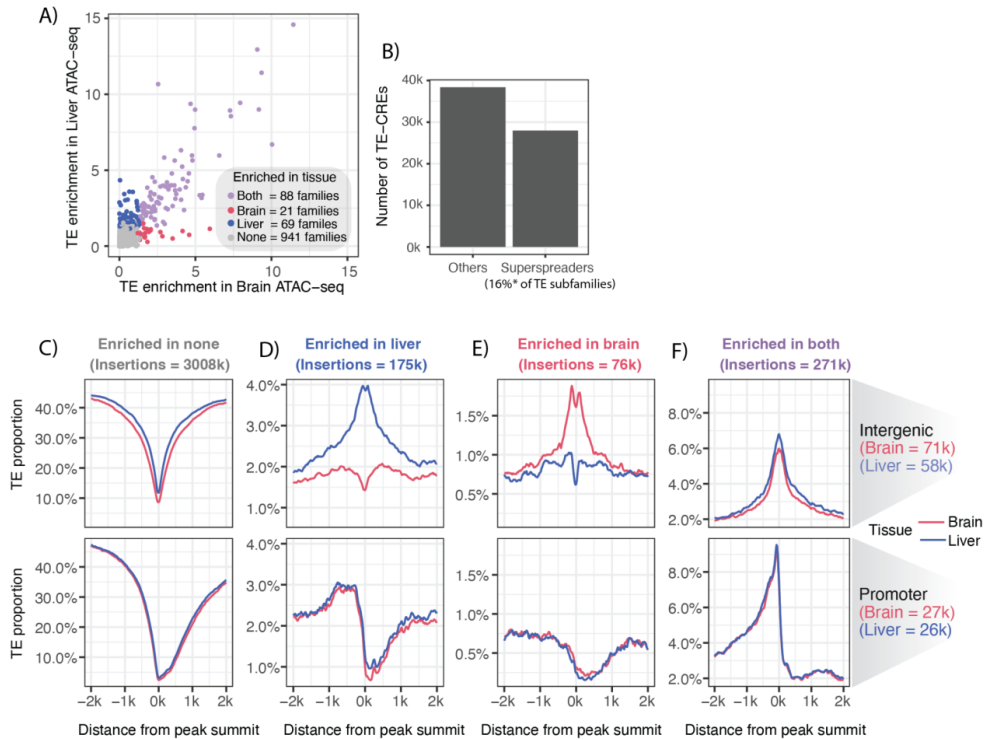


Figure 3. TE subfamilies enriched in open chromatin. A) TE subfamilies plotted according to fold-enrichment within ATAC-seq peaks in brain and liver. TE subfamilies are assigned into categories based on enrichment in liver, brain or both. B) Number of TE-CREs contributed by superspreaders vs. other TE subfamilies. *16% is calculated based on inclusion of large subfamilies (>500 copies) only and removal of four subfamilies after manual curation and inspection of TE consensus alignments. C-F) Proportion of bp overlapping TEs from each enrichment category around peak summits in intergenic or promoter regions.

Further dissecting the genomic context of tissue-shared and tissue-specific TE-CREs originating from superspreaders revealed that tissue-shared TE-CREs were more common closer to TSS (i.e. promoters) while tissue-specific TE-CREs were biased against intergenic regions (putative enhancer regions) (Figure 3 E-F, lower vs. upper panels). This is in line with CRE-evolution in mammals where tissue-shared CREs have been conserved for longer time-scales (Roller et al. 2021).

The evolutionary timing of TE-CRE superspreader activity

One of the key questions we set out to answer was the role of WGD-associated TE activity in TE-CRE evolution. To distinguish between TE subfamilies that were active before or after the WGD, we estimated the time of insertion based on sequence similarity, focussing on the TE-CRE superspreader subfamilies (Figure 4A). Since estimation of the age of an individual TE insertion is inherently difficult, we tried two different methods: a Nearest-neighbour calculation, where each insertion is compared with the five most similar copies, and the similarity to consensus reported from RepeatMasker. We assessed the two methods by testing the expectation that TE subfamilies with more recent activity (more active after WGD) should have a lower proportion of insertions at syntenic position in their duplicated region compared to TE subfamilies with highly divergent copies. Correlation between aligned TE-CREs and the rank order of TE activity age showed highly significant correlation for the Repeatmasker similarity estimate (Figure 4A), but none with the alternative method.

Moving forward with the RepeatMasker similarity estimates, we ranked TEs according to this proxy for TE activity age (Figure 4B). Given previous estimates of ~87% mean sequence similarity of duplicated genomic regions from the salmonid WGD ([Lien et al. 2016](#)), 63% of the CRE-superspreader TE subfamilies were likely mainly active after the WGD. Yet, at the level of subfamilies, we did not find a clear “TE activity shift” around the the ~87% for the TE-CRE superspreaders. In fact, the CRE-superspreader TEs were not active more recently in time than other TEs (Figure 4C). At the superfamily level, we find some differences in the proportion of TE-CRE superspreader subfamilies active prior to or following the salmonid WGD. The majority of CRE superspreader subfamilies belonging to hAT (DTA) and Tc1-Mariner (DTT) superfamilies, were active post WGD, while the highly potent CRE-superspreaders from the Nimb (RIN) superfamily were all likely active prior to the WGD (Figure 4D).

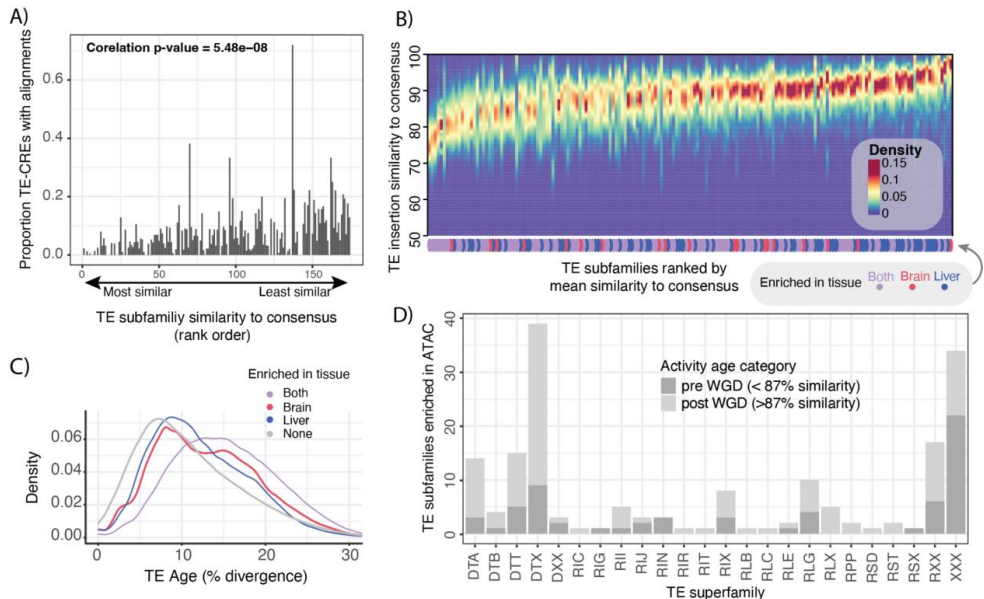


Figure 4. The relative age of TE-superspreaders. A) Barplot of proportion of TE-CREs with a genomic alignment in the duplicated genomic region. P-value of the correlation between proportion of aligned TE-CREs and RepeatMasker similarity estimates are reported. B) Heatmap of the similarity distributions of CRE-superspreader subfamilies to their annotating consensus. C) Distribution of sequence divergence among TE copies from the same TE subfamily. Colours indicate if TEs are classified as a TE-CRE superspreader (both, brain, liver) or not (None). D) Distribution of proportions of TE subfamilies enriched in open chromatin by pre- and post-WGD activity and superfamily.

TE-CRE distribution is not uniform within TE-sequences

One way TEs can impact TE-CRE evolution is if TEs contain particular sequences that function as transcription factor binding sites (TFBSs) and then spread these TFBSs throughout the genome as the TE reproduces [Sundaram and Wysocka 2020; Feschotte 2008; Chuong et al. 2017]. Under this scenario, we would expect that TE-CREs are distributed non-randomly along the TE sequence, and we would expect to observe a strong signal of accessible chromatin locally along the TE consensus sequences. To explore this we first calculated, for each base in consensus sequences of TE-CRE superspreaders, the number of genomic copies of this base in accessible chromatin. We then normalised for the TE consensus length and visualised this in a heatmap (Figure 5A). In general we found that chromatin accessibility varies locally across most of the TE consensus sequence, supporting that particular regions of these TEs harbour sequences that functions as TFBSs. However, it is important to note that consensus length normalisation impacts

the visual impression of how local the signal of chromatin accessibility is across the TE (Figure 5B). Hence, intense local signals in the longest TEs could be similar in size (in base pairs) to more diffuse signals (i.e. spread out) in shorter TEs.

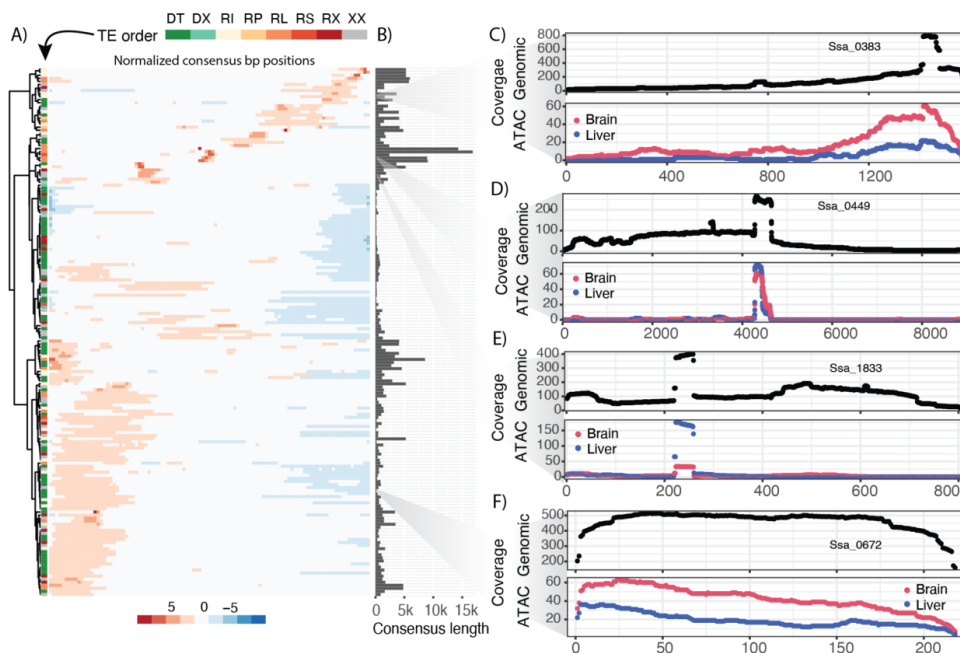


Figure 5. Coverage distribution of base pairs in open chromatin and genomic sequence for TE-CRE superspreaders. A) Heatmap of the ratio between open chromatin sequence- and genomic sequence coverage of each base pair of consensus sequences of “superspreader” TE-families, clustered based on ATAC-seq coverage profile. B-E) Examples of genomic sequence coverage (top panels with black dots) and open chromatin sequence coverage in brain (red) and liver (blue) for consensus sequences from four TE subfamilies selected as examples of different types of profile.

The distribution of accessible chromatin across each TE subfamily consensus sequence does not show strong associations with TE taxonomy (Figure 5A). However, TE subfamilies with accessible chromatin in the last half of the consensus sequence tend to be retrotransposons rather than DNA transposons (top 25% rows in Figure 5A). Manual inspection of individual TE subfamilies confirmed that the majority have highly non-uniform per-base coverage at the genomic level as well as at the level of bases in accessible chromatin. Three examples of this are the two retroelements Ssa_0383 (LTR-Gypsy) and Ssa_0449 (LINE-Togen), as well as the unclassified Ssa_1833 element (Figure 5C-E). The TE-CREs contributed by these TE subfamilies are clearly highly biased towards sequences from particular parts of the TE, and in some cases these parts have tissue-specific chromatin accessibility (Figure 5C and 5E). In contrast, the brain-enriched

Ssa_0672 subfamily (a DNA transposon of undetermined superfamily) is a good example of a TE having more uniform genomic coverage, but with parts of the consensus (in this case the start) biased towards being accessible chromatin and thus contributing more to TE-CREs.

TE-contribution to TFBS motifs

As many TE-families had a clear bias in which regions of the TE-sequence that contributed to TE-CREs (i.e. were in open chromatin, Figure 5), we expected these TE-family regions to contribute to the spread of different transcription factor binding sites (TFBS). To better understand how TEs have contributed to TFBS evolution, we used the ATAC-seq data to perform TF-footprinting. For each potential TFBS in the genome (from the jasper database [\(Castro-Mondragon et al. 2022\)](#)) we determined (i) if this was within a TE (or not) and (ii) if the TFBS was bound by a TF (determined by chromatin accessibility signal and TF-footprinting) or not. Using the entire genome as a background in a Fisher test, we then tested which TFBS were represented more than expected by chance in each TE subfamily. All 746 TFBS motifs from jasper were enriched in at least one TE subfamily in both tissues (FDR p-value<0.01, odds-ratio > 2) and about ~55% (996/1839) and ~66% (1207/1839) of TE subfamilies had at least one enriched TFBS in brain and liver, respectively (Figure 6 A-B). When classifying TEs as having enrichment of TFBSs in one or both tissues, the majority had enriched TFBSs in both tissues (minimum one in each tissue), but only a minority (3%) had brain-specific TFBSs enrichment (Figure 6C).

Since TE-CRE superspreaders (i.e., the 174 TE subfamilies enriched in accessible chromatin left after weeding out technical artefacts) have contributed to a large proportion of the TE-CRE landscape, we asked if this set of TEs is particularly enriched for TFBSs. Indeed, TE-CRE superspreading subfamilies had substantially higher number of enriched TFBSs compared to other TE-families (Figure 6D), and TEs enriched in open chromatin in both tissues had more enriched TFBS compared to TEs enriched in only a single tissue (Figure 6D). The numbers of enriched TFBS varied widely, from >600 to only a few per TE for the TE-CRE superspreader subfamilies. TF-footprint measurements are dependent on the level of chromatin accessibility, and thus, it is possible that very high numbers of TFBS enrichments could be explained by very high levels of chromatin accessibility. This was supported when looking at the relationship between per base mean coverage of accessible chromatin across TE-consensus sequences and the number of enriched TFBSs per TE subfamily (Figure 6E).

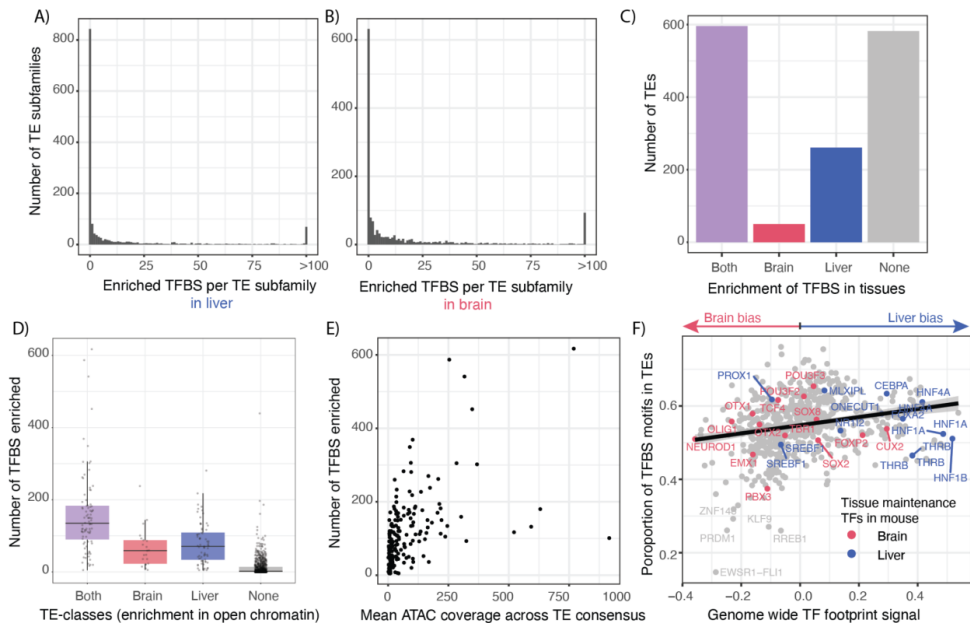


Figure 6. TE contribution to transcription factor binding sites. *A-B) Distributions of enriched TFBSs in TEs in liver and brain tissues. C) Number of TEs enriched with TFBS enriched in liver and brain (both), only single tissues, or no enriched TFBSs. D) Distribution of the number of TFBSs enriched per TE. TEs are categorised by enrichment of TEs in open chromatin. E) Relationship between the number of enriched TFBS per TE and the level of chromatin accessibility, calculated as mean ATAC-seq coverage across the entire TE consensus sequences across liver and brain. F) Correlation between the proportion of TFBS motifs found in TEs using FIMO and the genome wide TF footprint signal for the TFs predicted to bind these TFBSs.*

A key trend is the smaller role of TEs in TFBS evolution in brain compared to the liver (Figure 6 A-D), which is in line with the general patterns from all TE-CRE analyses (Figure 2C, 2D, 3A). One explanation could be that the brain gene regulatory networks are under stronger selective constraints. One expectation from this hypothesis would be that TFBSs with strong brain bias in occupancy of TFs, i.e. regulatory networks with particular importance for brain function, would be depleted in TE sequences (disregarding if the TFBS is in open or closed chromatin). To test this, we correlated the genome-wide TFBS occupancy signal from the TF-footprinting analyses with the proportion of TFBS in TE sequences (Figure 6F). As a quality control, we used known tissue maintenance TFs from mouse (Zhou et al. 2017) and confirmed that we recapitulate tissue-differences in TFBS occupancy for these TFs (Figure 6F). Indeed, the correlation between proportion of TFBS in TEs and genome-wide TFBS occupancy points to a clear trend; if TFBSs have a strong occupancy signature in brain compared to liver (data points to the left in Figure 6F),

these TFBSs were not commonly found in TEs. This result supports the idea that brain-specific selective constraints could explain the tissue-specific differences in TE-CRE evolution.

Discussion

Unique and shared features between salmon and mammalian TE-CRE landscapes

Our efforts to map out the TE-associated CRE-landscape in Atlantic salmon revealed strong similarities to trends seen in studies of mammalian genomes, but also highlighted unique features of the salmonid TE-CRE landscape. Our study identified 61,309 putative TE-derived CREs, which constituted about 15-20% of the total CREs in brain and liver (Figure 2A, 2C). Although quantitative comparisons of TE contribution to CREs are made somewhat complicated by differences in study systems and data, this agrees well with the most readily comparable study to ours ([Sundaram et al. 2014](#)) that finds around 20% of TE-CREs in mammalian cells, as defined by overlaps between ChIP-seq summits and transposable elements. Among the TE-CREs, the vast majority were classified as enhancers in that they were located far from transcription start sites of genes (Figure 2E). This has also been shown in previous studies ([Nishihara 2019](#); [Sundaram and Wysocka 2020](#)).

Relatively few (10%) TE subfamilies had contributed to ~40% of the TE-CREs. These were referred to as TE-CRE 'superspreaders' and classified according to their contribution to tissue-specific or tissue-shared TE-CREs. Consistent with previous findings ([Sundaram and Wysocka 2020](#)), TEs contributing a lot to the CREs landscape are relatively old (Figure 4C), and tissue-shared CREs have higher evolutionary stability (i.e. are older) compared to tissue-specific CREs (Figure 4B and 4C) ([Roller et al. 2021](#)). Hence, regulatorily inactive TEs are younger than TEs involved in liver-specific regulation, which are younger than those involved in brain-specific regulation which are themselves younger than those involved in regulation in both tissues (Figure 4B). This seems to generally reflect the rate of evolution of gene expression seen in mammals, where liver gene regulation evolves fast, brain gene regulation typically evolves much slower, and genes expressed across multiple tissues, such as housekeeping genes, have highly conserved regulation across deep evolutionary distances ([Brawand et al. 2011](#); [Berthelot et al. 2018](#); [Wang et al. 2020](#)).

Previous studies in mammals have found strong association between CRE-evolution and LINE elements in mammals ([Roller et al. 2021](#)). In our study we also find a LINE element, the Nimb superfamily elements, to be highly active in spreading CREs (see RIN elements in Figure 2 F and G), particularly in promoters (Figure 1 D). Nimb elements have been previously noted to have signals of exaptation in amniotes and humans, ([Frith 2022](#)) but the density we find here is remarkable. In contrast with studies of mammals, we do also find that some DNA transposons have contributed a lot to the CRE landscape, namely the hAT transposons (Figure 2G). This result is heavily influenced by a relatively few “superspreader” hAT subfamilies, and the precise reason for this enrichment is not obvious. It bears noting that the earliest discovered member of the hAT superfamily was *Ac*, or *Activator*, and that these were among the earliest transposons discovered ([Rubin et al. 2001](#); [McClintock 1950](#)), with a clear biological effect. Further studies into Nimb and hAT elements should focus on functional validation of these elements as transcriptional regulators.

No acute TE-associated spread of novel CRE elements following WGD

One of our working hypotheses was that the increased activity of Tc1-Mariner elements could have been important in rewiring the gene regulatory landscape following WGD. Many of our “superspreader” TE subfamilies appear to have been significantly active after the salmonid-specific whole genome duplication (Figure 4B), however, there is no clear increase in CRE-spreading TE-activity at the time of/shortly after the WGD (which is estimated to be around 10-13% similarity in Figure 4B based on sequence similarity estimates of duplicated regions from [Lien et al. 2016](#)). More interestingly, the Tc1-Mariner superfamily is quite depleted in TE-CREs compared to the number of genomic insertions (Figure 2G). This is in line with other studies which have shown that the Tc1-Mariner superfamily does not contain many TBFSs ([Zeng et al. 2018](#); [Simonti et al. 2017](#)). Yet, as the sheer number of Tc1-Mariner elements is huge, this superfamily has contributed to ~3500 TE-CREs. Hence a role in regulatory differentiation of duplicated genes after the WGD therefore remains plausible, and warrants a more careful examination integrating gene expression data with TE-CREs.

Selection on TE-CRE repertoire in Atlantic salmon

An interesting question is to what extent evolution of TE-CREs is non-neutral and shaped by selection. Although we cannot measure this experimentally in salmonids, several of our analyses

shed light on the role of selection on TE-CREs. Firstly, negative selection of TE insertions (i.e. selection to remove TEs) is one of the clearest patterns we find. TEs of all kinds are consistently underrepresented in open chromatin (Figure 2A), and there is also a clear signal of increased depletion of TEs towards the summit of ATAC-seq peaks (Figure 2C, Figure 3C). Furthermore, most abundant superfamilies are also significantly depleted in TE-CREs (Figure 2G), though that fact should not be overinterpreted: It may be that negative selection is relaxed for TEs not carrying sequences that function as a CREs or easily mutate into a CRE. Second, we identify clear tissue differences in TE-CREs evolution, with the brain having both absolute and proportionally fewer TE-CREs compared to the liver (Figure 2). One explanation for this could be stronger purifying selection on brain gene regulation than liver gene regulation ([Wang et al. 2020](#)). More interestingly, we also found that highly brain biased TFBS were very scarce in TEs in general (Figure 6F). Based on this observation we hypothesise that the TE sequences themselves are under strong selection not to contain TF binding sites that can interrupt critical functions. In fact, differences in numbers of TE-CREs found between tissues (absolute and proportion) (Figure 2) could therefore, at least partially, reflect that evolutionary successful TEs have sequences with fewer brain TFBS motifs.

We also find striking differences in which regions of the TE consensus that are conserved in many copies in the genome and contribute to the TE-CRE landscape. Among the ‘superspreaders’, the majority of TEs showed the tendency that only smaller regions of the consensus were commonly found in accessible chromatin (Figure 5). Such seemingly ‘non-random’ conservation of only smaller regions of TE could be a signature of differential purifying selection (i.e. conservation). However, it is difficult to distinguish this scenario from a scenario where a TE has undergone erosion during its history of propagation, for example by having some members of the subfamily becoming truncated and non-autonomous.

Methods

TE annotation

The TE library (`ssal_repeats_v5.1`) used to annotate TEs in this study is described in detail in ([Richard Minkley 2018](#)). To generate a TE annotation of the salmon genome (ICSASG v2 assembly) we used RepeatMasker version 4.1.2-p1 ([Smit et al. 2015](#)) ([Smit, Hubley, & Green 2013](#)) under default settings with the `ssal_repeats_v5.1` library. RepeatMasker takes a library of TE consensus sequences and detects whole and fragmented parts of these consensus across the genome using a BLAST-like algorithm. The output file contains the genomic coordinates of the

annotation, and various quality measures such as completeness, and divergence from consensus. The latter measure was used to estimate relative ages of TE activity. TE superfamilies were assigned a three letter tag based on the classifications from Figure 1 in [\(Wicker et al. 2007\)](#). Where there was no obvious categorisation, a literature review was conducted to determine the taxonomic status of a superfamily, and a new tag name introduced based on available letters (so e.g. Nimb is here called RIN as a superfamily of LINE elements).

Manual curation of specific TE subfamilies was done following an adapted version of Goubert et al's process [\(Goubert et al. 2022\)](#), under inspiration from Suh [\(Suh et al. 2018\)](#): Using BLASTn [\(Camacho et al. 2009\)](#), we aligned each transposable element consensus to the genome, extracted the twenty best matches and extended them by 2000bp upstream and downstream. We checked the extended matches against the RepBase [\(Bao et al. 2015\)](#) database using BLASTn and xBLAST with standard settings, before we aligned them using MAFFT's 'einsi' variant [\(Katoh and Toh 2008\)](#). Then, we inspected these alignments for structural features in BioEdit [\(Hall 1999\)](#) and, if conservation across the sequence was deemed interesting, in JalView [\(Waterhouse et al. 2009\)](#). In addition, we ran the TEaid package [\(Goubert 2021\)](#) on each consensus to help guide curation efforts and check each consensus according to its annotation profile and self-alignment. This helped screen for technical noise such as microsatellite sequences near sites of local annotation enrichment. If the annotating consensus was deemed to be incomplete (e.g., if parts of the extended sequence aligned well outside of the consensus), we used Advanced Consensus Generator [\(HIV sequence database 2018\)](#) to generate a new consensus from the best of the extracted alignments for classification.

Measuring TE consensus pileups and ATAC-seq coverage

We produced pileup figures for each TE consensus by using GenomicRanges R package [\(Lawrence et al. 2013; R Core Team 2022\)](#), splitting the base consensus into 1bp tiles using the "tiles" function before counting the numbers of overlaps with the annotation coordinate information for that consensus from RepeatMasker output. The same procedure was followed with both motifs and regions of open chromatin, except that there we first found overlap with (parts of) annotated fragments. Edge cases (such as non-matching annotation lengths and fragment length) were handled by setting the bp coordinate furthest upstream of the actual insertion as the start of the TE fragment, corrected for strandedness or filtered out.

ATAC-seq peak calling

To annotate regions of accessible chromatin we used ATAC-seq data from four brains and livers from Atlantic salmon (ENA project number PRJEB38052). The ATAC-seq reads were mapped to the salmon genome assembly (ICSASG v2, refseq ID: GCF_000233375.1) using BWA-MEM. Genrich v.06 (<https://github.com/jsh58/Genrich>) was then used to call open chromatin regions (also referred to as 'peaks') with default parameters, apart from '-m 20 -j' (minimum mapping quality 20; ATAC-Seq mode). Genrich uses all four replicates to generate peaks, resulting in one set of peaks for each tissue. The summit of each peak is identified as the midpoint of the peak interval with highest significance.

TE-CRE definition

To define TE-CREs we combined the ATAC-seq peak set with our TE annotations and classified an ATAC-seq peak as a TE-CRE if the peak summit is inside a TE-annotation. TE-CREs were defined as shared between tissues if (i) the brain ATAC-seq peak summit was within the liver ATAC-seq peak interval and (ii) both the liver and brain peak summits are inside the same TE annotation.

Defining genomic context

Based on the NCBI gene annotation (refseq ID: GCF_000233375.1), each part of the genome was assigned as promoter, exon, intron or intergenic. Slightly different method used for Figure 1D and for the TE-CREs. For Figure 1D the promoter was defined as 1000 bp upstream to 200 bp downstream of each transcription start site (TSS). Gene annotations can overlap, e.g. because of multiple transcript isoforms, so overlapping annotations were merged by prioritising promoter > exon > intron > intergenic. For TE-CREs (Figure 2E and 3C-F) each peak was classified as promoter if the summit is less than 500bp upstream or downstream from start of gene (i.e. first TSS per gene) or intergenic if summit is more than 500bp from any gene (exon and intron TE-CREs are not specifically mentioned).

Identification of TE subfamilies enriched in open chromatin

To identify TE subfamilies which had contributed more to TE-CREs than expected by chance we counted the number of ATAC-seq peak summits that are inside an annotated TE for each

subfamily and compare that with the total number of bases covered by that TE subfamily genome wide. The enrichment value for each subfamily was calculated as the proportion of summits in TEs divided by the proportion of basepairs in the genome that is annotated as TE. Subfamilies with less than 500 insertions were excluded. We defined TE subfamilies enriched in open chromatin as those containing more ATAC-seq peak summits than chance (binomial, $p < 0.05$), either in the ATAC-seq peak set from liver, brain, or in both tissues.

Estimating evolutionary timing of TE activity

To determine when a TE subfamily was active we use the sequence divergence between insertions, assuming that all new insertions are identical after a burst of activity. There are multiple ways to do this. The easiest is to use the divergence from consensus for each insertion as is reported by the RepeatMasker software ([Smit et al. 2015](#)). We also attempted an alternative approach that compares each insertion with the five most similar copies. This involved extracting the sequences of all insertions in each subfamily using BEDTools getfasta ([Quinlan and Hall 2010](#)), and using BLASTn ([Camacho et al. 2009](#)) with gap costs of 5/2 to blast all against all to identify the five closest matches and taking the mean of sequence identity between them. Since we were interested in whether the TEs were active before or after the salmonid whole genome duplication (WGD) we could use the fact that insertions prior to WGD would have been copied by the WGD while those inserted after will not. By aligning the duplicate genomic regions we got an estimate of how often we observe a WGD derived copy of the TE insertions (Figure 4A). Based on this we determined that the Repeatmasker software output produced the most accurate estimates of the distributions of subfamily-copy similarity.

Transcription factor binding and footprinting

We annotated transcription factor binding sites (TFBS) in two different ways. First, we used FIMO ([Grant et al. 2011](#)) on the whole genome with the JASPAR CORE vertebrates non-redundant motif database (<https://jaspar.genereg.net>). Secondly, we used the TOBIAS software, which uses a FIMO-like TFBS scan but also integrate ATAC-seq data to detect signals of local TF occupancy (i.e. a sudden, local drop in chromatin accessibility) and assigns each TFBS motif a “bound” or “not bound” status. We used the TOBIAS software to estimate a single genome wide TF binding score for each TFBS in liver and brain tissues ([Bentsen et al. 2020](#)).

Testing TE-TFBS enrichment

For each TE subfamily we counted the number of overlaps between each jaspar-TFBS motif (i.e. the entire motif within the annotated TE) and calculated these numbers for TFBSs “bound” by TFs and those “not bound” by TFs as according to the TOBIAS software ([Bentsen et al. 2020](#)) results. We then did the same counting for all TFBS instances outside the particular TE subfamily in question, and used this 2*2 contingency table (Table 1) in a Fisher exact test in R using the `fisher.test()` function in R ([R Core Team 2022](#)).

Table 1. Example of a 2*2 contingency table for fisher exact tests for TFBS-TE associations.

	TE subfamily	All other genomic positions
Not bound TFBS	x	y
Bound TFBS	z	w

Genomic alignments of duplicated genomic regions

To estimate if TEs deposited TE-CREs prior to or after the WGD we produced gene-centric genome alignments including flanking regions (100,000 bp upstream and downstream of gene) between duplicated genes in the Atlantic salmon genome (ICSASG_v2 assembly and NCBI refseq annotation). The identification of duplicated gene pairs from the salmonid WGD were carried out using a combination of orthology-predictions, synteny analyses, and ortholog gene tree parsing and filtering (https://gitlab.com/sandve-lab/salmonid_synteny). Alignments of genomic duplicated regions was done by running Cactus whole genome aligner ([Armstrong et al. 2020](#)) on the genes in each orthogroup including 100,000 bp flanking sequence. Finally, annotations of TE-CREs were used to identify if each TE-CREs had an alignment across duplicated genomic regions and if this alignment overlapped an annotation of the same TE subfamily in both duplicates.

Visualisations and code

We produced plots using base R's ([R Core Team 2022](#)) plot function, as well as the packages `ggplot2` ([Wickham 2016](#)) and `cowplot`. Both the Tidyverse ([Wickham et al. 2019](#)) and `data.table` ([Dowle and Srinivasan 2021](#)) packages were used for analysis, summary statistics and data management. Scripts available at GitLab repo: <https://gitlab.com/sandve-lab/TE-CRE>

Funding

This research was funded by NMBU and the Norwegian Research Council through the projects Transpose (275310) and Rewired (274669).

Bibliography

Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I.T., Novak, A.M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J., et al. (2020). Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* *587*, 246–251. <https://doi.org/10.1038/s41586-020-2871-y>.

Aziz, R.K., Breitbart, M., and Edwards, R.A. (2010). Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res.* *38*, 4207–4217. <https://doi.org/10.1093/nar/gkq140>.

Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* *6*, 11. <https://doi.org/10.1186/s13100-015-0041-9>.

Bentsen, M., Goymann, P., Schultheis, H., Klee, K., Petrova, A., Wiegandt, R., Fust, A., Preussner, J., Kuenne, C., Braun, T., et al. (2020). ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat. Commun.* *11*, 4267. <https://doi.org/10.1038/s41467-020-18035-1>.

Berthelot, C., Villar, D., Horvath, J.E., Odom, D.T., and Flicek, P. (2018). Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat. Ecol. Evol.* *2*, 152–163. <https://doi.org/10.1038/s41559-017-0377-2>.

Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., et al. (2018). Ten things you should know about transposable elements. *Genome Biol.* *19*, 199. <https://doi.org/10.1186/s13059-018-1577-z>.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., et al. (2011). The evolution of gene expression levels in mammalian organs. *Nature* *478*, 343–348. <https://doi.org/10.1038/nature10532>.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* *10*, 1213–1218. <https://doi.org/10.1038/nmeth.2688>.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* *10*, 421. <https://doi.org/10.1186/1471-2105-10-421>.

Castro-Mondragon, J.A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R.B., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., et al. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* *50*, D165–D173. <https://doi.org/10.1093/nar/gkab1113>.

Chuong, E.B., Elde, N.C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* *18*, 71–86. <https://doi.org/10.1038/nrg.2016.139>.

Cosby, R.L., Chang, N.-C., and Feschotte, C. (2019). Host-transposon interactions: conflict, cooperation, and cooption. *Genes Dev.* 33, 1098–1116. <https://doi.org/10.1101/gad.327312.119>.

Diehl, A.G., Ouyang, N., and Boyle, A.P. (2020). Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes. *Nat. Commun.* 11, 1796. <https://doi.org/10.1038/s41467-020-15520-5>.

Dowle, M., and Srinivasan, A. (2021). Package “data.table” (GitHub).

Elisaphenko, E.A., Kolesnikov, N.N., Shevchenko, A.I., Rogozin, I.B., Nesterova, T.B., Brockdorff, N., and Zakian, S.M. (2008). A dual origin of the Xist gene from a protein-coding gene and a set of transposable elements. *PLoS ONE* 3, e2521. <https://doi.org/10.1371/journal.pone.0002521>.

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* 9, 397–405. <https://doi.org/10.1038/nrg2337>.

Feschotte, C., and Pritham, E.J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.* 41, 331–368. <https://doi.org/10.1146/annurev.genet.40.110405.090448>.

Frith, M.C. (2022). Paleozoic protein fossils illuminate the evolution of vertebrate genomes and transposable elements. *Mol. Biol. Evol.* 39. <https://doi.org/10.1093/molbev/msac068>.

Fueyo, R., Judd, J., Feschotte, C., and Wsocka, J. (2022). Roles of transposable elements in the regulation of mammalian transcription. *Nat. Rev. Mol. Cell Biol.* 23, 481–497. <https://doi.org/10.1038/s41580-022-00457-y>.

Gillard, G.B., Grønvold, L., Røsæg, L.L., Holen, M.M., Monsen, Ø., Koop, B.F., Rondeau, E.B., Gundappa, M.K., Mendoza, J., Macqueen, D.J., et al. (2021). Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. *Genome Biol.* 22, 103. <https://doi.org/10.1186/s13059-021-02323-0>.

Goodier, J.L., and Davidson, W.S. (1994). Tc1 transposon-like sequences are widely distributed in salmonids. *J. Mol. Biol.* 241, 26–34. <https://doi.org/10.1006/jmbi.1994.1470>.

Goubert, C. (2021). TE-Aid: Annotation helper tool for the manual curation of transposable element consensus sequences (Github: Clement Goubert).

Goubert, C., Craig, R.J., Bilat, A.F., Peona, V., Vogan, A.A., and Protasio, A.V. (2022). A beginner’s guide to manual curation of transposable elements. *Mob. DNA* 13, 7. <https://doi.org/10.1186/s13100-021-00259-7>.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27, 1017–1018. <https://doi.org/10.1093/bioinformatics/btr064>.

Hall, T.A. (1999). BioEdit: A User-Friendly Biological Sequence Alignment Editor and Analysis Program for Windows 95/98/NT.

Katoh, K., and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinformatics* 9, 286–298. <https://doi.org/10.1093/bib/bbn013>.

Keene, M.A., Corces, V., Lowenhaupt, K., and Elgin, S.C. (1981). DNase I hypersensitive sites in *Drosophila* chromatin occur at the 5’ ends of regions of transcription. *Proc Natl Acad Sci USA* 78, 143–146. <https://doi.org/10.1073/pnas.78.1.143>.

Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H., and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* 42, 631–634. <https://doi.org/10.1038/ng.600>.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* 9, e1003118. <https://doi.org/10.1371/journal.pcbi.1003118>.

- Lien, S., Koop, B.F., Sandve, S.R., Miller, J.R., Kent, M.P., Nome, T., Hvidsten, T.R., Leong, J.S., Minkley, D.R., Zimin, A., et al. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature* 533, 200–205. <https://doi.org/10.1038/nature17164>.
- McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci USA* 36, 344–355. <https://doi.org/10.1073/pnas.36.6.344>.
- McGhee, J.D., Wood, W.I., Dolan, M., Engel, J.D., and Felsenfeld, G. (1981). A 200 base pair region at the 5' end of the chicken adult beta-globin gene is accessible to nuclease digestion. *Cell* 27, 45–55. [https://doi.org/10.1016/0092-8674\(81\)90359-7](https://doi.org/10.1016/0092-8674(81)90359-7).
- Morimoto, R.I. (1992). Transcription factors: positive and negative regulators of cell growth and disease. *Curr. Opin. Cell Biol.* 4, 480–487. [https://doi.org/10.1016/0955-0674\(92\)90015-5](https://doi.org/10.1016/0955-0674(92)90015-5).
- Nishihara, H. (2019). Retrotransposons spread potential cis-regulatory elements during mammary gland evolution. *Nucleic Acids Res.* 47, 11551–11562. <https://doi.org/10.1093/nar/gkz1003>.
- Qin, S., Jin, P., Zhou, X., Chen, L., and Ma, F. (2015). The role of transposable elements in the origin and evolution of miRNAs in human. *PLoS ONE* 10, e0131365. <https://doi.org/10.1371/journal.pone.0131365>.
- Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. <https://doi.org/10.1093/bioinformatics/btq033>.
- Richard Minkley, D. (2018). Transposable Elements in the Salmonid Genome. Master thesis. University of Victoria.
- Roller, M., Stamper, E., Villar, D., Izuogu, O., Martin, F., Redmond, A.M., Ramachanderan, R., Harewood, L., Odom, D.T., and Flicek, P. (2021). LINE retrotransposons characterize mammalian tissue-specific and evolutionarily dynamic regulatory regions. *Genome Biol.* 22, 62. <https://doi.org/10.1186/s13059-021-02260-y>.
- Rubin, E., Lithwick, G., and Levy, A.A. (2001). Structure and evolution of the hAT transposon superfamily. *Genetics* 158, 949–957.
- R Core Team (2022). R: A language and environment for statistical Computing <https://www.r-project.org/>.
- Simonti, C.N., Pavlicev, M., and Capra, J.A. (2017). Transposable Element Exaptation into Regulatory Regions Is Rare, Influenced by Evolutionary Age, and Subject to Pleiotropic Constraints. *Mol. Biol. Evol.* 34, 2856–2869. <https://doi.org/10.1093/molbev/msx219>.
- Smit, Hubley, & Green, A., R., & P. (2013). RepeatMasker open v4 (repeatmasker).
- Smit, A.F.A., Hubley, R., and Green, P. (2015). RepeatMasker.
- Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* 13, 613–626. <https://doi.org/10.1038/nrg3207>.
- Suh, A., Smeds, L., and Ellegren, H. (2018). Abundant recent activity of retrovirus-like retrotransposons within and among flycatcher species implies a rich source of structural variation in songbird genomes. *Mol. Ecol.* 27, 99–111. <https://doi.org/10.1111/mec.14439>.
- Sundaram, V., and Wysocka, J. (2020). Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 375, 20190347. <https://doi.org/10.1098/rstb.2019.0347>.
- Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., Snyder, M.P., and Wang, T. (2014). Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res.* 24, 1963–1976. <https://doi.org/10.1101/gr.168872.113>.
- Sundaram, V., Choudhary, M.N.K., Pehrsson, E., Xing, X., Fiore, C., Pandey, M., Maricque, B., Udawatta, M., Ngo, D., Chen, Y., et al. (2017). Functional cis-regulatory modules encoded by

mouse-specific endogenous retrovirus. *Nat. Commun.* *8*, 14550. <https://doi.org/10.1038/ncomms14550>.

Visel, A., Rubin, E.M., and Pennacchio, L.A. (2009). Genomic views of distant-acting enhancers. *Nature* *461*, 199–205. <https://doi.org/10.1038/nature08451>.

Wang, Z.-Y., Leushkin, E., Liechti, A., Ovchinnikova, S., Mößinger, K., Brüning, T., Rummel, C., Grützner, F., Cardoso-Moreira, M., Janich, P., et al. (2020). Transcriptome and translome co-evolution in mammals. *Nature* *588*, 642–647. <https://doi.org/10.1038/s41586-020-2899-z>.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* *25*, 1189–1191. <https://doi.org/10.1093/bioinformatics/btp033>.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* *8*, 973–982. <https://doi.org/10.1038/nrg2165>.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis (Use R)* (Cham: Springer).

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Golemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. *JOSS* *4*, 1686. <https://doi.org/10.21105/joss.01686>.

Zeng, L., Pederson, S.M., Kortschak, R.D., and Adelson, D.L. (2018). Transposable elements and gene expression during the evolution of amniotes. *Mob. DNA* *9*, 17. <https://doi.org/10.1186/s13100-018-0124-5>.

Zhou, Q., Liu, M., Xia, X., Gong, T., Feng, J., Liu, W., Liu, Y., Zhen, B., Wang, Y., Ding, C., et al. (2017). A mouse tissue transcription factor atlas. *Nat. Commun.* *8*, 15089. <https://doi.org/10.1038/ncomms15089>.

HIV Sequence Database <https://www.hiv.lanl.gov>.

Supplementary material

Table S1: **Curation notes and classification.** Every “superspreader” subfamily has been inspected manually as per the procedure in *Materials and Methods*. ‘Consensus’ is the ID of the annotating consensus in question, ‘Machine’ is the automatic classification, ‘tag’ is the post-curation three-letter ID, and ‘len’ is the estimated length of alignment. If ‘len’ is empty, it is because the length of the alignment is not easily found, typically because the curation alignment is rather fragmented.

consensus	Machine	Note	tag	len
Ssa_0015	DNA	Non-autonomous	DTX	79
Ssa_0018	DNA	Non-autonomous	DTA	879

Ssa_0020	DNA	Non-autonomous	DTT	458
Ssa_0021	DNA	Non-autonomous	DTA	142
Ssa_0022	DNA	Non-autonomous	DTT	142
Ssa_0035	DNA		DTX	768
Ssa_0044	DNA		DTB	1449
Ssa_0062	DNA/CMC- EnSpm	No signal	DTX	
Ssa_0070	DNA/Ginger	No signal	DTX	
Ssa_0090	DNA/hAT	Autonomous	DTA	3150
Ssa_0102	DNA/hAT	Non-autonomous	DTA	2886
Ssa_0105	DNA/hAT	Non-autonomous	DTA	486
Ssa_0108	DNA/hAT	Autonomous	DTA	2880
Ssa_0120	DNA/hAT		DTT	3361
Ssa_0134	DNA/hAT		DTA	1215
Ssa_0135	DNA/hAT		DTA	1408
Ssa_0142	DNA/hAT		DTT	372
Ssa_0144	DNA/hAT		DTA	2141
Ssa_0147	DNA/hAT	Maybe autonomous	DTT	1587
Ssa_0152	DNA/hAT		DTT	1233
Ssa_0156	DNA/hAT	Non-autonomous	DTA	814

Ssa_0165	DNA/hAT	Autonomous	DTA	1734
Ssa_0175	DNA/hAT	Non-autonomous	DTA	864
Ssa_0236	DNA/TcMar	Maybe autonomous	DTT	3617
Ssa_0262	DNA/PiggyBac	Autonomous	DTB	3088
Ssa_0264	DNA/Sola	No signal	DTX	
Ssa_0341	LINE/L2	Autonomous	RIR	3568
Ssa_0342	LINE/Nimb	Autonomous	RII	5224
Ssa_0343	LINE/Nimb	Autonomous	RII	5227
Ssa_0344	LINE/Nimb	Non-autonomous	RII	1540
Ssa_0345	LINE/Nimb	Maybe autonomous	RII	5099
Ssa_0350	LINE/Penelope		RPP	3619
Ssa_0351	LINE/Penelope		RPP	4068
Ssa_0357	LINE/Rex1	non-autonomous	RIJ	520
Ssa_0368	LINE/Rex1	Autonomous	RIJ	3392
Ssa_0383	LINE/Togen		RIX	1572
Ssa_0404	LTR/BEL		RLB	14638
Ssa_0412	LTR/Copia	Maybe autonomous	RLC	4799
Ssa_0425	LTR/ERV1		RLE	16871
Ssa_0449	LTR/Gypsy		RLG	

Ssa_0456	LTR/Gypsy		RLG	
Ssa_0458	LTR/Gypsy		RLG	
Ssa_0475	LTR/Gypsy		RLG	
Ssa_0479	LTR/Gypsy		RLG	
Ssa_0480	SINE/Deu		RSD	
Ssa_0497	SINE2/trNA		RST	
Ssa_0498	SINE2/trNA		RST	
Ssa_0552	LTR		RLG	
Ssa_0630	DNA		DTX	
Ssa_0643	DNA		DTX	
Ssa_0657	DNA		DTX	
Ssa_0658	DNA		DTX	
Ssa_0662	DNA		DTX	
Ssa_0664	DNA		DTX	
Ssa_0668	DNA		DTX	
Ssa_0669	DNA		DTX	
Ssa_0670	DNA		DTX	
Ssa_0672	DNA		DTX	
Ssa_0684	DNA		DTX	

Ssa_0685	DNA		DTX	
Ssa_0686	DNA		DTX	
Ssa_0693	DNA		DTT	
Ssa_0696	DNA		DTX	
Ssa_0701	DNA		DTX	
Ssa_0727	DNA/hAT		DTA	
Ssa_0776	DNA/PiggyBac		DTB	
Ssa_0796	LINE/CR1		RIC	
Ssa_0844	LINE/Nimb		RIN	
Ssa_0845	LINE/Nimb		RIN	
Ssa_0846	LINE/Nimb		RLG	
Ssa_0849	LINE/Nimb		RIN	
Ssa_0872	LINE/Rex1		RIJ	
Ssa_0875	LINE/RTEX		RIT	
Ssa_0907	LTR/ERV		RLE	
Ssa_0934	LTR/Gypsy		RLG	
Ssa_0958	LTR/Gypsy		RLG	
Ssa_0959	LTR/Gypsy		RLG	
Ssa_0972	Unknown	Satellite noise	XXX	

Ssa_1131	Unknown		XXX	
Ssa_1137	Unknown	Satellite noise	XXX	
Ssa_1146	Unknown		XXX	
Ssa_1153	Unknown		DTX	
Ssa_1159	Unknown		RXX	
Ssa_1161	Unknown	no	DTX	637
Ssa_1169	Unknown	no	DTX	155
Ssa_1174	Unknown	no	DTX	175
Ssa_1175	Unknown	maybe	RLX	1385
Ssa_1181	Unknown	maybe	XXX	1637
Ssa_1183	Unknown	no	RIX	537
Ssa_1184	Unknown	no	DTX	400
Ssa_1185	Unknown	no	DXX	441
Ssa_1186	Unknown	no	DTX	508
Ssa_1206	Unknown	maybe	XXX	
Ssa_1211	Unknown		XXX	153
Ssa_1218	Unknown		RIX	853
Ssa_1239	Unknown		DTA	167
Ssa_1246	Unknown		RXX	

Ssa_1250	Unknown		RLX	1006
Ssa_1258	Unknown		XXX	
Ssa_1259	Unknown		RIX	568
Ssa_1263	Unknown		RXX	181
Ssa_1270	Unknown		XXX	
Ssa_1282	Unknown		RIX	1348
Ssa_1301	Unknown		RII	
Ssa_1305	Unknown		DTT	
Ssa_1327	Unknown		DTX	
Ssa_1345	Unknown		DXX	536
Ssa_1354	Unknown		RXX	502
Ssa_1367	Unknown		DTX	128
Ssa_1379	Unknown	cryptic	XXX	1755
Ssa_1381	Unknown	cryptic	XXX	
Ssa_1382	Unknown	Interesting CA-motif ca 400bp	DTT	1428
Ssa_1385	Unknown		RXX	213
Ssa_1388	Unknown		RXX	262
Ssa_1395	Unknown		DTX	128
Ssa_1396	Unknown		XXX	779

Ssa_1397	Unknown		XXX	470
Ssa_1399	Unknown		RLX	1457
Ssa_1402	Unknown		DTX	620
Ssa_1403	Unknown		RXX	513
Ssa_1418	Unknown		RXX	424
Ssa_1419	Unknown		RXX	374
Ssa_1420	Unknown		RXX	834
Ssa_1421	Unknown		XXX	
Ssa_1423	Unknown		XXX	693
Ssa_1426	Unknown		DTX	615
Ssa_1427	Unknown	not autonomous	RXX	251
Ssa_1429	Unknown	not autonomous	XXX	497
Ssa_1430	Unknown		XXX	
Ssa_1431	Unknown		DTT	941
Ssa_1432	Unknown		DTX	1013
Ssa_1433	Unknown	not autonomous	DTX	494
Ssa_1436	Unknown	not autonomous	DTX	804
Ssa_1437	Unknown		XXX	
Ssa_1439	Unknown		XXX	

Ssa_1446	Unknown	not autonomous	DTX	185
Ssa_1456	Unknown		XXX	393
Ssa_1471	Unknown		NOT TE	
Ssa_1507	Unknown		RXX	1493
Ssa_1523	Unknown	not autonomous	RXX	510
Ssa_1524	Unknown		RXX	1224
Ssa_1570	Unknown		RXX	
Ssa_1574	Unknown	not autonomous	RIX	758
Ssa_1575	Unknown	not autonomous	DTT	745
Ssa_1584	Unknown		DTX	
Ssa_1603	Unknown	not autonomous	RXX	855
Ssa_1631	Unknown		XXX	
Ssa_1647	Unknown		XXX	
Ssa_1657	Unknown	No signal	XXX	
Ssa_1674	Unknown	not autonomous	DXX	124
Ssa_1681	Unknown	No signal	XXX	
Ssa_1682	Unknown		XXX	
Ssa_1683	Unknown	not autonomous	DTX	114
Ssa_1684	Unknown		XXX	351

Ssa_1699	Unknown	not autonomous	XXX	137
Ssa_1751	Unknown		RLG	1349
Ssa_1778	Unknown	No signal	XXX	
Ssa_1781	Unknown	No signal	XXX	
Ssa_1782	Unknown	not autonomous	DTX	145
Ssa_1784	Unknown	needs new consensus	DTX	192
Ssa_1807	Unknown		RIX	
Ssa_1825	Unknown		XXX	
Ssa_1833	Unknown		XXX	806
Ssa_1839	Unknown		XXX	
Ssa_1850	Unknown		DTT	987
Ssa_1899	Unknown	not autonomous	DTX	
Ssa_1901	Unknown	no	RIX	small
Ssa_1925	Unknown	no	RLX	92
Ssa_1959	Unknown		XXX	
Ssa_1973	Unknown		DTB	785
Ssa_1976	Unknown	maybe	RLX	1313
Ssa_1978	Unknown		DTT	872
Ssa_1983	Unknown	maybe	DTX	1250

Ssa_1984	Unknown	no	DTX	191
Ssa_1985	Unknown	no	DTT	1011
Ssa_1990	Unknown	no	DTX	839
Ssa_2003	Unknown		RXX	
Ssa_2004	Unknown	maybe	RSX	

PAPER 3

Structural Variation in Atlantic salmon Strongly Correlates with Telomere Associated Tandem Repeats

Øystein Mosen, Kristina R. S. Stenløkk, Simen R. Sandve & Sigbjørn Lien

INTRODUCTION

Genomic structural variants (SVs) are defined as differences in the genome structure larger than 50 base pairs (bps), encompassing deletions, insertions, inversions and duplications (Mahmoud et al. 2019). SVs are common in the genome, affecting many times more base pairs than single nucleotide polymorphisms (Kosugi et al. 2019), and may have significant phenotypic consequences, e.g. by impacting gene regulation and gene copy number variation (Henkel et al. 2019; Jeffares et al. 2017; Perry et al. 2007). SVs are also shown to contribute substantially to speciation (Zhang et al. 2021) and local adaptation (Wellenreuther et al. 2019). Despite their potentially large impact on trait variation, the SV landscape are not well characterized in many species.

One common source of SVs is repetitive DNA (Garrido-Ramos 2017; Miga 2019; Riethman 2008; Tørresen et al. 2019), typically constituting 50-90% of eukaryote genomes (de Koning et al. 2011; Garrido-Ramos 2017; Liu et al. 2019; Mehrotra & Goyal 2014; Platt et al. 2016). There are two main categories of repetitive DNA: tandem repeats (TRs) and transposable elements (TEs). TRs are defined as adjacently repeated stretches of DNA with the length of the repeated unit (array size) and sequence composition varying widely (Benson 1999; Lu et al. 2021; Sulovari et al. 2019). A common class of TRs is satellite DNA (satDNA), which are large arrays (here defined as at least 10 000bps) of repeated DNA motifs, typically enriched around telomeric and centromeric regions playing important roles in cell division-related processes such as recombination and cytokinesis (Garrido-Ramos 2017). For the purposes of this text, satDNA are large, contiguous regions of highly repetitive DNA whereas TRs include phenomena such as local duplications and incomplete motifs. TRs are believed to give rise to SVs in different ways, including polymerase slippage (Pearson et al. 2005; Raz et al. 2019), tandem duplication (Farnoud et al. 2019) and template switching (Course et al. 2020). The second main class of repetitive DNA, TEs, are mobile, self-replicating elements present in all eukaryotic genomes

(Bourque et al. 2018). They span from dozens to thousands of bps in size and replicate through either copy-paste (type 1: retrotransposon) or cut-and-paste (type 2: DNA transposon) mechanisms. The role of TEs in generating SV is mostly linked to their transposition activity, resulting in insertion and deletion variation (Mun et al. 2021). Scattered and highly similar TE-copies can also result in ectopic recombination giving rise to inversions or translocations (Kent et al. 2017).

Traditionally, genome wide detection of SVs has relied mainly on the analysis of short-read sequencing data, having clear limitations especially for detecting longer inversions and insertions (Bertolotti et al. 2020; Mahmoud et al. 2019). With the introduction of long-read sequencing technologies, the ability to detect SVs has improved drastically, both in terms of sensitivity and precision (Mahmoud et al. 2019; Sedlazeck et al. 2018). Numerous studies involving such technology have detected ten to hundreds of thousands of novel SVs at a wide size range (Alonge et al. 2020; Beyter et al. 2021; Chawla et al. 2021; Kou et al. 2020; Liu et al. 2020; Weissensteiner et al. 2020). As SVs are widespread in repeat dense regions in which long reads increase the quality (Logsdon et al. 2020; Murigneux et al. 2020), the increase in SV-detection efficiency seems especially pertinent for studying such regions (Weckselblatt & Rudd 2015).

The Atlantic salmon (*Salmo salar*) is an economic and culturally important species distributed in the North Atlantic Ocean. The ancestor of salmonid fishes underwent a whole genome duplication 90-120 Mya (Gundappa et al. 2021; Macqueen & Johnston 2014) which generated increased functional redundancy coinciding with a burst of TE-activity (Lien et al. 2016). Compared to other vertebrates, salmonid genomes are highly repetitive with an estimated repeat content of around 58-60% (McCluskey & Postlethwait 2015), from which ~65% of the repeats have a clear TE origin (Lien et al. 2016) However, it is important to note that satellites and tandem repeats have so far not been characterized in a long-read Atlantic Salmon genome assembly, which is known to have a better representation of these classes of repetitive DNA. Recently, short-read sequencing data was used to study the SV landscape across 493 wild and farmed Atlantic salmon. Among ~165,000 SVs identified, only about 9% (15,483) were classified as high confidence SVs (Bertolotti et al. 2020), highlighting the need for long-read sequencing approaches to portray the landscape of SVs in the duplicated Atlantic salmon genome.

In this study, we aim to address the shortcomings of short-read sequencing technologies by utilising nanopore long-read sequencing methods to characterise the repeat landscape in

Atlantic salmon sampled across the three phylogeographic groups in Europe. We describe the genome wide distribution of SVs and explore the role of genome duplication and a complex repeat landscape in the evolution of SVs.

RESULTS

Assembly

To characterise the landscape of SVs and repeat content in the Atlantic salmon genome (*Salmo salar*), we first constructed a chromosome-level reference genome using Oxford Nanopore sequencing technology (ONT) and the Flye (Kolmogorov et al. 2019) genome assembler. This resulted in 4,222 contigs with a N50 of more than 28 Mbp (Table 1). The draft assembly was error polished using Illumina short-reads and scaffolded into chromosome sequences using Hi-C (Ssal_v3.1; GCA_905237065.2). Compared to the previous Atlantic salmon reference genome (ICSASG_v2), based on sanger sequencing and Illumina short-reads, the assembly contiguity as measured by contig N50 increased over 450-fold (Table 1). The number of contigs was reduced from 965,912 to 4,222. Gene completeness as measured by BUSCO also increased slightly from 92.3 % to 95.7 % (Table 1).

Table 1. Assembly metrics for the first public Atlantic salmon genome reference (ICSASG_v2, [GCA_000233375.4](#)) and the new long-read based genome reference (Ssa_v3.1, [GCA_905237065.2](#)).

Assembly	Total length (Gbp)	Chromosomes anchored (Gbp)	Contigs	Contig N50 (Mbp)	BUSCO (%)
ICSASG_v2	3.06	2.08	965 912	0.04	92.3
Ssal_v3.1	2.76	2.50	4 222	28.06	95.7

Furthermore, the ICSASG_v2 suffered from high proportion of contigs not assigned to chromosomes. Although the new assembly had less sequence in total, it contained 420 Mbp more sequence anchored to chromosomes (Table 1).

The structural variation landscape across European Atlantic salmon

To characterise the SV-landscape in Atlantic Salmon of European origin. We generated ONT sequence data for six wild Atlantic salmon, representing the three phylogeographic groups in

Europe (Atlantic, Barents/White Sea and Baltic) (Fig 1 A). The mean read depth for the samples ranged between 16-42x genome coverage (Table S1). By long read based SV calling, we identified a total of 717,234 SVs, A PCA on the SV dataset confirmed the expected phylogeographic relatedness between these samples (Fig 1B).

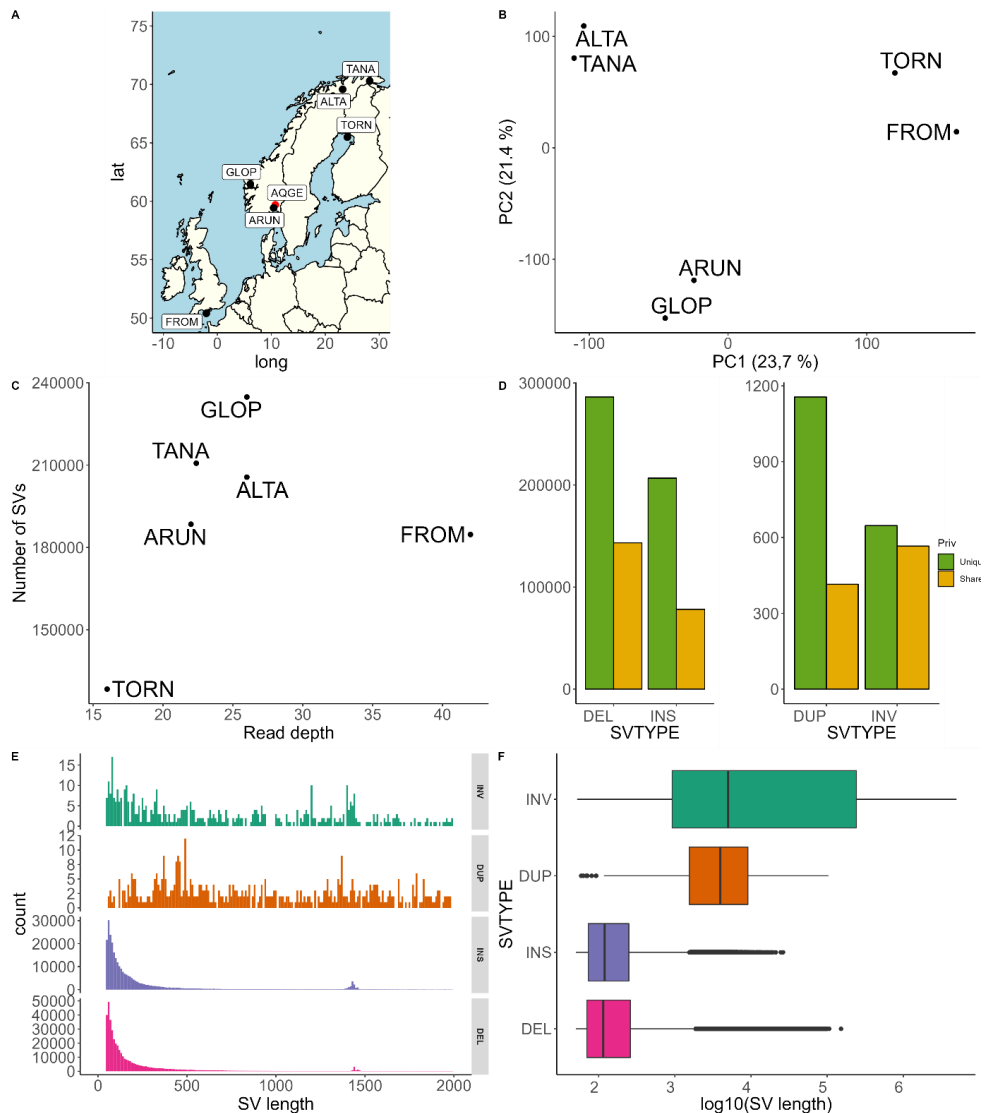


Figure 1. Long read genome assemblies and structural variation. A) Sampling sites of seven long-read sequences Atlantic salmon samples used for SV detection, including the sample (AQGE) used for construction reference genome

sequence (Ssal_v3.1). B) PCA plot of shared SVs. C) Number of SVs plotted per read depth. SV size distribution for four SV classes D) Number of SVs by class and uniqueness E) histogram up to 2kbp and D) boxplot on log10-scale.

The number of SVs per fish ranged 1.8-fold, from 128,369 to 234,822 (Fig 1C), and weakly correlates (spearman 0.32) with sequencing coverage. Among all SVs, deletions (59,9 %, n = 429,556) and insertions (39,7 %, n = 284,894) made up the majority of the entries, with duplications (0,2 % n = 1,571) and inversions (0,2 %, n = 1,213) contributing modestly to the total SV landscape. The majority (69%) of all SVs were unique to one sample, with the proportion unique and shared SV varying between SV types (Fig 1D). Deletions, insertions and duplications had substantially higher proportion unique SVs, while the inversion variants were about the same proportion unique and shared (Fig 1D/E).

The SV length distributions also varied considerably between classes of SVs (Fig 1E and F). Duplications and inversions had median lengths of 3,960bp and 5,019bp, while the median insertions and deletions were 120bp and 1150bp, respectively. We found a distinct peak near 1,400 bp for insertions and deletions (Fig1E) (Bertolotti et al. 2020; Lien et al. 2016).

Contribution of genomic repeats to the SV-landscape

To further investigate the relationship between SVs and repetitive DNA, we first preformed both transposable element- and tandem repeat annotation of the Ssal_v3.1 genome reference (GCA_905237065.2). The total repeat content (TRs and TEs) of the genome was 60-70%. TEs made up 40.61%, with the largest TE group being Tc1-mariner elements (see Table S2). Apart from Tc1-mariner elements (11,6% of bps in the genome), there is a relatively large number of unclassified DNA transposons (5,8%), and LINE-Jockey-like elements (8,3%), a superfamily previously reported mainly in arthropods (Tambones et al. 2019). TE-families were unevenly distributed in divergence from annotating consensus (Figure 2C), and annotation was heavily fragmented (Figure 2B).

Using two different approaches for annotating tandem repeats, we estimated the satellite DNA content to be 20,2% and total tandem repeat content 34% of the genome; the difference is likely made up of shorter arrays of less-prevalent repeats, as well as local duplications. Both categories of non-TE repeat content were consistently enriched near the telomeres (Figure 2D). No other variable seemed relevant for predicting the genome distribution of TRs.

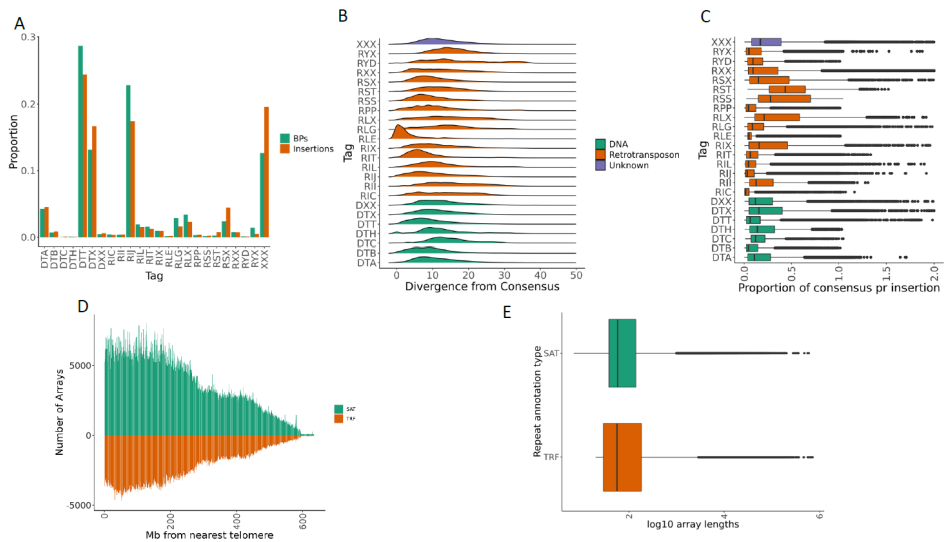


Figure 2: Summary of the repeat annotation. A) Distributions of different superfamilies of TEs by number of insertions and base-pair proportion. B) Divergence from applied consensus per superfamily. The greater the divergence, the greater the estimated age of a given insertion. C) How intact the consensus are in the annotated insertions. The relatively fragmented nature of the TEs is partially a sign of sequence length inflation in the TE library and partially a simple consequence of RepeatMasker's annotation algorithm, which tends to break up insertions into smaller, proximate 'chunks'. D) The number of discrete tandem repeat arrays per million base pairs from the nearest telomere, both active and ancestral. The satDNA annotation has a greater number of discrete arrays altogether. There is a general tendency for repetitive DNA to accumulate around the telomeres, in accordance with expectations. E) Array lengths according to annotations. The somewhat more conservative SAT annotation tends a little longer and a little more consistent, but most arrays are relatively small.

Next, we examined the overlap between SVs and different categories of repeat. TEs as a general category were found to be somewhat underrepresented in indels and inversions (Table 2). However, some families were notably enriched compared to the mean, in particular sequences associated with the peak identified in Figure 1C. This element bears close resemblance to a previously reported active PiggyBac-like transposon (Bertolotti et al. 2020; Lien et al. 2016). The element has clear sequence similarity to the EF685967.1 element (de Boer et al. 2007) (19% of deletions in this size range matched with >99% identity over their full length) and was masked by three similarly overlapping consensus sequences. Additionally, one LINE family showing sequence similarity to the Keno-1_SSa element was mildly overrepresented in

deletions compared to its genomic average, and an unclassified TE was overrepresented in duplications.

For TRs the pattern was reversed, with higher proportions of TRs compared to the genomic average (Table 2). The proportion of SVs overlapping TRs was substantially higher than the proportion of SV base-pair content overlapping TRs, indicating that a large proportion of smaller SVs correlates with variation in TRs. This was not detected in either TEs or satDNA more specifically. The fraction of total base pair overlap between deletions and satDNA has increased from 14,3 % in the previous Atlantic salmon SV annotation (Bertolotti et al. 2020) to 37 % here. TR and satDNA density increased towards both historic (i.e. telomeres belonging to chromosomes which have since merged with others and no longer act as telomeres in cell division) and presently active telomeres, whereas TE density was reduced in these regions (Figure 3a). Indeed, proximity to telomeres appears to be the main predictor of both TR density and TR-SV overlap (Figure 3b). Interestingly, this also holds for historic telomeres.

Table 2: Proportion of overlap between SVs and repeat DNA annotations. Numbers in parenthesis are fraction of total bases in each SV-type overlapping repeat DNA. As insertions are single genomic coordinates, the proportion of annotations and proportions of bases are the same.

Repeat DNA classes	Deletions	Insertions	Inversions	Duplications	Total genome portion
Tandem repeats	0.83 (0.52)	0.77*	0.13 (0.38)	0.51 (0.27)	0.34
Satellite DNA	0.48 (0.37)	0.36*	0.35 (0.31)	0.24 (0.22)	0.18
Transposab le elements	0.19 (0.19)	0.19*	0.45 (0.36)	0.41 (0.41)	0.40

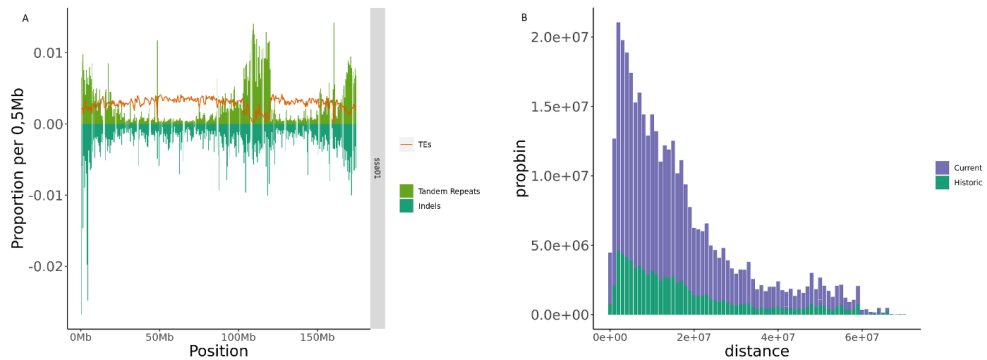


Figure 3. Genomic positioning of repeats A) The distribution of aggregated TE annotations, insertions and deletions and detected TRF output on chromosome 1 of the assembly. Each bin represents the proportion of the whole category within 500000bps. The ancestral telomere from around 105Mb to around 120Mb is enriched in both indels and TRs but depleted for TEs. B) Average number of base-pairs of TR/SV-overlap per Mb depending on distance from telomeres. TR/SV-overlap accumulates around both historic and current telomeres.

TRs and SVs are enriched in telomeric regions

A recent study found overrepresentation of TRs in telomeric regions of the human genome (Beyter et al. 2021). The intricate structure of salmonid chromosomes following whole genome duplication makes the Atlantic salmon TR landscape particularly interesting as it allows one to examine the fate of repeat-dense regions such as centromeres and telomeres after rediploidisation. In order to investigate the potential effects of overlaps between TRs and SVs on rediploidisation, we plotted positions of both TRs and SVs in the chromosome sequences. Figure 4 shows the duplicated nature of the genome with homeologous chromosome arms following Ss4R. When we compare the pattern of TRs and SVs (the outer and middle track, respectively), we see a similar enrichment of both TRs and SVs in telomeric regions. We also see the same pattern for 10 chromosome regions serving as telomeres (historical telomeres) prior to the known chromosome fusions in the Atlantic salmon genome (Phillips & Rab 2001). Specifically, these fusion points are located on ssa01;119Mbp, ssa09;106Mbp, ssa10;59Mbp, ssa11;59Mbp, ssa14;51Mbp, ssa16;59Mbp, ssa17;47Mbp, ssa18;49Mbp, ssa19;32Mbp and ssa20;43Mbp in Ssal_v3.1. When comparing Ssal_v3.1 with ICSASG_v2, we see an overrepresentation of TRs and SVs (Figure 4, Tracks c. and d.) in chromosome regions in Ssal_v3.1 missing in ICSASG_v2 (red colour in Figure 4, Track a.), suggesting that large proportion of TRs and SVs were undetectable in the ICSASG_v2 assembly. Within the 420Mbp of chromosome sequence added in ssa1_v3.1

there is a 1.23-fold enrichment of satellite elements, but a 0.8-fold depletion in TE derived sequences. These regions are also enriched in structural variants, containing 38.5% of structural variations in the assembly in regions totalling around 23% of assembly length. There is a significant correlation between TRs-count and distance to telomere for the given chromosome arm (p -value $< 2.2e-16$) with an adjusted R-squared of 0.53.



Figure 4. Circos plot links showing homeologous regions in the Atlantic salmon genome. Track a. regions with average mapping depth of less than 1 (red) per Mbp between ICSASG_v2 and Ssa1_v3.1. Black circles represent the centromere position Track b. Sequence similarity between homologous blocks in the genome ranging from 80% (green) to 100% (red) Track c. SVs per Mbp. Track d. TRs per Mbp.

DISCUSSION

In this study we generated a new long-read assembly of Atlantic salmon from aquaculture and sequenced an additional 6 samples with long-read sequencing technology representing wild populations spanning the North-Atlantic distribution (Figure 1A). In total we identified 717,234 SVs, a massive increase from 15,5k SVs previously reported in Atlantic salmon (Bertolotti et al. 2020), in line with increased sensitivity of long-read assemblies to detect SVs (Mahmoud et al. 2019). We found significant variation in number of SVs detected in different assemblies (128k-235k). The assembly from the Finnish TORN population had much fewer SVs compared to the other assemblies (Figure 1C), but this is likely not explained by increased genetic similarity between assemblies, but rather technical artifacts. Firstly, the TORN assembly had the lowest genomic coverage of all assemblies (17x, Table S1). It is known that Identification of SVs is highly dependent on sequencing coverage, and the methods used in this study require >20x genomic coverage for efficient SV-detection (Jiang et al. 2021; Sedlazeck et al. 2018). Secondly, the TORN individual was base called using an earlier version of the base caller Guppy (Wick et al. 2019) compared to the other individuals in this study. As sequence read quality is also known to impact SV detection negatively (Jiang et al. 2021), this could also have contributed to the low SV count in the TORN individual.

Our new catalogue of SV was overwhelmingly dominated by deletions and insertion (99%, Figure 1D), which is very similar to what was recently found in *Coregonus*, another salmonid fish genus (Mérot et al. 2022). One of our initial hypotheses was that TE-activity had contributed to the SV landscape. Indeed, we find that 19% of SVs overlap TE, but except for the previously described Tc1-Mariner expansion in salmonids (Krasnov et al. 2005; Lien et al. 2016), we did not identify any large-scale expansions TE-derived SV in Atlantic salmon. This is different from what is found in the recent comparison of the young species pair within *Coregonus*, where several retroelements had contributed significantly to the insertion/deletion SV landscape. Also in rainbow trout, TEs is reported to have contributed substantially to the SV landscape (Liu et al. 2021), with recent expansions of Tc1-Mariner elements as well as Gypsy retroelements making up a substantial proportion of the SV catalogue. However, the latter study on rainbow trout used only short-reads data to call SVs, and this is likely to have biased the SV detection.

The recent developments in long-read sequencing have revolutionized our ability to sequence and assemble tandem repeats (Nurk et al. 2022). This is also evident for our new long-read Atlantic salmon genome assembly which had increased levels of satellite DNA in the fraction of the genome that was unique to the long-read assembly. We find a strong association between SVs and tandem repeats, with 83% and 77% of deletion and insertions overlapping TR annotations,

respectively. This large proportion of SVs that overlap with TRs is much higher than previous estimates based on short-read identification of SVs in Atlantic salmon (Bertolotti et al. 2020). This highlights the inherent difficulty of assembly, read mapping, and annotation associated with TR DNA. Telomeres are typically dense in TR-DNA and could be potent sources of TR-SV (Audano et al. 2019; Garrido-Ramos 2017). Indeed, this was also the case in our study where both historic (Lien et al. 2016) (telomeres from prior to karyotype shuffling after WGD) and current telomeres were particularly enriched in TR-associated SVs (Figure 3B).

MATERIALS AND METHODS

Genome assembly

To build the *Atlantic salmon* reference genome (GCA_905237065.2), we sequenced a male from Norwegian aquaculture strain (AquaGen) to approx. 70x genome coverage on a long-read Oxford Nanopore PromethION sequencing platform. Long-read library was prepared using the SQK-LSK109 kit following the Genomic DNA by ligation protocol (Table S3). Base calling was performed with Guppy (Wick et al. 2019) (versions used are listed in Table S3). We made a draft assembly with Flye (v2.7 and v2.8) (Kolmogorov et al. 2019), with different sequence overlaps (5, 10, 15, 20 and 30kb) that was combined into one assembly by merging contig ends overlapping with at least 20kb. Overlaps were determined by LASTZ alignments (Harris 2007). The combined assembly was polished with long-reads using PEPPER (v 0.0.6) (Shafin et al. 2021) and Illumina short-reads using pilon (v1.23) (Walker et al. 2014). Hi-C data was used to build chromosome sequences juicer pipeline (v1.5.7, (Durand et al. 2016)).

Long-read based structural variant detection

The SV detection was based on mapping nanopore long reads from six samples of Atlantic salmon of European origin (with a read depth 17-27x) to Ssal_v3.1 using Winnowmap2 v(2.0) (Jain et al. 2020) (with the --MD flag). Sam-files were sorted and converted into BAM-files with samtools (version 1.3.1) (Li et al. 2009). The SV-detection was performed with three SV-calling softwares (Sniffles (Sedlazeck et al. 2018), SVIM (Heller & Vingron 2019) and NanoVar (Tham et al. 2020) using with default settings. Only SVs detected by at least two methods were retained. The filtered SVs merged with Jasmine (Kirsche et al. 2021). To account for the variable read depth between samples when running Sniffles, the minimum number of reads that support a SV to be reported (-s) was set to 1/3 of the median read depth, calculated using Mosdepth (version 0.2.6) (Pedersen & Quinlan 2018). The SVs of type "breakpoint" were removed with custom R scripts, and read names were added to preserve insertion sequences in the final VCF. For

refining the insertion sequences, Iris (Kirsche et al. 2021) was employed. Finally, we merged the VCFs across sample with Jasmine (Kirsche et al. 2021) . Custom scripts are available at https://github.com/kristinastenlokk/long_read_SV.

Repeat annotation

The Atlantic salmon genome is known to be heavily repetitive, and rich in both transposable elements and tandem repeats. A good annotation is therefore crucial. We annotated repetitive DNA using a library-based method.

Repeat Masking

We annotated the repeat content of Ssal_v3.1 in two rounds. First, we produced a library of satellite DNA consensus sequences from TAREAN (v 2.3.7) (Novák et al. 2017), a part of the RepeatExplorer pipeline (Novák et al. 2013). Next, we merged it with output of Tandem Repeat Finder (Benson 1999) filtered to take arrays of at least 10 kb and maximum period size 2 kb after a reciprocal RepeatMask run to filter out any redundancy. This was then annotated using RepeatMasker (Smit et al. 2015) on default settings and masked again using our TE library.

A library of TE consensus sequences for Atlantic salmon was already available from the ICSASG_v2 assembly (Lien et al. 2016). However, since the long-read-based assembly is likely to detect additional TE families, we decided to make a new annotation on the Ssal_v3.1 assembly. To this end, we used three *de novo* pipelines to generate TE libraries: RepeatModeler2 (Flynn et al. 2020), REPET's (Flutre et al. 2011) TEdenovo and PASTEC (Hoede et al. 2014) suites, and a merged EDTA/DeepTE method (Bell et al. 2021). Each of these libraries was reciprocally BLASTed using BLAST+/2.10.1 (Camacho et al. 2009), masked using RepeatMasker, and grouped according to the 80-80-80 rule of thumb (i.e., 80% similarity over 80% of the sequence down to at least 80bp) (Wicker et al. 2007). Every extant library entry was compared to the *de novo* libraries and corrected if there was consensus among the automated classifications that the sequence was misclassified. In addition, every satellite DNA or "simple repeat" entry was removed, and finally well-characterised sequences not present in the previously extant library but detected by at least two *de novo* methods had their most reasonable-looking consensus added to the library.

Finally, in order to cast a somewhat wider net in our search for repetitive DNA, we simply ran Tandem Repeat Finder on the assembly, extracted all arrays of more than three complete monomers and took those as a separate TR annotation. This means that our SatDNA annotation is the product of this earlier two-round masking, whereas the TR annotation is the product of another, much simpler method.

The assembly has therefore been masked twice: Once for satDNA and simple repeats and then specifically for transposable elements.

Analyses were carried out in R version 4.1.2 using the Tidyverse (v 1.3.1), GenomicRanges (v 1.44.0), data.table (v 1.14.2), maps (v 3.4.0), cowplot (v 1.1.1), RColorBrewer (v 1.1-2) and ggplot2 (v 3.3.5) packages. The circos plot was constructed using circis (v 0.69.8) (Krzywinski et al. 2009). Scripts for repeat analyses mainly available at https://gitlab.com/sandve-lab/rep_SV_scripts.

- Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. E., Dougherty, M. L., Nelson, B. J., Shah, A., Dutcher, S. K., et al. (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell*, 176 (3): 663-675.e19.
- Bell, E. A., Butler, C. L., Oliveira, C., Marburger, S., Yant, L. & Taylor, M. I. (2021). Transposable element annotation in non-model species: The benefits of species-specific repeat libraries using semi-automated EDTA and DeepTE de novo pipelines. *Molecular Ecology Resources*, n/a (n/a).
- Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research*, 27 (2): 573-580.
- Bertolotti, A. C., Layer, R. M., Gundappa, M. K., Gallagher, M. D., Pehlivanoglu, E., Nome, T., Robledo, D., Kent, M. P., Røsæg, L. L., Holen, M. M., et al. (2020). The structural variation landscape in 492 Atlantic salmon genomes. *Nature Communications*, 11 (1): 5176.
- Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H. P., Bjornsson, E., Jonsson, H., Atlason, B. A., Kristmundsdottir, S., Mehringer, S., Hardarson, M. T., et al. (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nature Genetics*, 53 (6): 779-786.
- Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., et al. (2018). Ten things you should know about transposable elements. *Genome Biology*, 19 (1): 199.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10 (1): 421.
- Course, M. M., Gudsnuk, K., Smukowski, S. N., Winston, K., Desai, N., Ross, J. P., Sulovari, A., Bourassa, C. V., Spiegelman, D., Couthouis, J., et al. (2020). Evolution of a Human-Specific Tandem Repeat Associated with ALS. *The American Journal of Human Genetics*, 107 (3): 445-460.
- de Boer, J. G., Yazawa, R., Davidson, W. S. & Koop, B. F. (2007). Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. *BMC Genomics*, 8 (1): 422.
- de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A. & Pollock, D. D. (2011). Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLOS Genetics*, 7 (12): e1002384.
- Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S. & Aiden, E. L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell systems*, 3 (1): 95-98.

- Farnoud, F., Schwartz, M. & Bruck, J. (2019). Estimation of duplication history under a stochastic model for tandem repeats. *BMC Bioinformatics*, 20 (1): 64.
- Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. (2011). Considering Transposable Element Diversification in De Novo Annotation Approaches. *PLOS ONE*, 6 (1): e16526.
- Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C. & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, 117 (17): 9451.
- Garrido-Ramos, M. A. (2017). Satellite DNA: An Evolving Topic. *Genes*, 8 (9): 230.
- Gundappa, M. K., To, T.-H., Grønvold, L., Martin, S. A. M., Lien, S., Geist, J., Hazlerigg, D., Sandve, S. R. & Macqueen, D. J. (2021). Genome-Wide Reconstruction of Rediploidization Following Autopolyploidization across One Hundred Million Years of Salmonid Evolution. *Molecular Biology and Evolution*.
- Harris, R. S. (2007). *Improved pairwise alignment of genomic DNA*: The Pennsylvania State University.
- Heller, D. & Vingron, M. J. B. (2019). SVIM: structural variant identification using mapped long reads. *Bioinformatics*, 35 (17): 2907-2915.
- Henkel, J., Saif, R., Jagannathan, V., Schmocker, C., Zeindler, F., Bangerter, E., Herren, U., Posantzis, D., Bulut, Z., Ammann, P., et al. (2019). Selection signatures in goats reveal copy number variants underlying breed-defining coat color phenotypes. *PLOS Genetics*, 15 (12): e1008536.
- Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V. & Quesneville, H. (2014). PASTEC: An Automatic Transposable Element Classification Tool. *PLOS ONE*, 9 (5): e91929.
- Jain, C., Rhie, A., Hansen, N., Koren, S. & Phillippy, A. M. (2020). A long read mapping method for highly repetitive reference sequences. *bioRxiv*: 2020.11.01.363887.
- Jeffares, D. C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J. & Sedlazeck, F. J. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications*, 8 (1): 14061.
- Jiang, T., Liu, S., Cao, S., Liu, Y., Cui, Z., Wang, Y. & Guo, H. (2021). Long-read sequencing settings for efficient structural variation detection based on comprehensive evaluation. *BMC Bioinformatics*, 22 (1): 552.
- Kent, T. V., Uzunović, J. & Wright, S. I. (2017). Coevolution between transposable elements and recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372 (1736): 20160458.
- Kirsche, M., Prabhu, G., Sherman, R., Ni, B., Aganezov, S. & Schatz, M. C. (2021). Jasmine: Population-scale structural variant comparison and analysis. *bioRxiv*: 2021.05.27.445886.
- Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37 (5): 540-546.
- Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M. & Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology*, 20 (1): 117.
- Krasnov, A., Koskinen, H., Afanasyev, S. & Mölsä, H. (2005). Transcribed Tc1-like transposons in salmonid fish. *BMC Genomics*, 6 (1): 107.
- Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. & Marra, M. (2009). CIRCOS: an information aesthetic for comparative genomics. *Genome research*, 19: 1639-45.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25 (16): 2078-2079.
- Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., Hvidsten, T. R., Leong, J. S., Minkley, D. R., Zimin, A., et al. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature*, 533 (7602): 200-205.

- Liu, Q., Li, X., Zhou, X., Li, M., Zhang, F., Schwarzacher, T. & Heslop-Harrison, J. S. (2019). The repetitive DNA landscape in *Avena* (Poaceae): chromosome and genome evolution defined by major repeat classes in whole-genome sequence reads. *BMC Plant Biology*, 19 (1): 226.
- Liu, S., Gao, G., Layer, R. M., Thorgaard, G. H., Wiens, G. D., Leeds, T. D., Martin, K. E. & Palti, Y. (2021). Identification of High-Confidence Structural Variants in Domesticated Rainbow Trout Using Whole-Genome Sequencing. *Frontiers in Genetics*, 12.
- Logsdon, G. A., Vollger, M. R. & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, 21 (10): 597-614.
- Lu, T.-Y., Munson, K. M., Lewis, A. P., Zhu, Q., Tallon, L. J., Devine, S. E., Lee, C., Eichler, E. E., Chaisson, M. J. P. & The Human Genome Structural Variation, C. (2021). Profiling variable-number tandem repeat variation across populations using repeat-pangenome graphs. *Nature Communications*, 12 (1): 4250.
- Macqueen, D. J. & Johnston, I. A. (2014). A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proceedings of the Royal Society B: Biological Sciences*, 281 (1778): 20132881.
- Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C. & Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. *Genome Biology*, 20 (1): 246.
- McCluskey, B. M. & Postlethwait, J. H. (2015). Phylogeny of Zebrafish, a "Model Species," within *Danio*, a "Model Genus". *Molecular Biology and Evolution*, 32 (3): 635-652.
- Mehrotra, S. & Goyal, V. (2014). Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. *Genomics, Proteomics & Bioinformatics*, 12 (4): 164-171.
- Mérot, C., Stenløkk, K. S. R., Venney, C., Laporte, M., Moser, M., Normandeau, E., Árnýasi, M., Kent, M., Rougeux, C., Flynn, J. M., et al. (2022). Genome assembly, structural variants, and genetic differentiation between lake whitefish young species pairs (*Coregonus* sp.) with long and short reads. *Molecular Ecology*, n/a (n/a).
- Miga, K. H. (2019). Centromeric Satellite DNAs: Hidden Sequence Variation in the Human Population. *Genes*, 10 (5): 352.
- Mun, S., Kim, S., Lee, W., Kang, K., Meyer, T. J., Han, B.-G., Han, K. & Kim, H.-S. (2021). A study of transposable element-associated structural variations (TASVs) using a de novo-assembled Korean genome. *Experimental & Molecular Medicine*, 53 (4): 615-630.
- Murigneux, V., Rai, S. K., Furtado, A., Bruxner, T. J. C., Tian, W., Harliwong, I., Wei, H., Yang, B., Ye, Q., Anderson, E., et al. (2020). Comparison of long-read methods for sequencing and assembly of a plant genome. *GigaScience*, 9 (12): g1aa146.
- Novák, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. (2013). RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, 29 (6): 792-793.
- Novák, P., Ávila Robledillo, L., Koblížková, A., Vrbová, I., Neumann, P. & Macas, J. (2017). TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Research*, 45 (12): e111-e111.
- Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. *Science*, 376 (6588): 44-53.
- Pearson, C. E., Edamura, K. N. & Cleary, J. D. (2005). Repeat instability: mechanisms of dynamic mutations. *Nature Reviews Genetics*, 6 (10): 729-742.
- Pedersen, B. S. & Quinlan, A. R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics (Oxford, England)*, 34 (5): 867-868.
- Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A., Mountain, J. L., Misra, R., et al. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*, 39 (10): 1256-1260.
- Phillips, R. & Rab, P. (2001). Chromosome evolution in the Salmonidae (Pisces): an update. *Biological Reviews*, 76 (1): 1-25.

- Platt, R. N., II, Blanco-Berdugo, L. & Ray, D. A. (2016). Accurate Transposable Element Annotation Is Vital When Analyzing New Genome Assemblies. *Genome Biology and Evolution*, 8 (2): 403-410.
- Raz, O., Biezuner, T., Spiro, A., Amir, S., Milo, L., Titelman, A., Onn, A., Chapal-Ilani, N., Tao, L., Marx, T., et al. (2019). Short tandem repeat stutter model inferred from direct measurement of in vitro stutter noise. *Nucleic acids research*, 47 (5): 2436-2445.
- Riethman, H. (2008). Human Telomere Structure and Biology. *Annual Review of Genomics and Human Genetics*, 9 (1): 1-19.
- Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A. & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, 15 (6): 461-468.
- Shafin, K., Pesout, T., Chang, P.-C., Nattestad, M., Kolesnikov, A., Goel, S., Baid, G., Kolmogorov, M., Eizenga, J. M., Miga, K. H., et al. (2021). Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nature Methods*, 18 (11): 1322-1332.
- Smit, A., Hubley, R. & Green, P. (2015). *RepeatMasker Open-4.0. 2013–2015*.
- Sulovari, A., Li, R., Audano, P. A., Porubsky, D., Vollger, M. R., Logsdon, G. A., Consortium, H. G. S. V., Warren, W. C., Pollen, A. A., Chaisson, M. J. P., et al. (2019). Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 116 (46): 23243-23253.
- Tambones, I. L., Haudry, A., Simão, M. C. & Carareto, C. M. A. (2019). High frequency of horizontal transfer in Jockey families (LINE order) of drosophilids. *Mobile DNA*, 10 (1): 43.
- Tham, C. Y., Tirado-Magallanes, R., Goh, Y., Fullwood, M. J., Koh, B. T. H., Wang, W., Ng, C. H., Chng, W. J., Thiery, A., Tenen, D. G., et al. (2020). NanoVar: accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing. *Genome Biology*, 21 (1): 56.
- Tørresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., Gruca, A., Grynberg, M., Kajava, A. V., Promponas, V. J., et al. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research*, 47 (21): 10994-11006.
- Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., et al. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE*, 9 (11): e112963.
- Weckselblatt, B. & Rudd, M. K. (2015). Human Structural Variation: Mechanisms of Chromosome Rearrangements. *Trends in Genetics*, 31 (10): 587-599.
- Wellenreuther, M., Mérot, C., Berdan, E. & Bernatchez, L. (2019). Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. *Molecular Ecology*, 28 (6): 1203-1209.
- Wick, R. R., Judd, L. M. & Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology*, 20 (1): 129.
- Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8 (12): 973-982.
- Zhang, L., Reifová, R., Halenková, Z. & Gompert, Z. (2021). How Important Are Structural Variants for Speciation? *Genes*, 12 (7).

LORecess

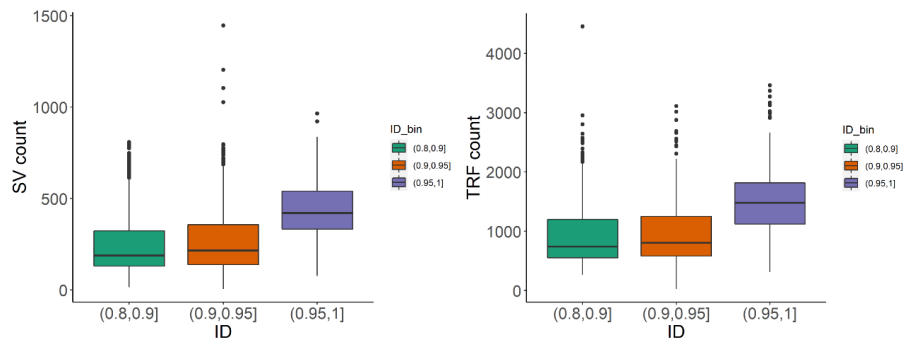


Figure S1: enrichment of variation and repetitiveness by similarity across ohnologous regions.

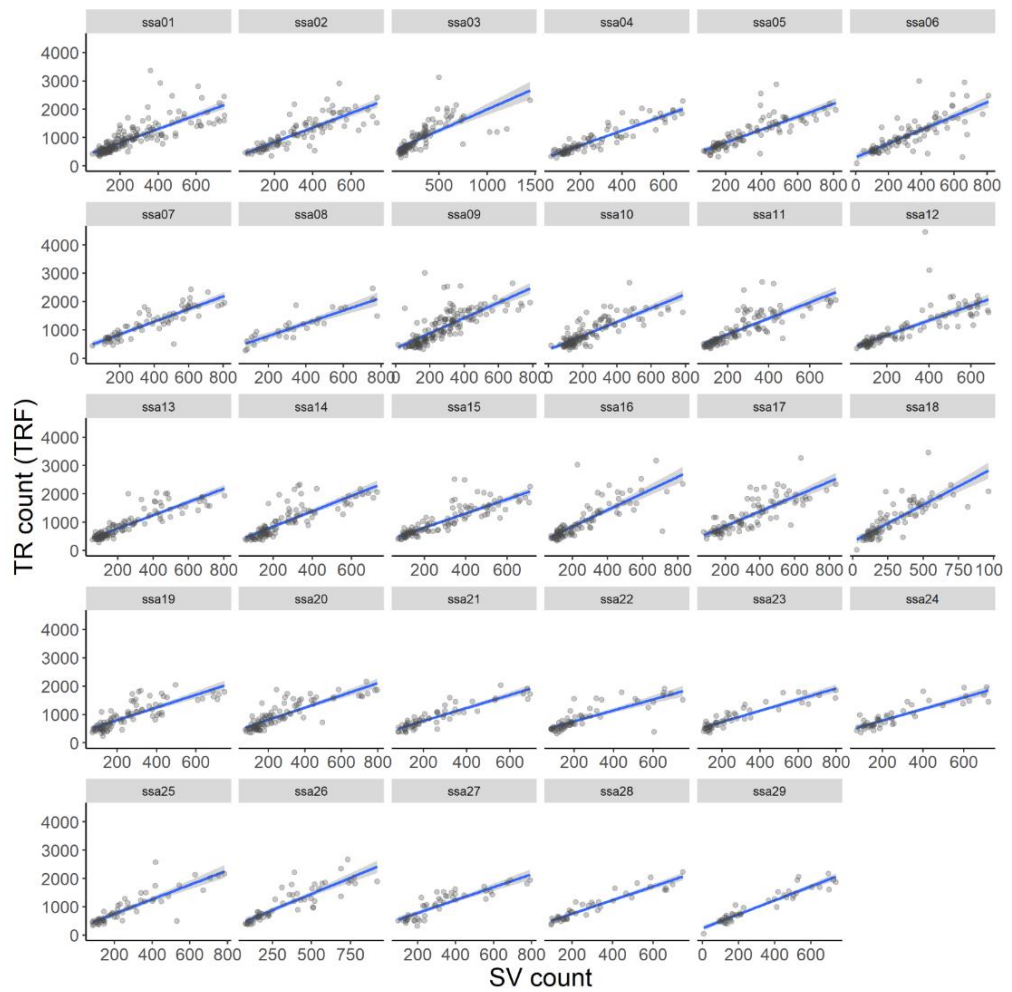


Figure S2: Correlation between SV and TR count per mega base by chromosome

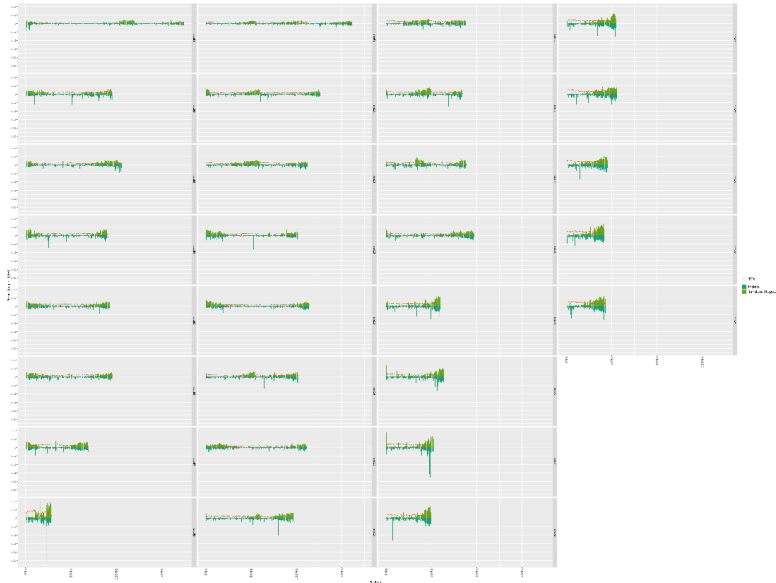


Figure S3: All TE/SV/indeldistros by binned chromosomes (can be made BIG BIG BIG)

Table S1: Metadata for Atlantic salmon samples sequenced with Nanopore long-read technology. The samples represent the phylogeographical groups Atlantic (ATL), Barents/White Sea (BWS), Baltic (BAL) and one aquaculture sample (AQU).

Sample ID	River name	Phylogeographical group	Country	Gender	Population type	Lat, Long	Mean read depth
AQGE	-	AQU	Norway	Male	Aquaculture	-	70
GLOP	Gloppenelva	ATL	Norway	Male	Anadromous	61.46N, 6.12E	26
ARUN	Årungselsva	ATL	Norway	Male	Anadromous	59.43N, 10.43E	22
ALTA	Altaelva	BWS	Norway	Male	Anadromous	69.58N, 23.22E	26
TANA	Tanaelva	BWS	Norway	Male	Anadromous	70.29N, 28.23E	22
FROM	River Frome	ATL	UK	Male	Anadromous	50.41N, 2.05W	42

TORN	Tornio	BAL	Finland	Male	Anadromous	65.49N, 24.09E	16
------	--------	-----	---------	------	------------	-------------------	----

Table S2: Distribution of TEs by family, incl. summaries. E.g. “DNA transposons” is the sum of all categorised and uncategorised DNA transposons in the genome.

TYPE	NO. INSERTIONS	BPs	% OF GENOME
Retroelements	989281	385493162	15,42
SINEs	184238	26950020	1,08
LINEs	654096	287294018	11,49
R1/LOA/Jockey	486952	207565864	8,30
RTE/Bov-B	26139	17346465	0,69
L1/CIN4	24902	10906299	0,44
LTR elements	150947	71249124	2,85
BEL/Pao	758	480521	0,02
Gypsy/DIRS1	63674	46459553	1,86
Retroviral	6315	1712719	0,07
DNA transposons	1526496	479546375	19,19
hobo-Activator	135832	44458628	1,78
Tc1-IS630-Pogo	805085	289286954	11,57
PiggyBac	26804	7201568	0,29
Tourist/Harbinger	1652	854835	0,03
Unclassified	570559	133701458	5,35

Table S3: Sequencing and base calling details.

Sample ID	Kit	Flow cell	MinKnow Core (first run)	MinKnow Core (last run)	Guppy (first run)	Guppy (last run)
TORN	LSK-109	FLO-PRO002	18.08.2	18.08.2	2.2.3	2.2.3
TANA	LSK-109	FLO-PRO002	19.06.9-2	19.06.9-2	3.0.5	3.0.5

GLOP	LSK-109	FLO-PRO002	3.6.8	3.6.8	3.2.10	3.2.10
ARUN	LSK-109	FLO-PRO002	3.6.8	4.0.5	3.2.10	4.0.11
ALTA	LSK-109	FLO-PRO002	4.0.5	4.0.5	4.0.11	4.0.11
FROM	LSK-109	FLO-PRO002	4.0.5	4.0.5	4.0.11	4.0.11

ISBN: 978-82-575-2025-0

ISSN: 1894-6402



Norwegian University
of Life Sciences

Postboks 5003
NO-1432 Ås, Norway
+47 67 23 00 00
www.nmbu.no