The Genome and Linkage Map of the Northern Pike (*Esox lucius*): Conserved Synteny Revealed between the Salmonid Sister Group and the Neoteleostei

Eric B. Rondeau[1], David R. Minkley[1], Jong S. Leong[1], Amber M. Messmer[1], Johanna R. Jantzen[1], Kristian R. von Schalburg[1], Craig Lemon[2], Nathan H. Bird[1], Ben F. Koop[1]*

[1] Department of Biology, Centre for Biomedical Research, University of Victoria, Victoria, British Columbia, Canada, [2] The Charles O. Hayford Hackettstown State Fish Hatchery, Hackettstown, New Jersey, United States of America

Abstract

The northern pike is the most frequently studied member of the Esociformes, the closest order to the diverse and economically important Salmoniformes. The ancestor of all salmonids purportedly experienced a whole-genome duplication (WGD) event, making salmonid species ideal for studying the early impacts of genome duplication while complicating their use in wider analyses of teleost evolution. Studies suggest that the Esociformes diverged from the salmonid lineage prior to the WGD, supporting the use of northern pike as a pre-duplication outgroup. Here we present the first genome assembly, reference transcriptome and linkage map for northern pike, and evaluate the suitability of this species to provide a representative pre-duplication genome for future studies of salmonid and teleost evolution. The northern pike genome sequence is composed of 94,267 contigs (N50 = 16,909 bp) contained in 5,688 scaffolds (N50 = 700,535 bp); the total scaffolded genome size is 878 million bases. Multiple lines of evidence suggest that over 96% of the protein-coding genome is present in the genome assembly. The reference transcriptome was constructed from 13 tissues and contains 38,696 transcripts, which are accompanied by normalized expression data in all tissues. Gene-prediction analysis produced a total of 19,601 northern pike-specific gene models. The first-generation linkage map identifies 25 linkage groups, in agreement with northern pike's diploid karyotype of 2N = 50, and facilitates the placement of 46% of assembled bases onto linkage groups. Analyses reveal a high degree of conserved synteny between northern pike and other model teleost genomes. While conservation of gene order is limited to smaller syntenic blocks, the wider conservation of genome organization implies the northern pike exhibits a suitable approximation of a non-duplicated Protacanthopterygiian genome. This dataset will facilitate future studies of esocid biology and empower ongoing examinations of the Atlantic salmon and rainbow trout genomes by facilitating their comparison with other major teleost groups.

Citation: Rondeau EB, Minkley DR, Leong JS, Messmer AM, Jantzen JR, et al. (2014) The Genome and Linkage Map of the Northern Pike (*Esox lucius*): Conserved Synteny Revealed between the Salmonid Sister Group and the Neoteleostei. PLoS ONE 9(7): e102089. doi:10.1371/journal.pone.0102089

Editor: Tongming Yin, Nanjing Forestry University, China

Received April 10, 2014; **Accepted** June 14, 2014; **Published** July 28, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All sequence assemblies are available on the website web.uic.ca/grasp/pike. Genome sequence assemblies have been submitted to NCBI, under project GenBank:AZJR00000000, while the Trinity RNA-seq assembly is under project GenBank:GATF00000000. Any additional relevant data is provided within the paper and its Supporting files.

Funding: This study was primarily supported by a grant from the Natural Sciences and Engineering Research Council of Canada (www.nserc-crsng.gc.ca) to BFK; additional resources were provided by the University of Victoria (www.uvic.ca). Support was provided by NSERC through a postgraduate scholarship to EBR and by the government of British Columbia to support DRM. Compute Canada (www.computecanada.ca) contributed computing resources. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: bkoop@uvic.ca

Background

Of the nine currently available teleost genome assemblies (www.ensembl.org), all fall into either the Neoteleostei or the Ostariophysi [1]. This is not necessarily surprising as combined these two clades comprise more than 70% of the approximately 26,840 extant teleost species [2]. While not as large, other groups contain a number of commercially important species and thus receive a significant amount of scientific attention. This observation is particularly true of the Protacanthopterygii (the salmonids, esocids and marine smelts) [2]. The Protacanthopterygii are placed in an interesting evolutionary position, having diverged from the Neoteleostei approximately 200 million years ago (MYA) [1] following an even more ancestral split from the lineage containing the Ostariophysi (250 MYA; see Figure 1). Thus, the generation of a genome sequence representing the Protacanthopterygii provides an intermediate phylogenetic branch of significant scientific interest.

While not conclusively resolved [1,3–6] the Protacanthopterygii generally include the orders Salmoniformes, Esociformes (pikes and mudminnows) and Argentiniformes (marine smelts). The Salmoniformes (grayling, ciscoes, whitefish, trout, char, and salmon) are the most studied of the Protacanthopterygii and nearly all constituent species have significant economic, ecological, conservation, environmental and societal value [7,8]. In the last 60 years, over 70,000 publications [7] have examined fundamental

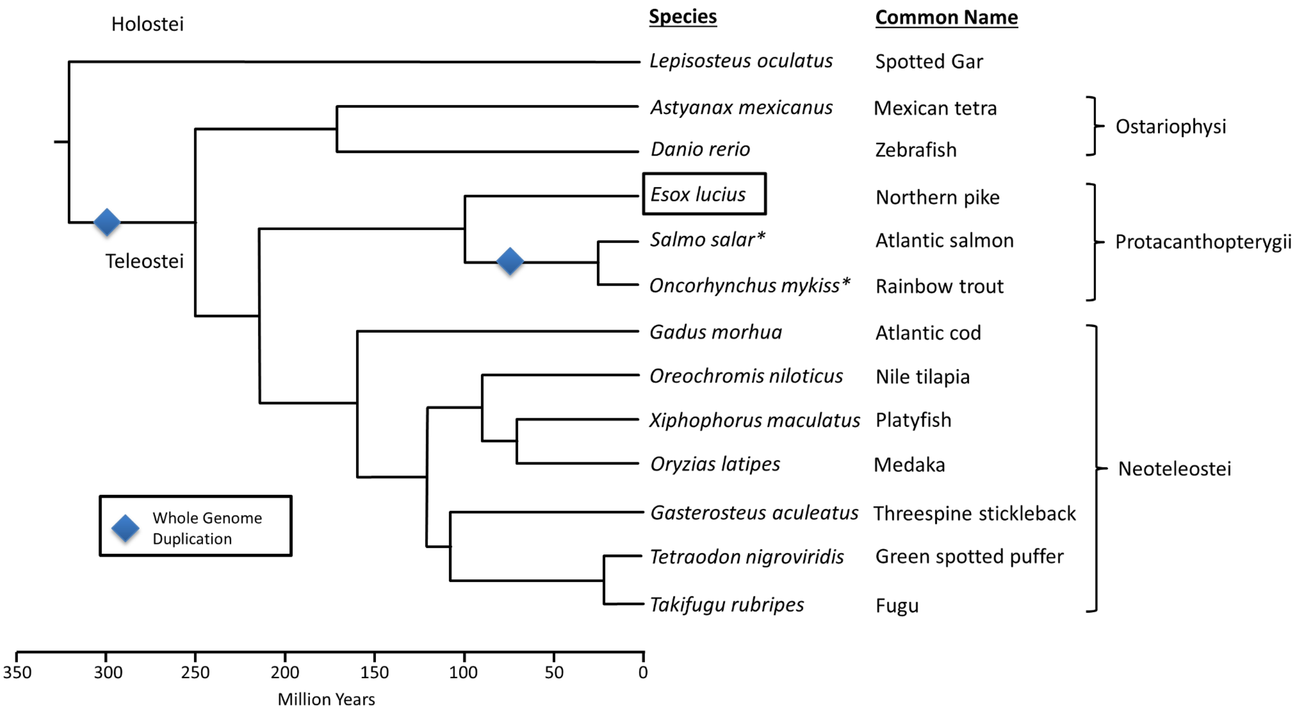


Figure 1. The phylogenetic relationship of teleosts with publicly available genome assemblies, as well as select species with ongoing genome projects (*). Divergence times and Subcohort naming based on Betancur et al. [1]. Diamond shapes represents presumed whole-genome duplications. *Lepisosteus oculatus* (Spotted gar) used as an outgroup.
doi:10.1371/journal.pone.0102089.g001

and applied questions relating to their fisheries, aquaculture, physiology, ecology, evolution, disease, resistance, reproduction, growth, tolerance to physical factors, and general immunity.

The Salmoniformes are a rapidly evolving order, with most species exhibiting in excess of 92% similarity in the coding DNA [9,10]. The biological diversity and complexity observed within salmonids is undoubtedly a reflection of their complex genomes. While two rounds of whole-genome duplication (WGD) are thought to have preceded the origin of vertebrates and a third round (3R) is believed to have occurred in basal teleosts [11–15], the common ancestor of salmonids has purportedly experienced an additional, fourth round (4R) WGD approximately 60–100 MYA [7,8,16–18]. Modern salmonids are considered pseudotetraploid and various independent rediploidizations and genome restabilizations have occurred in different salmonid lineages as they revert to a stable diploid state [7,8,16,17]. Although this recent duplication makes the salmonids difficult to use in the analysis of major teleost lineage evolution, it does make them ideal organisms for examining some of the early impacts of gene and genome duplications. These processes are thought to have played pivotal roles in promoting increased genetic diversity, the development of novel functions and functional specialization. The way in which a genome is reorganized to cope with duplicated chromosomes and the importance of gene duplications for evolution and adaptation represent long-standing biological questions that remain unresolved [7,19–22]. Additionally, the role of repetitive DNA in facilitating post-duplication genomic reorganization is also of increasing interest in light of the recent expansions of transposable elements which have been identified within salmonids [23]. The salmonid-specific WGD, the resulting pseudotetraploid genomic state, and recent waves of transposable element activity all contribute to the increased complexity of relating genetic mechanisms to physical characteristics in salmonid species. The presence of duplicated genes has made it more difficult to clearly identify genetic mechanisms underlying physiological, immunological, ecological and evolutionary traits within the commercially and culturally valuable salmonids. Presently, the International Collaboration to Sequence the Atlantic Salmon Genome (ICSASG) represents researchers, funding agencies and industry from Canada, Chile and Norway. It has the goal of generating a robust salmonid reference genome sequence that facilitates the study of all salmonids [7]. This work in Atlantic salmon (*Salmo salar*) [7] is ongoing. During review for this work, the genome assembly for the rainbow trout (*Oncorhynchus mykiss*) was released [24] providing an additional dataset with which to analyse the results of genome duplication in the salmonids.

The genetic diversity in the Protacanthopterygii is reflected in their variable genome sizes and highly variable karyotypes that frequently differ from the presumed ancestral karyotype of 48 or 50 chromosomes [25]. This primitive karyotype is seen in a few members, including within two genera of the Esociformes. Total DNA content comparisons [26] and several gene studies [9,10] suggest that Esociformes diverged from Salmoniformes prior to the salmonid WGD and that a genome of the Esociformes may closely approximate the ancestral genomic state. Therefore, it is important to determine how closely the organization of a representative Esociformes genome may reflect that of the ancestral Protacanthopterygiian. This information will help determine if comparisons to the Esociformes are appropriate in helping to resolve patterns of genome recombination and changes in gene expression that are implicated in the rediploidization of the salmonids.

Esociformes, an old lineage with fossils dating back to 70 MYA, contains a limited number of species (~13) of which most literature concerns the northern pike (*Esox lucius*). The karyotype for the *Esox* genus is 2N = 50 (NF = 50) [27,28], similar to the presumed karyotype of the teleost ancestor (2N = 48–50) [25]. This compares to the larger Salmoniformes karyotypes, which range from 2N = 52 (*Oncorhynchus gorbuscha*) to 2N = 102 (*Thymallus thymallus*) [26]. Esociformes outside the *Esox* genus have varying karyotypes of 2N = 22–78, with only *Novumbra hubbsi* exhibiting an ancestral-like diploid chromosome number of 2N = 48 (although the fundamental number indicates there are 62 chromosomal arms, suggesting significant rearrangement may have occurred [29]). A previous assessment of the suitability of northern pike as an outgroup for the study of duplicated genes in the salmonids examined 408 genes in Atlantic salmon (*Salmo salar*) and *E. Lucius* [10] and obtained results consistent with northern pike having a pre-WGD genome and being much more closely related to salmon than to other teleost species (zebrafish, medaka, fugu or stickleback). Such data suggest that the esocids provide an important intermediate comparison between salmonids and other more distantly-related fish genome sequences. For the present study *E. lucius* was chosen to represent Esociformes because of its availability, general interest and literature support.

The northern pike is a northern hemisphere, circumpolar, freshwater species that is typically the top predator in many lakes [30,31]. It is an important northern recreational fishery and forms the basis of small commercial fisheries in some locations [32]. Northern pike are morphologically diverse and highly adaptable, they play a major role in structuring freshwater communities and are used in stocking programs to improve water quality [32]. Despite their wide distribution, northern pike populations have very low levels of polymorphism and divergence [31], indicative of a recent common ancestor. Genetic studies suggest that 1–3 post-glacial refugia in North America and several significant refugia in Europe [30,33] contributed to the reduced levels of genetic diversity within this group.

In order to complement ongoing genome studies of Atlantic salmon and rainbow trout we sought to develop an extensive set of genomic resources for northern pike. To this end we have obtained a comprehensive set of northern pike transcripts and characterized their expression levels in various tissues. Further, we have produced a genome assembly that establishes the structure of genes, gene order, and repeat element content in northern pike. We have constructed a first-generation linkage map using highly variable microsatellite markers and mapped the relative position of both genomic scaffolds and expressed transcripts to linkage groups. Finally, we have examined the syntenic relationships between this representative esocid and distantly-related model organisms within the Neoteleostei and Ostariophysi in order to determine whether or not northern pike represent an appropriate model of a pre-duplication genome for use in the analysis of the salmonids. This work presents a significant resource for future studies of northern pike and provides the basis for invaluable comparisons in future salmonid studies, particularly once salmonid genomes become available.

Results and Discussion

A high quality genome assembly for *Esox lucius*

Libraries from 180 bp DNA fragments were Illumina paired-end sequenced (465.9 million 100 bp overlapping reads, ~50X genome coverage) (Michael Smith Genome Sciences Centre, Vancouver, BC, CANADA). Libraries from 2 kb (448 million reads, ~45X coverage) and 5 kb fragments (59.7 million reads, ~6X coverage) were mate-pair sequenced (BGI). Assembly using ALLPATHS-LG [34] produced 94,275 contigs with 823,910,316 bp. Subsequent filtering (vector, duplicates, mitochondrial sequences) produced a finalized dataset comprising 823,673,596 bp in 94,267 contigs (mean length = 8,738, median length = 4,729, contig N50 = 16,909). Contigs were further joined by ALLPATHS-LG to include the estimated gaps between contigs based on mate-pair data, which produced 5,688 scaffolds with a total scaffolded genome size of 877,777,613 bp. The longest scaffold was 5.1 Mb, and the overall scaffold assembly had an N50 of 700,535 bp. In all, 199 scaffolds were over 1 Mb in length and 1,600 scaffolds were over 100,000 bp. These sequences have been uploaded to NCBI under bioproject PRJNA221548, accession GenBank:AZJR00000000, as well as to web.uvic.ca/grasp/pike in FASTA format. As the genome assembly improves, updates to the assembly will continue to be provided on the website. Major genome assembly statistics are summarized in Table 1.

The most recent figures based on the c-value [35–37] estimate the *Esox lucius* haploid genome size to be approximately 1.09–1.12 Gb; if correct, our scaffold assembly covers approximately 78–79% of the genome. As there are known or suspected biases in next-generation sequencing (NGS) library construction (see [38] for example), it is possible that a small portion of the genome is missing from the sequencing libraries constructed. Published genome size estimates, however, range from 0.83–1.37 Gb [26], implying scaffold coverage could range anywhere from 64–100% of the genome. Therefore, because of the wide range of genome size estimates, a further estimate of genome completeness based on published EST transcript coverage [10] was performed. Of 11,662 EST contig sequences considered, 11,263 (96.6%) were successfully mapped to the scaffold assembly. Further, ALLPATHS-LG reported an estimated genome size of 911,413,085 bp based on the kmer distribution of raw genomic reads (k = 25), or 96.3% coverage. Together, these results suggest that the assembly represents a very large percentage of the genome.

While we cautiously expect that some chimeric scaffolds occur within our assembly, none were observed in our analysis of the largest 50 scaffolds (see synteny section below). Therefore, this initial version of the *E. lucius* genomic sequence provides an excellent framework to assess gene structure, gene content and general genome organization. The assembly statistics for both contigs and scaffolds are in-line with recently published fish genomes sequenced using NGS technologies, such as Atlantic cod (*Gadus morhua*; Contig N50 = 2,311 bp, Scaffold N50 = 393,166 bp) [39] and African coelacanth (*Latimeria chalumnae*; Contig N50 = 12,671 bp, Scaffold N50 = 924,513 bp) [40]. This assembly is the first genome assembly for a non-salmonid member of the Protacanthopterygii and represents an important evolutionary bridge between the two largest teleost subcohorts, Neoteleostei and Ostariophysi.

The genome sequence was examined for low-complexity repeat and transposable element content. This analysis allowed for the identification of major transposable element families present in the genome, and more importantly, provides a consensus library of transposable element sequences for use in filtering non-genic RNA sequences. An estimated 18.1% of the genome sequence is derived from transposable elements. This repetitive fraction is dominated by class II elements of the Tc1-Mariner superfamily (at least 9.42% of the genome is derived from Tc1-Mariner elements). The genomic abundance of observed transposable element taxa are reported in Table S1. An additional 2.0% of the genome consists of low-complexity sequences for a total of 20.1% of the genome annotated as repetitive sequence. Additional studies are currently underway to more thoroughly profile the transposable element sequences in *E. lucius* and to compare them to those of the Atlantic salmon. As a relative observation, 20% repeat-derived genome content is within the range represented in other teleost

Table 1. *Esox lucius* genome assembly statistics. Minimum contig and scaffold size ≥200 bp.

	Number (≥200 bp)	Bases (Mb)	N50 (Sequence length)	N50 (number of sequences)	Maximum Length (bp)
Contigs	94,267	824	16,909	13,483	232,364
Scaffolds	5,688	878	700,535	318	5,140,982

doi:10.1371/journal.pone.0102089.t001

genomes; while repeat elements represent ≥50% of the well-characterized zebrafish genome [41], only 25% of the stickleback [42] and Atlantic cod genomes [39], 17.5% of the medaka genome [43], and <5% fugu genome [44] are TE-derived.

Transcriptome assembly and Gene Identification

A total of 677,321,182 100 bp reads were obtained from 13 different tissues of a one year old individual and assembled into 413,679 contigs using the Trinity assembler [45]. Potential gene candidates were identified from the Trinity raw assembly. Those that were potential full-length annotated genes, had homology to a UniprotKB/Swiss-Prot or Gene Ontology protein, or those that possessed an unannotated ORF ≥300 bp were retained. All potential gene candidates were further filtered for the longest unique, non-overlapping location in the genome sequence. To reduce isoforms, alleles, recent duplicates, and possible sequencing errors, the potential gene candidates were screened for redundancy (≥98%, ≥300 bp). 38,696 potential gene candidates are represented in this curated non-redundant set, comprised of full-length annotations (9,553), UniprotKB/Swiss-Prot homology (14,538), Gene Ontology homology (604), and unannotated ORFs ≥300 bp (14,001). These gene candidates represent our *E. lucius* reference transcriptome set. These sequences have been submitted to NCBI under project accession GenBank: GATF00000000; the assembly described in this work is the first version, GenBank: GATF01000000.

An additional dataset incorporating all available transcriptome (including ESTs) and genome sequence data was prepared using the gene-modelling program MAKER2 to produce an *ab initio* dataset of 19,601 putative transcripts. It can be noted that the number of transcripts predicted by modelling is less than those from the de novo assembly. In some instances, this is a result of predicting a single gene model that spans multiple, non-overlapping, RNAseq contigs. In other cases, it appears to be due to limits in the data used to predict the models such as the size of the genomic contigs used; the short sequences limit the ability to successfully model a transcript. As the genome assembly is improved, so should the gene-modelling results. Datasets have thus been made available for download at web.uvic.ca/grasp/pike, and any future updates will be distributed here as well; the current MAKER2 dataset is additionally presented as File S1.

With previous EST studies and the current RNAseq data, we attempt to provide a comprehensive view of the number and identity of genes that make up the genome of *Esox lucius*. In combination with transcriptome data from other fish species, these gene lists will help facilitate future estimates of the identity and number of genes in the ancestral Protacanthopterygii lineage, before the whole-genome duplication occurred in the salmonid ancestor.

A snapshot of tissue-specific gene expression

To determine the expression levels of hypothesized genes in each of the 13 tissues examined, reads from each tissue library were mapped back to the reference transcriptome and normalized expression values (fragments per kilobase of transcript per million mapped reads = FPKM) were determined for each sequence. All FPKM values from the non-redundant protein-coding transcriptome are presented in Table S2, while results based on the MAKER2 transcriptome can be found in Table S3. The primary intent of this initial characterization of gene expression is to provide a basic resource that can empower future work in *E. lucius* while also providing a baseline comparison for research in salmonids that seeks to understand the impact of gene duplication on transcriptional regulation and gene network organization.

To illustrate the general utility of gene expression values, the genetic complexity of the different tissues was investigated by analysing the magnitude and specialization of expression in each tissue. We discuss results based on the Trinity-assembled transcriptome, although findings using the MAKER2 transcriptome generally reflect the same trends. Unsurprisingly, the most common highly-expressed genes are NADH-ubiquinone oxidoreductase, actin, cytochrome c and the ribosomal protein-coding genes (Table S2); these occur within the top 10 highest-expressed genes in almost all tissues. We also looked at specialized expression patterns for each gene by identifying any tissues in which FPKM values were more than three standard deviations greater than the average across all tissues. The number of specialized transcripts was counted for each tissue (Table S4) and the order of tissues by decreasing complexity is: brain (5,408 transcripts), testis (2,784 transcripts), spleen (1,100 transcripts), eye (1,055 transcripts), nose (804 transcripts), gut (659 transcripts), heart (630 transcripts), muscle (430 transcripts), kidney (406 transcripts), gill (374 transcripts), head kidney (338 transcripts), liver (332 transcripts) and stomach (140 transcripts). The brain represents the most complex/specialized tissue with the largest number of genes expressed and the largest number of specialized genes. We found that the testis were the second-most complex and specialized tissue, a result that is in accord with the work of others [46]. The spleen, eye and nose regions have the next highest number of specialized genes expressed.

In examining the ten most specialized genes in each tissue (Table S4) it was found that chymosin, gastricsin, and pepsin were highly expressed in the stomach tissue, rhodopsin and gamma-crystallin in the eye, immunoglobulins in the head kidney and complement in the spleen, genes often associated with these tissues. The muscle was the least complex tissue in terms of the number of transcripts expressed, however the stomach tissue expressed the smallest number of specialized genes. Perhaps one surprise was the high level of ice-structuring protein expressed in the gut. Though antifreeze protein (AFP) gene expression is known to occur in many tissues, it is typically produced in the liver and transferred to the blood [47]. However, as the gut is a possible route whereby nucleating ice might enter the fish, the finding of high transcription of AFP in the gut is informative and may help to explain some of the protective physiology of survival in high northern climates.

Documenting the expression values for each of the transcripts from the different tissues provides a single snapshot of expression in one set of conditions and is therefore bound to vary in different circumstances. This snapshot does, however, provide a basal atlas of gene expression estimates that can be used by the fish community for many experimental and comparative purposes.

E. lucius homologues suggest a pre-duplication genome

To investigate patterns of gene duplication in *E. lucius*, we plotted the percent similarity of transcript homologues between and within northern pike, Atlantic salmon and threespine stickleback (*Gasterosteus aculeatus*) (Figure 2). To facilitate this analysis three transcript datasets were used: 1) 38,696 reference transcripts identified within our RNA-seq assembly; 2) previously published Atlantic salmon EST data [9,10], and; 3) *G. aculeatus* transcripts available from the UCSC genome browser. Transcript homologues were identified within and between datasets using reciprocal best BLASTN hits following a similar approach previously used to analyse Atlantic salmon homologues [9]. In this analysis percent similarity was used as a proxy for evolutionary time; lower percent similarity between two paralogues suggests more time has passed since their origin by duplication.

Two peaks of percent similarity are apparent in the northern pike paralogue comparisons (Figure 2A): one above 97% and one centred at 82%. The analogous plot of stickleback paralogue similarity exhibits a similar profile (with a nearly identical peak at 82% identity). If a similar rate of mutation is assumed within the northern pike and stickleback lineages, the overlapping peaks occurring between 77% and 89% suggest a common origin. These peaks may represent the remnants of the 3R basal teleost genome duplication, which would have occurred in a distant common ancestor of both lineages. Atlantic salmon paralogue comparisons, in contrast, show much higher percent similarities and exhibit a notable peak at 94% identity, a region in which very few northern pike or stickleback paralogues appear. This upward shift in the peak of percent similarity is most easily explained by the fourth-round (4R), salmonid-specific WGD. While Atlantic salmon does not exhibit a distinct peak in the region identified in the other two fish species (77–89% similarity), the right tail of its percent distribution is considerably drawn out. The notably extended tail may be a result of a) paralogues produced following the 4R WGD with a much increased rate of sequence divergence relative to the average, or b) remnants of the 3R duplication in which one copy of each paralogous gene pair produced following a salmon-specific 4R duplication has been lost. Together, these observations support the hypothesis that an Atlantic salmon ancestor experienced a comparatively recent whole-genome duplication event that was not experienced in the pike or stickleback lineages, and supports the use of the northern pike as a WGD outgroup.

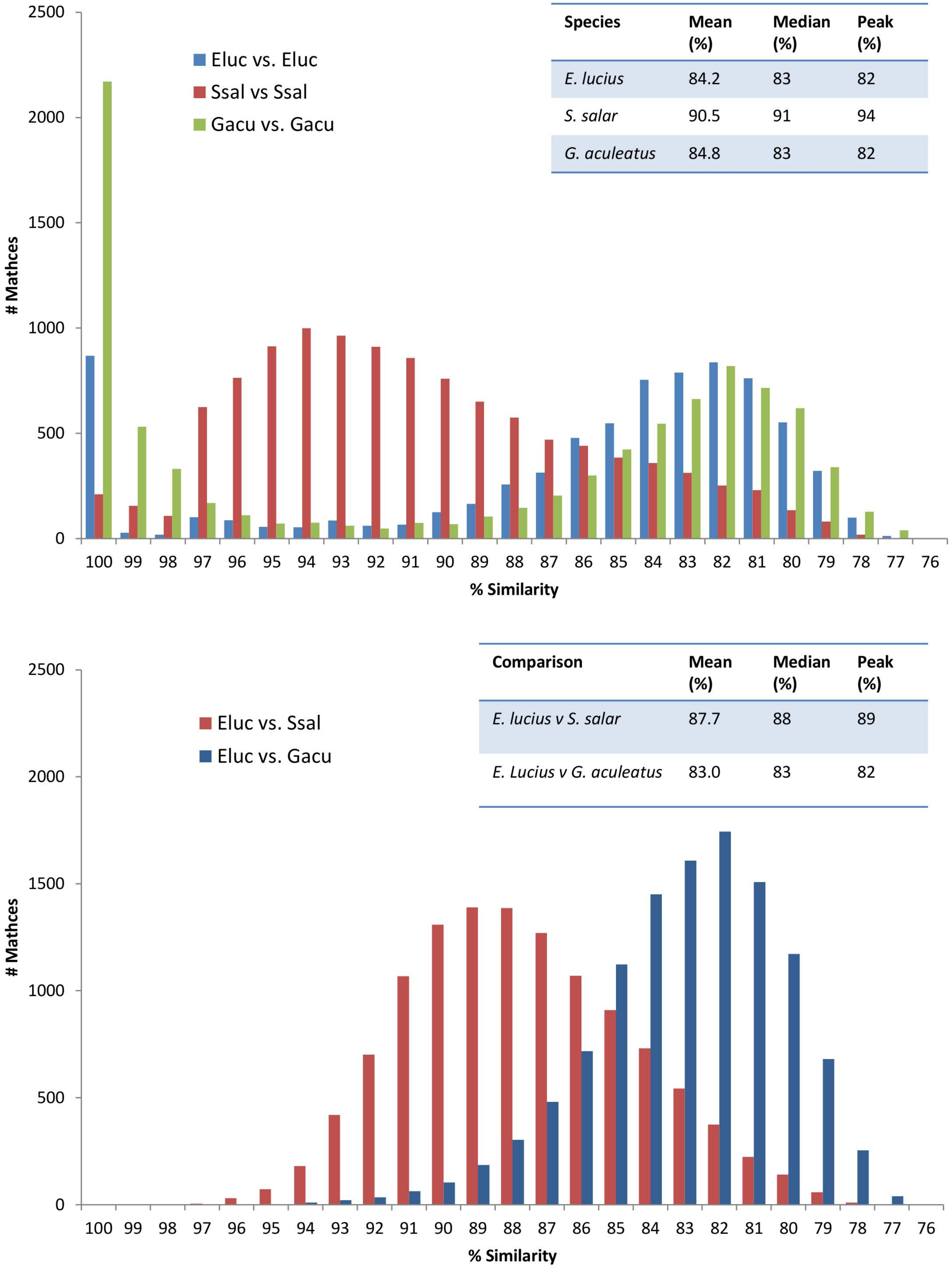
The peaks occurring above 97% similarity in all within-species comparisons imply very recent gene duplication events that are unique to each species. Although gene duplication is known to occur at very high rates, the vast majority of those duplicate pairs that are fixed in a lineage are thought to experience rapid loss of one or both copies, occurring within a few million years [48,49]. The high-similarity peaks observed in our within-species comparisons likely result from this dynamic of creation and loss; many paralogues are almost identical, indicating recent duplication, however fewer paralogues are observed as similarity decreases. These young paralogues may have been created by a recent series of segmental duplications [50], or may simply be an artefact of sequencing errors; we cannot with the current data distinguish between the two possibilities.

The most divergent peak observed in the northern pike within-species comparisons occurs from 77% to 89% similarity and contains 5,885 total transcripts. The annotations of these transcripts were examined in order to identify significantly over-represented functional categories (Table S5). Many of the terms identified fall into the development, signalling and regulatory classifications Brunet et al. [51] found to be enriched in *Tetrodon nigroviridis* gene duplicates likely to be from the 3R duplication. Of the Gene Ontology (GO) terms identified in the presumed 3R orthologues in northern pike however, the most striking are phosphate metabolism biological processes and nucleoside-binding and nucleotide-binding molecular functions. Interestingly, nucleotide-binding molecular function category was significantly under-represented in 3R-duplicated genes examined in the *Tetraodon* pufferfish. Examining more recent duplicates in carp, Wang et al. [52] identified immune-related terms and pathways as the predominantly enriched biological functions relative to in the recently duplicated common carp (*Cyprinus carpio*) genome, relative to the Zebrafish; immune-related terms were not a significant component of the GO terms identified in this study. The diversity of enriched GO categories implied by these studies suggests that different taxonomic groups may retain gene duplicate pairs from different functional categories, thus contributing to overall taxonomic diversity. Alternatively, the observed differences could indicate that different functional categories of gene duplicates are retained over different lengths of time.

Utilizing the same approach as that used for within-species comparisons, northern pike transcripts were compared with their orthologues in both the Atlantic salmon EST dataset and the threespine stickleback transcriptome (Figure 2B). Unsurprisingly, Atlantic salmon and northern pike orthologues are on average more similar (distribution maxima at 89%) than northern pike transcripts are to their respective orthologues in threespine stickleback (distribution maxima at 82%). These data are consistent with the expectation that the pike are more closely related to the salmonids than to species in the Neoteleostei, including stickleback. As northern pike and Atlantic salmon are estimated to have diverged from one another 100–130 million years ago (MYA) [1,4], the value of 89% provides a very broad reference point with which to estimate the timing of the orthologue divergence in the transcriptomes of the Salmoniformes and Esociformes.

RAD-tag based linkage mapping inefficient in pike

Our attempts to produce a RAD-tag based genetic map from a single full-sib family of *E. lucius* were unsuccessful. From 33,588 *Sbf*I-associated sequence loci across all family members, 351 informative polymorphic SNP markers were identified, or about 175 informative SNPs per sex. Further, the combination of very few useful markers and the bi-allelic nature of the SNPs did not allow for an efficient merge of the two sex-specific maps. Attempts to produce a useful linkage map from this data proved fruitless. Our initial decision to pursue this approach was based on its comparative success in other fish species. Other studies using the same protocol and enzyme produced 8,406 mapped SNPs in the spotted gar [53], 6,458 mapped SNPs in Atlantic salmon [54] and 5,703 mapped SNPs in Atlantic halibut [55]. The poor performance of this method in our study is almost certainly due to the very low degree of polymorphism observed in the northern pike data; the ALLPATHS-LG genome assembler reported an estimated polymorphism rate of 1 SNP for every 10,830 bp and an independent analysis of our RAD-tag dataset indicated an average polymorphism level of approximately 1 SNP for every 6,000 bp. The polymorphism rate estimates from our data are consistent with the very low levels of genetic polymorphism and heterozygosity noted in previous genetic studies of North American pike populations [33]. It is possible that future attempts utilizing a



Species	Mean (%)	Median (%)	Peak (%)
E. lucius	84.2	83	82
S. salar	90.5	91	94
G. aculeatus	84.8	83	82

Comparison	Mean (%)	Median (%)	Peak (%)
E. lucius v S. salar	87.7	88	89
E. Lucius v G. aculeatus	83.0	83	82

Figure 2. The percent similarities identified through transcriptome blastn comparisons. Blast results from non-redundant, repeat mask transcripts >300 bp with minimum 100 bp alignment. Results plotted by 1% intervals, grouped into bins. Results are graphed as A) within species (paralogues) and B) between species (Orthologues). Eluc = *Esox lucius*, Ssal = *Salmo salar*, Gacu = *Gasterosteus aculeatus*.
doi:10.1371/journal.pone.0102089.g002

restriction enzyme which cuts more frequently could restore the usefulness of this RAD-tag approach. Such changes, however, would significantly increase the number of sequencing lanes required, thereby reducing some of the cost-advantages of the RAD-tag technique.

A first-generation, Microsatellite based linkage map

In an attempt to increase the detection and integration levels of polymorphism, we utilized a half-sibling family (48 progeny from one father and 44 from a second) in combination with microsatellite markers. Utilizing 32,833 previously published northern pike EST sequences [9,10] we identified long near-perfect di- and tri-nucleotide repeats from which we generated 776 microsatellite primer pairs, of which 187 (24%) were found to be polymorphic in at least one of the parents. Of these polymorphic primer pairs, 180 were successfully mapped back to a single *E. lucius* genomic scaffold. To further extend our database of polymorphic primer pairs we similarly screened the largest 900 genomic scaffolds for long near-perfect di- and tri-nucleotide microsatellites. This approach allowed us to screen an additional 835 scaffold-designed primer pairs against the parents of our half-sib family. We identified 378 primer pairs that were polymorphic in at least one of the parents with between 1 and 6 markers identified per scaffold. From a total of 1,611 screened primer pairs we identified 565 polymorphic markers that cover 429 genomic scaffolds; 7 such markers could not be mapped to a scaffold.

The 565 polymorphic primer pairs we have identified represent a very significant increase in the available marker data for *E. lucius*; prior to this work only 70 markers were available [56–60]. Given the low level of polymorphism observed in northern pike, particularly in North American populations, the additional marker availability should be useful to future studies that seek to delineate population structure or trace unauthorized introductions of this fish species.

Analysis of the genotypes found within our half-sib family allowed 526 of the 565 polymorphic primer pairs (and 406 scaffolds) to be mapped across 25 linkage groups spanning 1289.3 cM, the expected number of linkage groups based on previous *E. lucius* karyotyping [27]. The final merged linkage map is presented in Figures 3–5. An additional 12 markers were polymorphic but remained unlinked at the LOD cutoff of 3.0. Of the 28 markers with more than 15% missing data, all but one were dropped from the linkage analysis: u0028c was maintained in the data set to join two groups of markers into the single linkage group LG-04. This join was further verified at LOD 3.0 using u0046a, u0028c and u0330a in an unrelated family of northern pike (Whiteshell hatchery, MB, Canada). While most of the mapping procedure was straightforward, some difficulty was observed in the merging of the map due to a low level of polymorphism across multiple individuals. Of the 526 mapped markers, 301 were only informative in a single parent, leaving relatively few common markers with which to merge the maps. Therefore, true marker order can only be determined from the individual maps in Figure S1; the merged map is presented as a best interpretation of the available data.

Sex-specific maps spanned 1245.4 cM in the female and 1166.0 cM in the male, an average recombination ratio of 1.07:1 between females and males. Regional recombination differed across and within particular linkage groups between the sexes, with one region in particular standing out. In the male maps, the top of LG14 shows strongly repressed recombination in the males relative to the females, with ~11X greater rate of recombination between scaffold 260 and scaffold 11 in the female map relative to the male. While the average recombination rates diverge slightly from even between the two sexes in pike, the recombination ratio is not nearly as skewed as in other teleost species; in Atlantic salmon, recently published ratios ranging from 1.3–1.7:1 (female:male) [54,61], in rainbow trout (*Oncorhynchus mykiss*) 1.68:1 [62], and in zebrafish (*Danio rerio*) 2.74:1 [63].

In one of the fathers we noted the unusual observation that of the 29 markers mapped to LG-19, none were found to be polymorphic (individual male-1 in Figure S1). In early karyotyping of northern pike various chromosome counts were observed [27], particularly in the range of 48–50 chromosomes; in concert with our data these observations suggest that the father was possibly missing a single chromosome, or distributed only a single chromosome to its progeny. It is also possible, depending on the degree of inbreeding within our sampled population, that this particular father inherited a separate but nearly identical chromosome from each of its parents; no observable polymorphism would be displayed in this situation. We are unable to determine the biological source of this unusual observation, as the available tissues are insufficient for karyotyping.

The low level of polymorphism implied by our data produces an interesting trade-off: genome assembly proved relatively straightforward due to a lack of sequence ambiguity introduced by polymorphic sites, while this same lack of polymorphic diversity ultimately limited the number of scaffolds that could be placed and oriented on the linkage map. Originally, we identified and targeted at least two microsatellite markers on each of the longest genome scaffolds; however, very few scaffolds ultimately possessed more than one such marker that was usefully polymorphic. Only 41 marker pairs derived from the same scaffold were mapped in such a way as to imply definitive scaffold orientation in at least one of the sex-specific linkage maps. Generally, such marker pairs were from the largest scaffolds and almost all were found on scaffolds greater than 1 Mb in length. The 406 scaffolds mapped by at least one microsatellite marker to this first-generation linkage map contribute to a total of 46% of scaffold assembly bases being associated with a linkage group. Additional northern pike families, further polymorphic marker identification, directed sequencing of scaffold gaps, and/or additional whole-genome sequencing will be required in order to improve this value in future work.

Synteny analyses suggest Neoteleostei-like genome in northern pike

Given that the majority of teleost species fall into either the Ostariophysi or the Neoteleostei subcohorts, it is understandable that all of the nine teleost genome sequences available in the Ensembl genome browser (December 2013) fall into one of these two clades. While the Ostariophysi are thought to have diverged from the Euteleosteomorpha approximately 250 MYA, the next branch point possessing a representative genome assembly occurs within the Neoteleostei, approximately 160 MYA when the Paracanthomorphacea (represented by the Atlantic cod) diverged from the Euacanthomorphacea [1]. Usefully, the Protacanthopterygii are thought to have branched off from other Euteleosteomorphs approximately 200 MYA, and as such they represent a

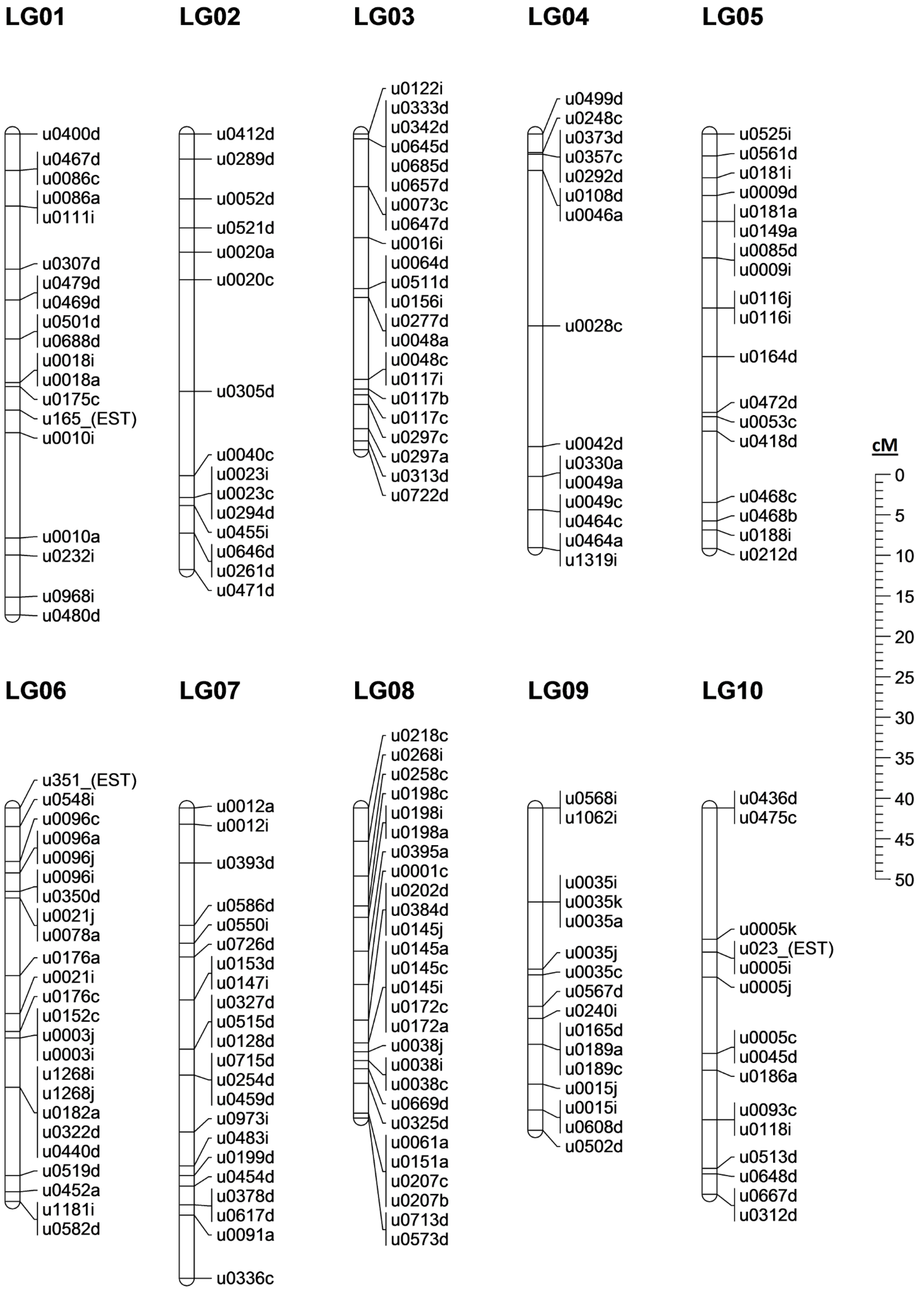


Figure 3. The northern pike merged linkage map: LG-01 to LG-10. Genomic scaffold identified by first 4 digit number, with multiple scaffold markers distinguished by final letter. Markers not mapped to a scaffold identified by "(EST)"; sex-specific linkage maps in Figure S1. Marker information in Table S7.
doi:10.1371/journal.pone.0102089.g003

phylogenetically intermediate branch point between the fish species with currently available genomic data. When considering the reconstruction of ancestral genomes for evolutionary studies, the addition of branch points and the decreasing of branch lengths can only improve the quality of the results. If a phylogenetically useful Protacanthoptyerygiian genome is desired, the selection of an order that did not experience the salmonid-specific genome duplication is desirable in order to limit complications; such an order could be reasonably expected to provide the most straightforward and ancestral-like genome that can be obtained

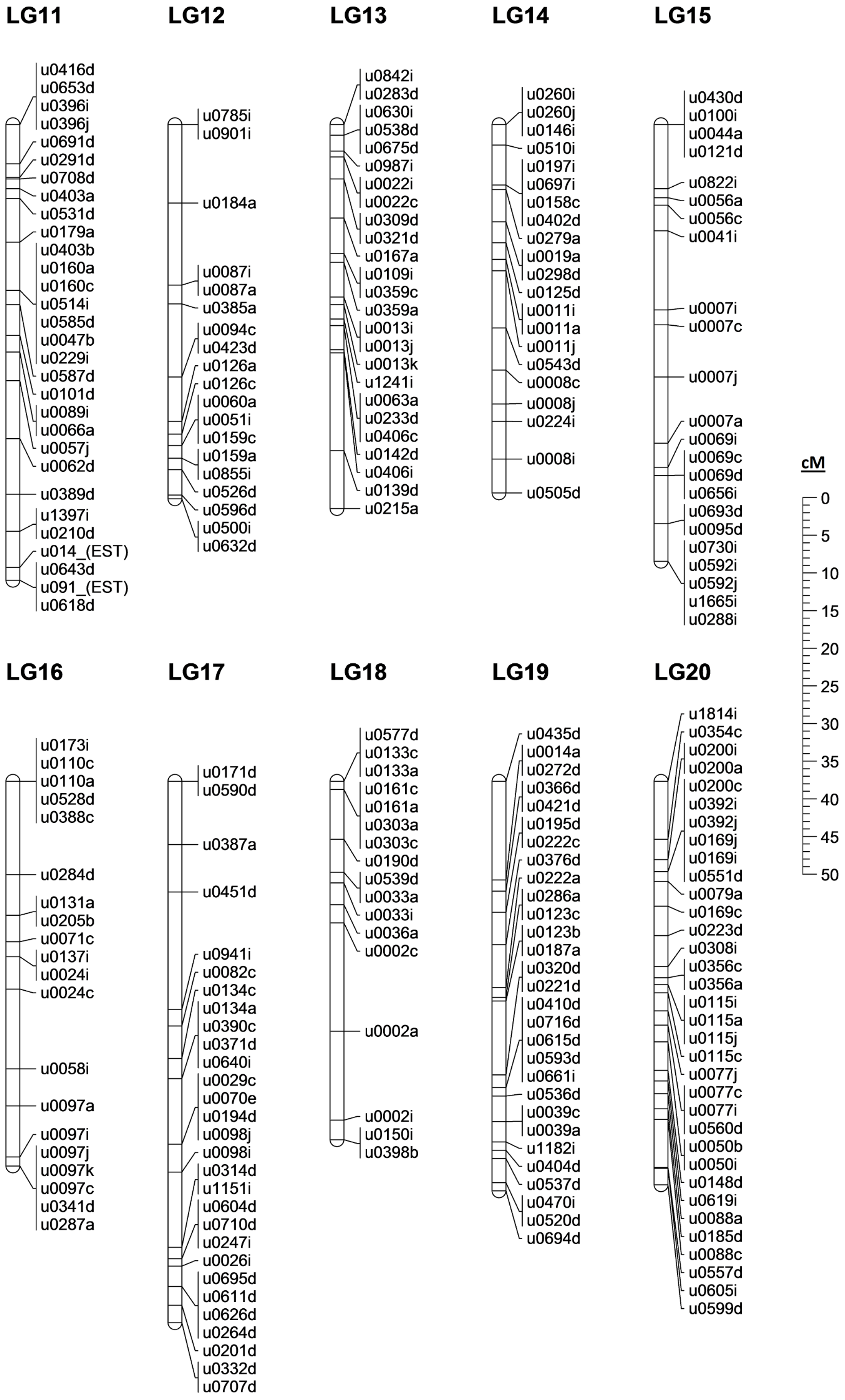

LG11
u0416d
u0653d
u0396i
u0396j
u0691d
u0291d
u0708d
u0403a
u0531d
u0179a
u0403b
u0160a
u0160c
u0514i
u0585d
u0047b
u0229i
u0587d
u0101d
u0089i
u0066a
u0057j
u0062d
u0389d
u1397i
u0210d
u014_(EST)
u0643d
u091_(EST)
u0618d
LG12
u0785i
u0901i
u0184a
u0087i
u0087a
u0385a
u0094c
u0423d
u0126a
u0126c
u0060a
u0051i
u0159c
u0159a
u0855i
u0526d
u0596d
u0500i
u0632d
LG13
u0842i
u0283d
u0630i
u0538d
u0675d
u0987i
u0022i
u0022c
u0309d
u0321d
u0167a
u0109i
u0359c
u0359a
u0013i
u0013j
u0013k
u1241i
u0063a
u0233d
u0406c
u0142d
u0406i
u0139d
u0215a
LG14
u0260i
u0260j
u0146i
u0510i
u0197i
u0697i
u0158c
u0402d
u0279a
u0019a
u0298d
u0125d
u0011i
u0011a
u0011j
u0543d
u0008c
u0008j
u0224i
u0008i
u0505d
LG15
u0430d
u0100i
u0044a
u0121d
u0822i
u0056a
u0056c
u0041i
u0007i
u0007c
u0007j
u0007a
u0069i
u0069c
u0069d
u0656i
u0693d
u0095d
u0730i
u0592i
u0592j
u1665i
u0288i
cM
0
5
10
15
20
25
30
35
40
45
50
LG16
u0173i
u0110c
u0110a
u0528d
u0388c
u0284d
u0131a
u0205b
u0071c
u0137i
u0024i
u0024c
u0058i
u0097a
u0097i
u0097j
u0097k
u0097c
u0341d
u0287a
LG17
u0171d
u0590d
u0387a
u0451d
u0941i
u0082c
u0134c
u0134a
u0390c
u0371d
u0640i
u0029c
u0070e
u0194d
u0098j
u0098i
u0314d
u1151i
u0604d
u0710d
u0247i
u0026i
u0695d
u0611d
u0626d
u0264d
u0201d
u0332d
u0707d
LG18
u0577d
u0133c
u0133a
u0161c
u0161a
u0303a
u0303c
u0190d
u0539d
u0033a
u0033i
u0036a
u0002c
u0002a
u0002i
u0150i
u0398b
LG19
u0435d
u0014a
u0272d
u0366d
u0421d
u0195d
u0222c
u0376d
u0222a
u0286a
u0123c
u0123b
u0187a
u0320d
u0221d
u0410d
u0716d
u0615d
u0593d
u0661i
u0536d
u0039c
u0039a
u1182i
u0404d
u0537d
u0470i
u0520d
u0694d
LG20
u1814i
u0354c
u0200i
u0200a
u0200c
u0392i
u0392j
u0169j
u0169i
u0551d
u0079a
u0169c
u0223d
u0308i
u0356c
u0356a
u0115i
u0115a
u0115j
u0115c
u0077j
u0077c
u0077i
u0560d
u0050b
u0050i
u0148d
u0619i
u0088a
u0185d
u0088c
u0557d
u0605i
u0599d

Figure 4. The northern pike merged linkage map: LG-11 to LG-20. Genomic scaffold identified by first 4 digit number, with multiple scaffold markers distinguished by final letter. Markers not mapped to a scaffold identified by "(EST)"; sex-specific linkage maps in Figure S1. Marker information in Table S7.
doi:10.1371/journal.pone.0102089.g004

from this group. The Esociformes, represented in the present study by northern pike, are such a suitably placed order; the northern pike genome will be an excellent intermediate resource for use in phylogenetic studies of teleosts.

Using the genetic map we set out to determine whether or not the gene distribution and genome structure of northern pike imply synteny with the hypothesized genome of the ancestral teleost, as is tentatively implied by karyotype morphology. To this end, the syntenic relationships were examined between four species representing three major subcohorts of the infraclass Teleostei: northern pike (*Esox lucius*) was chosen to represent the Protacanthopterygii, zebrafish (*Danio rerio*) was chosen to represent the Ostariophysi, and both threespine stickleback (*Gasterosteus aculeatus*) and medaka (*Oryzias latipes*) were selected to represent the Neoteleostei. First, transcripts from northern pike were mapped to those northern pike genomic scaffolds that had previously been assigned to a group on the linkage map. Putative orthologues for these mapped transcripts were then identified in the transcriptomes of the three other representative teleosts using a reciprocal BLAST approach. A linkage group was subsequently inferred for each mapped transcript based on that of its parent scaffold. Finally, this *E. lucius* transcript linkage group was compared to the chromosomal location of the transcript's orthologue in stickleback (Figure 6A), medaka (Figure 6B) and zebrafish (Figure 6C); see Table S6 for specific numbers.

Considering the time since their last common ancestor there exists a clear association between linkage groups in *E. lucius* and chromosomes in *G. aculeatus* and *O. latipes*. In general, each northern pike linkage group predominantly associated with a single stickleback chromosome. In four cases there were two *E. lucius* linkage groups that associated with only a single stickleback chromosome. Of these four cases, only one was analogously observed in *Oryzias latipes* (Figure 6B), a species that possesses a karyotype equivalent to the presumed ancestral karyotype in the Neoteleostei [25]. Specifically, only chr. 1 of *O. latipes* (orthologous to *G. aculeatus* chr. IX) shows a strong association with two separate *E. lucius* linkage groups. This observation suggests that the other three instances in which a single stickleback chromosome corresponds to two northern pike linkage groups (ie. *G. aculeatus* chrs. I, IV and VII) are best described as stickleback lineage-specific fusions of three pairs of ancestral-like chromosomes. This interpretation is further supported by recent work using sablefish (*Anoplopoma fimbria*), a species which is much more closely related to stickleback than is northern pike or medaka [64].

Two *E. lucius* linkage groups consistently associate with only a single chromosome of the other representative teleosts. As previously discussed, transcripts from northern pike linkage groups LG-09 and LG-25 predominantly possess orthologues on stickleback chr. IX and medaka chr. 1; an analogous situation is

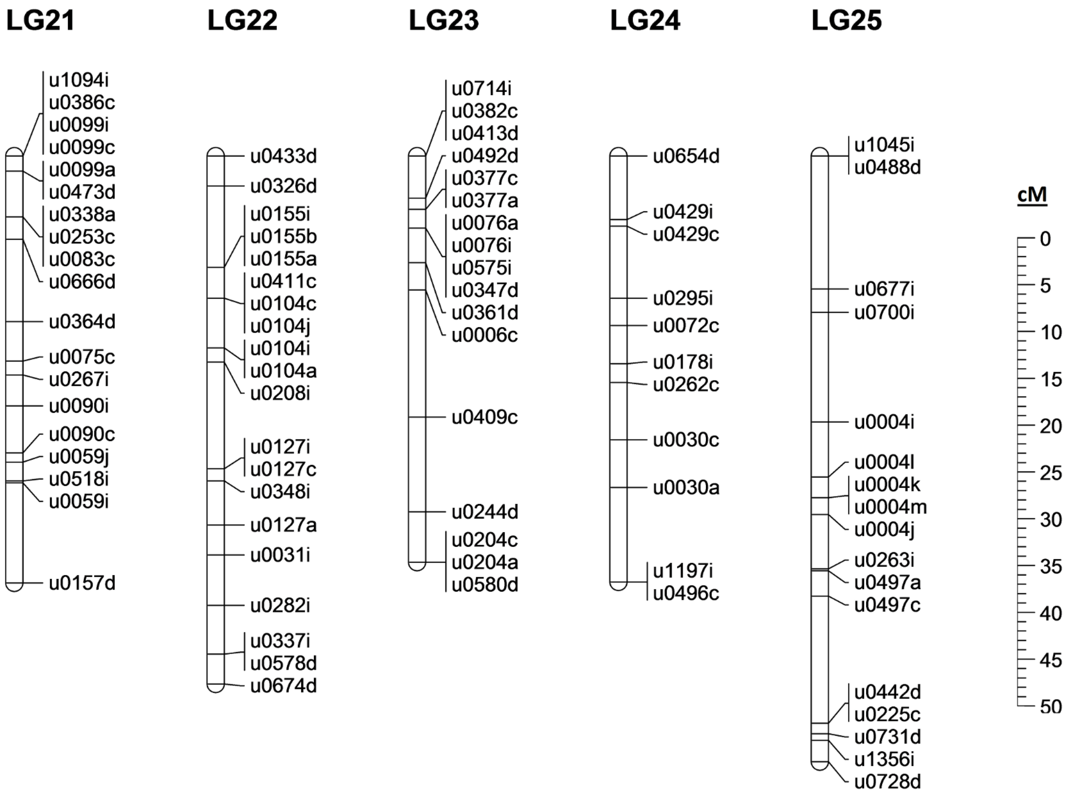


Figure 5. The northern pike merged linkage map: LG-21 to LG-25. Genomic scaffold identified by first 4 digit number, with multiple scaffold markers distinguished by final letter. Markers not mapped to a scaffold identified by "(EST)"; sex-specific linkage maps in Figure S1. Marker information in Table S7.
doi:10.1371/journal.pone.0102089.g005

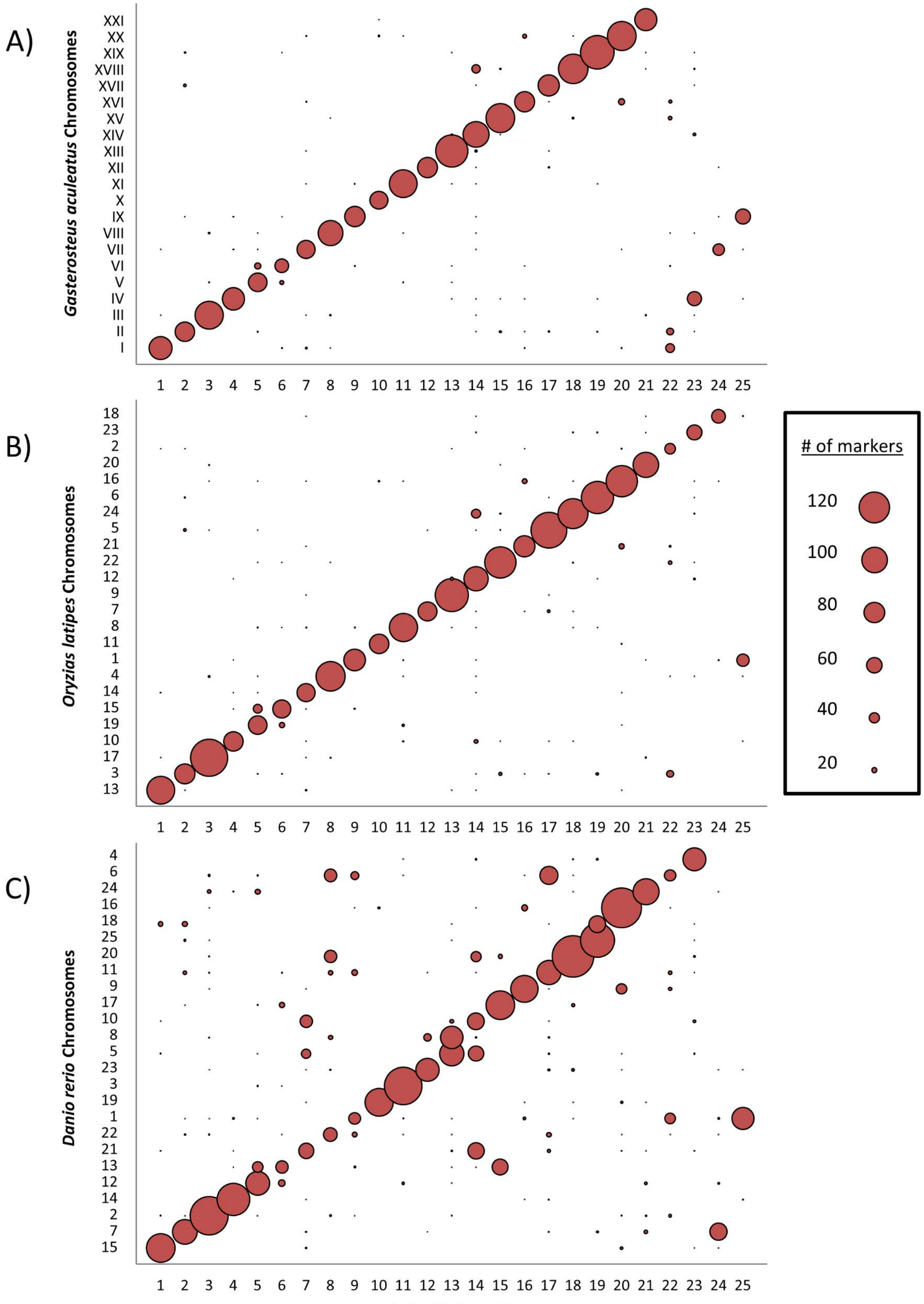


Figure 6. The synteny between the northern pike and model teleosts. Reciprocal best hit results (≥50% of total length, ≤1e-5) displaying synteny between the 25 pike linkage groups and three model teleost genomes, threespine stickleback (*Gasterosteus aculeatus*), medaka (*Oryzias latipes*), and zebrafish (*Danio rerio*). Sphere diameter reflects the number of matched transcripts. For numerical results, see Table S6.
doi:10.1371/journal.pone.0102089.g006

observed with zebrafish chr. 1. These observations suggest that the common ancestor of the Neoteleostei and the Protacanthopterygii most likely possessed a haploid karyotype of 24 chromosomes; this is in contrast to the haploid karyotype of 25 chromosomes observed in members of the *Esox* genus [27] and in the majority of the members of the Ostariophysi [25]. This analysis further implies that the two *E. lucius* chromosomes represented by LG-09 and LG-25 were formed as a result of the fission of a single ancestral chromosome, and that this fission occurred after the divergence of the Protacanthopterygii lineage. Syntenic analyses of orthologous sequence shared by stickleback and Atlantic salmon [61] suggest that the ancestral chromosome that eventually evolved into stickleback chromosome IX had already split in two prior to the salmonid-specific genome duplication. Therefore, this chromosomal fission likely occurred prior to the Esociformes/Salmoniformes divergence and it can be reasonably concluded that the common ancestor of the Esociformes and Salmoniformes possessed a 2N = 50 karyotype which was secondarily derived from the more ancestral karyotype of 2N = 48.

While the distribution of orthologous genes to chromosomes was particularly well conserved between the Protacanthopterygii and the Neoteleostei, the gene order along chromosomes was not as consistently similar. The positions of mapped genes in northern pike and the three model teleost genomes were plotted for the largest 50 scaffolds (Figure S2). In some instances, such as scaffolds 4 and 6, syntenic blocks in excess of three 3 million bp can be found where gene order appears to be conserved between medaka, stickleback and northern pike. In other cases, such as those observed in scaffolds 13 and 15, there are only small blocks of 500,000 bp or less where conservation of gene order can be observed. There are qualitatively many fewer occurrences of conserved gene order between zebrafish and pike, however there do exist large tracts of conserved order in certain scaffolds; one such scaffold is scaffold 1, where sections of 1–2 million bp maintain consistent ordering. In fact, gene order appears to be more consistently similar between zebrafish and pike than between pike and medaka in scaffold 1.

While the relatively small number of reciprocal blast hits per scaffold makes quantifying the degree of gene order conservation difficult, the number of consistently ordered transcripts from a single syntenic chromosome lends confidence to the quality of the assembled scaffolds. In three of the largest 50 scaffolds (scaffolds 8, 9 and 31), however, northern pike transcripts are consistently mapped across two or three chromosomes in all three model teleosts; this raises the possibility of chimeric scaffold assembly. An examination of the scaffolds mapped to either side of scaffold 9 in LG-05 implies that the discrepancy in this particular scaffold represents a genuine chromosomal translocation, in which the break point falls in the middle of the scaffold. Scaffolds to one side of LG-05 exhibit blast hits predominately to *G. aculeatus* chr. V, while scaffolds on the other show blast hits to *G. aculeatus* chr. VI. Similarly, scaffold 8 (mapped to LG-14) also appears to contain a genuine chromosomal translocation rather than a chimeric assembly artefact. This is inferred because a) the majority of other scaffolds on LG-14 possess best blast hits to *G. aculeatus* chr. XIV and b) scaffold 224, mapped adjacent to the chr. XVIII 'end' of scaffold 8 in LG-14, exhibits blast results dominated by hits to *G. aculeatus* chr. XVIII. Given the pronounced split between these regions it seems likely these translocations are quite recent and that the genes involved have not yet had time to be re-distributed to other regions across the chromosome. The third scaffold exhibiting potential assembly chimerism, scaffold 31 (located on LG-22), is not flanked by scaffolds that support a similarly distinct breakpoint. However, scaffolds mapped to LG-22 possess genes with significant numbers of orthologues on at least four different chromosomes in *O. latipes* and *G. aculeatus* (Figure 6A and B; Table S6); genes from different chromosomal sources are also mixed and distributed over the entire length of this linkage group. Together, these observations suggest that LG-22 possesses an old collection of pieces from multiple ancestral chromosomes that have had significant time to redistribute themselves throughout the linkage group. This characterization of LG-22 supports the idea that scaffold 8 reflects the biological reality and is not the result of inappropriately linked assembly contigs. While we are unable to rule out the possibility of chimeric joins having occurred in our genomic assembly, the three large scaffolds deemed most likely to have been chimeric show evidence suggesting that they are indeed real.

Although gene order is highly variable, chromosomal synteny remains high between the northern pike and the Neoteleostei representatives. While small sections appear to have undergone chromosomal translocations, and LG-22 may be highly reorganized, there is generally a 1:1 relationship between linkage groups identified in northern pike and the chromosomes in the Neoteleostei. This implies the genome of *E. lucius* is indeed ancestral-like and will provide a good overall representative of the pre-duplication Protacanthopterygii genome structure in wider studies of genome evolution.

The usefulness in salmonid duplication analysis

While inferring ancestral states is made easier with an ancestral-like genome, the examination of more recent events, such as the WGD event and subsequent rediploidization in the salmonid lineage, can be made much easier using this genetic dataset. Although numerous rearrangements have undoubtedly had a substantial effect on genome organization, many of the greatest genomic changes must be seen as the consequence of the formation and evolution of paralogous genes resulting from a WGD (such gene pairs are termed ohnologues [65]). As the genome attempts to revert to a stable diploid state, ohnologues are presumed to evolve like individual gene duplicates through subfunctionalization, neofunctionalization or loss of one of the gene copies. Determining the mode and rate of gene and genome evolution requires an understanding of where the two copies began. Gaining this understanding would be most easily accomplished using the ancestral, pre-duplicated sequence of the gene or genome segment; since this sequence is unavailable it is most readily inferred using an appropriate pre-duplication sister group. Even given the significant amount of time which has passed since their divergence from each other over 100 MYA [1] the Esociformes remain the most closely related order to the Salmoniformes and therefore represent the best system of species from which a pre-duplication genomic state can be inferred.

Nucleotide BLASTN alignments between the flanking sequence of 5,918 mapped genomic markers from the Atlantic salmon linkage map [61] and the northern pike genome scaffolds produced 3,241 high-quality alignments. Of these alignments, 1,803 involved scaffolds assigned a linkage group on the northern pike genetic map. An analysis of these alignments clearly indicates that each *E. lucius* linkage group possesses significant conserved synteny with two Atlantic salmon chromosomes (Table 2). This observation is consistent with a historical duplication of each chromosome in the salmonid lineage that did not occur in an esocid ancestor. These results corroborate previous analyses which compared Atlantic salmon to stickleback [61]. Indeed, the increased relatedness and shared Protacanthopterygiian-specific rearrangements between Atlantic salmon and northern pike

Table 2. Sequences from mapped Atlantic salmon SNPs [61] blasted against the *Esox lucius* genome (≤1e-10).

		Esox lucius **Linkage Groups**																								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**
Salmo salar **Chromosomes**	**Ssa01**					8	**28**								**44**	**41**										
	Ssa02									**45**							13				**40**					
	Ssa03			**37**								**43**														
	Ssa04							**28**																		**24**
	Ssa05				**28**												10				**47**					
	Ssa06											**52**							**43**							
	Ssa07																		3					**17**	**10**	
	Ssa08																									**15**
	Ssa09	**32**			**34**											**29**										
	Ssa10								**68**	3										**36**						
	Ssa11		**35**					10							**31**											
	Ssa12		3							**18**								**55**								
	Ssa13							**20**					**28**		11											
	Ssa14			**18**							**25**															
	Ssa15												**13**	5	8				**33**							
	Ssa16	3							22											**20**			**23**			
	Ssa17																						**22**	**23**		
	Ssa18					4	**29**																		**16**	
	Ssa19					**17**	3															**21**				
	Ssa20	**18**												**36**												
	Ssa21																**25**				6					
	Ssa22		4															**56**								
	Ssa23								**33**											12						
	Ssa24													**39**												
	Ssa25																**25**									
	Ssa26		**46**									3								3						
	Ssa27										**41**															
	Ssa28					**24**	8																			
	Ssa29																					**24**				
Total blast hits		**57**	**89**	**58**	**65**	**59**	**71**	**63**	**131**	**67**	**72**	**108**	**43**	**87**	**103**	**73**	**77**	**118**	**84**	**74**	**97**	**46**	**49**	**45**	**27**	**40**

The Atlantic salmon chromosome linked to each mapped SNP was plotted against the linkage group of the top *E. lucius* scaffold blast match. Results <3 for a given LG/chromosome comparison are removed for clarity.
doi:10.1371/journal.pone.0102089.t002

demonstrate an even greater example of shared synteny between these two species that are separated by a genome duplication.

This work in northern pike provides a core dataset with which to examine genome evolution following whole-genome duplication. The reference transcriptome allows for the functional analysis of ohnologue evolution while the genome assemblies facilitate the study of intron and promoter evolution as well as post-WGD gene and genome rearrangements. The role of transposable elements in both esocid and salmonid genome evolution and speciation can be further explored by inferring pre-duplication genomic repeat content, by dating periods of transposable element activity and by characterizing the abundance and character of transposable elements in this pre-duplication genome.

Conclusions

This work represents the largest currently available collection of genomic data for a member of the Esociformes and the first genome assembly for a non-salmonid member of the Protacanthopterygii, an economically important superorder which bridges the evolutionary gap between the Ostariophysi and Neoteleostei. We present a genome assembly of 824 Mbp across 94,267 contigs and 878 Mbp across 5,688 scaffolds, representing an estimated 96% of the genome; 50% of the assembly is represented in the largest 318 scaffolds. Approximately 20% of the genome sequence is made up of repeat elements. Two transcriptome assemblies from multi-tissue RNA-seq are presented and FPKM values provide a baseline dataset of expression levels across 13 individual tissues. A microsatellite-based linkage map places 526 markers across 25 linkage groups, the expected number of linkage groups based on previous karyotypes. These markers further facilitate the mapping of 46% of the scaffold assembly sequence onto linkage groups. Comparisons between linkage-mapped scaffolds and three teleost genomes show that northern pike has a chromosomal distribution of genes very similar to that of the Neoteleostei, suggesting relatively low chromosomal translocation activity and an ancestral-like gene distribution. Gene order along individual chromosomes, however, exhibits significant reorganization. This dataset represents an important resource for future ecological and evolutionary work in northern pike and will assist in understanding the effects of whole-genome duplication on genome evolution, especially once the upcoming salmonid genomes become available.

Methods and Materials

Animal Care

In accordance with the Canadian Council on Animal Care Guidelines, the Animal Care Committee at the University of Victoria did not require ethical review as only archival tissue or tissue from fish that were harvested by various government agencies for their own purposes, was used in the course of the experiments.

Whole-Genome Sequencing

For whole-genome sequencing, genomic DNA was extracted from the spleen of a single male pike individual from an introduced population to Charlie Lake, British Columbia, CANADA [66], using pooled DNA from both DNeasy (QIAGEN) and phenol [67] extractions. DNA libraries of 180 bp were prepared and sequenced (paired-end) by the Michael Smith Genome Sciences Centre; libraries of 2 kb and 5 kb fragments were constructed and sequenced by BGI (http://www.genomics.cn/en/index); all sequences were produced using Illumina HiSeq2000 instruments (San Diego, CA).

Following ALLPATHS-LG genome assembly, contigs were checked for long runs of ambiguous bases and trimmed/split if necessary. We removed contigs that were primarily vector sequence (BLASTN 1e-50). Furthermore, all contigs were trimmed internally and at terminal ends for any vector sequence according to NCBI's VecScreen search recommendations (www.ncbi.nlm.nih.gov/tools/vecscreen/). Duplicate contigs were removed and mitochondrial contigs were removed (BLASTN 1e-5). This resulting contig set was compared to GenBank's non-redundant nucleotide database for contamination (BLASTN 1e-25). Potential non-vertebrate sequence matches were flagged and manually inspected to determine classification. These contig sequences including scaffold information have been uploaded to NCBI under BioProject ID PRJNA221548, accession AZJR00000000.

Transposable element library creation and genomic repeat annotation

The process used to create the northern pike transposable element (TE) library was based on that used by Wegrzyn et al. [68] to identify TE sequences in the Loblolly Pine genome. First, a custom repeat 'seed' library (CSL) containing 33,304 TE sequences was assembled from three sources: 1) a quality-checked repeat consensus library produced by the *de novo* repeat-finding program REPET (version 1.3.9.1; default settings; [69]); 2) the RepBase database of TE sequences (REPET-formatted v18.08; [70]); and 3) TE sequences previously identified in Esociformes or Salmoniformes and deposited in Genbank.

Following CSL creation, the CENSOR program [71] was used to identify all non-redundant instances of CSL sequences in the northern pike genome contigs. Subsequent filtering removed all CENSOR hits that were smaller than 80% of the length of the respective CSL query sequence. Sequences from long CENSOR hits were clustered using the uclust.global program of USEARCH [72]. USEARCH parameters were such that a sequence was only added to an existing cluster if it possessed at least 80% nucleotide similarity over 80% of its length to the cluster's representative 'centroid'. Following clustering, a representative sequence was obtained for each cluster; if there were fewer than three sequences in the cluster the longest sequence was chosen as a representative, otherwise a consensus sequence was generated from a multiple sequence alignment (MSA). MSAs were built using the T-COFFEE program [73], which used alignment data from two sources: global alignment information from MUSCLE [74,75] and local alignment information from USEARCH.

The representative sequences from all clusters were combined and sequences with strong BLASTX alignments (E-value $\leq$1e-10) to the UniprotKB/SwissProt protein database (July 2013 release) were manually examined and, if appropriate, removed as non-TE host genes. The remaining library sequences were classified in accordance with the Wicker et al. classification system [76] based on the occurrence of structural motifs and/or similarity to previously identified TEs. The classification process relied on both manual and automated processing of output from NCBI's BLAST+ software and the PASTEClassifier.py script included with REPET. The final northern pike TE library contained 26,628 sequences of which 14,820 were classified to at least the subclass level under Wicker's taxonomic system. Many of the in-house scripts used to create and validate the TE library utilized modules included in the BioPython library [77,78].

The total amount of low-complexity DNA and TE-derived sequence within the genome was estimated using RepeatMasker (v4.0.3 [79]); RepeatMasker was configured to use RMBlast (version 2.2.23+) as its internal search engine. The abundance estimations for individual TE taxa were aggregated from RepeatMasker's annotation output file.

RNA-seq and assembly

All tissues for RNA-seq were extracted from a single, juvenile, male pike caught November 1, 2011 in the Portage Diversion/Assiniboine River Floodway near Portage la Prairie, MB, Canada (N50.00747 W98.38181) as the water was beginning to freeze. The animal was euthanized and then frozen on dry ice until sampling, and subsequently at −80°C. RNA from all 13 tissues – brain, eye, gill, hind gut, head kidney, heart, kidney, liver, muscle, nose, stomach, spleen and testis - were extracted by mixer-mill (Retsch) homogenization in Trizol (Invitrogen), followed by column clean-up using the RNeasy kit (QIAGEN). Total RNA was submitted directly to BGI for Illumina sequencing; indexed samples were pooled (4 per RNA-seq lane, pooled along with other projects) and sequenced on an Illumina HiSeq2000.

Paired-end raw reads were quality-trimmed using Trimmomatic [80]. Trimmed reads from all tissues were combined and a *de novo* transcriptome assembly was created using the Trinity assembler [45]. Trinity used a minimum kmer count of 3 in order to reduce noise introduced by sequencing errors.

In order to generate a non-redundant reference set of transcripts, the resulting set of Trinity transcripts were filtered. These raw putative transcripts were reduced by retaining those that were characterized as full-length [10], those that had a significant BLASTX [81] match to the UniProtKB/Swiss-Prot or Gene Ontology protein databases (≤1e-5) without being transposable element annotations, and those that did not show any sequence homology to a known protein but had a predicted open reading frame ≥300 bp. This reduced set was mapped to our genome assembly using BLAT [82]. Transcripts mapping uniquely to a single loci were retained. In cases where multiple transcripts (from alternatively-spliced mRNAs) were mapped to the same genomic location, only the longest transcript was retained. We masked repeats using RepeatMasker [79] and our *Esox lucius* repeat library, removing open reading frames that were no longer ≥300 bp. To remove possible alleles, recent duplicates, and sequencing errors in our de novo assembly, we took a single representative of transcripts that were ≥98% similar over a minimum length of 300 bp, as determined by BLASTN. This curated set represents our RNA-seq reference transcriptome.

We produced an *ab initio* transcript set from our genome contig assembly using MAKER2 [83]. An Augustus [84] gene model specific for *Esox lucius* was generated using contigs from the RNA-seq assembly and the previously published EST assembly as evidence. MAKER2 was run for two rounds to produce an *Esox lucius* SNAP model. A final round of MAKER2 was run using the Augustus gene model, SNAP model, and the mRNA evidence.

Homology analysis of protein-coding mRNA data utilized the non-redundant protein-coding northern pike transcriptome (repeat-masked), Atlantic salmon EST data [9] and *Gasterosteus aculeatus* transcripts obtained from the UCSC genome browser. Intra-transcriptome comparisons for northern pike, Atlantic salmon and threespine stickleback, as well as between-species comparisons (northern pike vs. Atlantic salmon and northern pike vs. threespine stickleback) were performed using a reciprocal best BLASTN approach. Only BLASTN alignments longer than 100 bp and possessing an E-value ≤1e-5 were considered in establishing homology. DAVID analyses [85] were performed using paralogues from *Esox lucius* that exhibited 77–89% similarity in the aforementioned BLASTN alignments. Results were categorized by molecular function and biological process, and sorted by EASE-Score (p-value).

RAD-tag based linkage mapping

A RAD-tag-based approach for linkage map construction was attempted following the protocol of Amores et al. [53] using the *Sbf*I restriction enzyme. Tissues from a family consisting of a single father and mother (fin clips) and 94 progeny (yolk-sac larvae) from the Hackettstown Fish Hatchery, New Jersey, USA were extracted and used to construct three barcoded Illumina libraries, each containing data from 32 individuals. These libraries were paired-end sequenced using an Illumina HiSeq-2000 sequencer (Michael Smith Genome Sciences Centre, Vancouver, BC, CANADA). Subsequent analysis to identify and score SNPs was performed with Stacks [86].

Microsatellite Primer Design

A second family was obtained from the Hackettstown Fish Hatchery, New Jersey. Tissues stored in 95% ethanol from 103 progeny (either yolk-sac larvae or fin clips from 4-month old juveniles) of a half-sibling family produced from 3 fathers and a single mother were extracted by placing 2 mm^3 of tissue in 5% Chelex 100 (Biorad), 0.2% SDS and 0.27 mg/ml proteinase K (Invitrogen) for 2 hours at 55°C followed by 95°C for 10 minutes [87]. Extractions were diluted 1/100 in water for use in PCR. Offspring were screened using microsatellite primers to determine parentage (see below for primer design); progeny were distributed 106:45:0 between the three fathers. The family was therefore used as a two-father/single mother half-sibling family in further analysis, split 48:44 progeny per father to fit a 96-well format.

776 primer pairs were designed to target di- and tri-nucleotide repeats identified by Tandem Repeats finder [88] from *Esox* EST sequences [10], with an additional 835 primer pairs designed to target microsatellite repeats identified from genome scaffolds (1–3 primer pairs evaluated per scaffold). The unlabelled EST-designed primer pairs, u001-u296 were evaluated on a 5% polyacrylamide gel prior to re-ordering HEX or 6-FAM directly-labelled forward primers. The forward primer from each of the remaining primer pairs (u297–u776) and all scaffold primers were ordered with one of three tails, Tail-A, Tail-B or Tail-C for fluorescently labelling products with a labelled universal third primer, adapted from Blackett et al. protocol [89]. The universal primers were labelled: Tail-A [90] with 6-FAM, Tail-B [90] with HEX, Tail-C [89] with NED, 6-FAM and HEX.

For directly labelled primers, each PCR reaction contained: ~2.5 ng template DNA, 200 µM each dNTP (Promega), 0.5 µM each forward (HEX or 6-FAM labelled) and reverse primer (IDT), 0.25 U Hot Start *Taq* DNA polymerase (Promega), 1X GoTaq Flexi Colorless PCR buffer (Promega), 2.0 mM $MgCl_2$, to a final volume of 10 µl with DNase/RNase free H_2O (Gibco). Reactions were cycled on a TC-412 thermocycler (Techne) following the profile: 3 min at 95°C followed by 35 cycles of 30 s at 95°C, 30 s at 52°C and 30 s at 72°C, and a final extension of 72°C for 10 min with a hold at 4°C. For "Tailed" reactions: ~2.5 ng template DNA, 200 µM each dNTP (Promega), 0.15 µM forward primer (IDT), 0.5 µM reverse primer (IDT), 0.2 µM labelled universal tail primer (IDT), 0.25 U Hot Start *Taq* DNA polymerase (Promega), 1X GoTaq Flexi Colorless PCR buffer (Promega), 2.0 mM $MgCl_2$, and final volume of 10 µl with DNase/RNase free H_2O (Gibco). Tailed reactions were cycled on the same profile as above, except the 52°C annealing step was replaced with 56°C for 1 minute 30 seconds.

Following PCR, genotyping reactions consisted of 1.0 µl each PCR product, 9.9 µl Hi-Di Formamide (Life Technologies) and

0.1 µl GeneScan -500 ROXSize Standard (Life Technologies); two or three PCR reactions with different dyes were pooled into a single genotyping reaction. Following denaturation for 3–5 minutes at 95°C, samples were cooled on ice for 5–10 minutes before analysis on a 3730 DNA analyzer (Life Technologies). The generated electropherograms were analysed using GeneMapper V4.0 (Life Technologies), with automated scoring reviewed manually for errors in all samples. Samples for which clear peaks were not identified were removed from the dataset in subsequent analyses. All primer pairs were evaluated on the mother and two fathers of the family. Markers identified as polymorphic in the parents were used to amplify and score the distribution of alleles in the progeny. Polymorphic marker information is summarized in Table S7.

Linkage Mapping

Linkage mapping was performed using the LINKMFEX package, v 2.3 (R. Danzmann, University of Guelph, http://www.uoguelph.ca/~rdanzman/software.htm) following the standard protocol; markers with more than 15% of genotypes missing were omitted from the analysis (apart from u0028c; see below). Linkage groups were identified at an LOD threshold of 4.0; linkage groups in the individual maps were further joined at an LOD of 3 if in another individual the linkage was recognized using the stricter cut-off. Further estimation of distances below LOD = 3 was used to create merged, sex-specific maps; these are represented by dotted lines in Figure S1. Merged sex-specific and final merged maps were produced using the MERGE programs in the LINKMFEX package. Maps were visualized using the program MAPCHART [91]. The linkage of u0028c (with more than 15% missing data in original family) to adjacent markers in LG-04 was confirmed using a separate, unrelated family from the Whiteshell fish hatchery (West Hawk Lake, Manitoba, CANADA), extracted, amplified and analysed using the previously described procedures.

Synteny analyses

Four datasets were used for synteny analyses: 1) the reference *Esox lucius* transcriptome (see 'RNA-seq and assembly' section of Methods); 2) stickleback protein sequences from the Feb. 2006 Broad/gasAcu1 release; 3) zebrafish protein sequences from the Jul. 2010 Zv9/danRer7 release and; 4) medaka protein sequences from the Oct. 2005 NIG/UT MEDAKA1/oryLat 2 release. Stickleback, zebrafish and medaka sequences and their associated genomic location information were obtained from the UCSC Genome Browser [92]. Scaffold locations for northern pike transcripts were obtained through mapping using GMAP [93]; linkage group assignments followed if the host scaffold had been previously mapped to a group in the genetic map. Using the BLASTX and TBLASTN programs, BLAST alignments (E-value ≤1e-5) were obtained between the northern pike transcripts and the proteins of each other fish species. Orthology between northern pike transcripts and other fish protein sequences was determined using the reciprocal best hit (RBH) paradigm requiring at least 50% of each sequence was covered in non-overlapping BLAST alignments (HSPs) from the other. Synteny between two species (Figure 6) and scaffold continuity (Figure S2) were examined by plotting the genomic locations of each sequence in a relevant orthologue pair.

The analysis of synteny between northern pike and Atlantic salmon (Table 2) was performed by obtaining the flanking sequence of chromosome-associated SNPs in Atlantic salmon and identifying the strongest BLASTN hits (E-value ≤1e-10) between these sequences and northern pike scaffolds with a known linkage group.

Supporting Information

Figure S1 Individual male and sex-specific linkage maps for the northern pike. Sex-specific maps and how markers are merged into the final, merged linkage map are presented. Individual linkage maps produced from each of the two fathers are also presented, as well as positions for which a distance was estimated below an LOD of 3.0 for efficient merging. Star symbol denotes markers designed from the same scaffold that appear disrupted by one or more additional scaffolds in the merged linkage map, due to lack of shared informative markers in merging sex-specific maps.
(PDF)

Figure S2 Conservation of gene order within the largest 50 northern pike genomic scaffolds relative to *Gasterosteus aculeatus*, *Oryzias latipes*, and *Danio rerio*. Reciprocal best hit results (≥50% of total length, ≤1e-5) and mapped scaffold position plotted. Average position ((Start position + End position)/2) taken to represent a point position for each mapped transcript.
(PDF)

Table S1 Contribution of major Transposable Element categories to the *Esox lucius* genome. Produced by parsing results from Repeat masker output; variation between the direct Repeat masker output (18.1% of genome) and the sum of the annotated categories is likely due to overlap in the masked elements.
(XLSX)

Table S2 Trinity-assembly FPKM results for 13 *Esox lucius* tissues. All values submitted to NCBI GEO profile database.
(XLSX)

Table S3 MAKER2-assembly FPKM results for 13 *Esox lucius* tissues.
(XLSX)

Table S4 Top 10 expressed specialized transcripts from each tissue in *Esox lucius*. Sorted by highest FPKM for transcripts >3 standard deviations above average across all tissues; based on results from 38K Trinity-assembled transcriptome.
(XLSX)

Table S5 Pathways Identified in duplicated transcripts. Significant DAVID results (sorted by p-value) for duplicated transcripts between 77–89% similarity in northern pike. Gene_Ontology terms presented from the molecular function (GOTERM_MF_FAT) and biological process (GOTERM_BP_-FAT) categories.
(XLSX)

Table S6 Comparative synteny between *Esox lucius* and three model organisms. Reciprocal blast results (≥50% of total length, ≤1e-5) between scaffold-linked *E. lucius* transcripts and published genome assemblies of A) *Gasterosteus aculeatus* (gasAcu1 – Feb 2006), B) *Oryzias latipes* (oryLat2 – Oct 2005) and C) *Danio rerio* (ZV9 - Jul 2010). The table provides numerical results used to produce Figure 6.
(XLSX)

Table S7 Microsatellite marker information and primer sequences used in construction of first-generation linkage map. Includes relevant information regarding the

location of the microsatellite marker within the scaffolds, the primers used to amplify the repeats, and the results from tandem repeats finder representing information on the repeats, based on the EST/scaffold sequence used to design the primer pairs targeting the repeats.
(XLSX)

File S1 19K MAKER2-assembled transcriptome in FASTA format.
(GZ)

Acknowledgments

We would like to thank Doug Watkinson for tissues used in RNA-seq and the staff at the Whiteshell hatchery (West Hawk Lake, Manitoba, CANADA) and the Hackettstown Fish Hatchery (Hackettstown, New Jersey, USA) for their work in producing and collecting the pike families used in linkage mapping.

Author Contributions

Conceived and designed the experiments: BFK EBR. Performed the experiments: EBR DRM JSL AMM JRJ KRVS NHB. Analyzed the data: EBR DRM JSL AMM. Contributed reagents/materials/analysis tools: BFK CL. Contributed to the writing of the manuscript: EBR DRM JSL BFK.

References

1. Betancur RR, Broughton RE, Wiley EO, Carpenter K, Lopez JA, et al. (2013) The tree of life and a new classification of bony fishes. PLoS Curr 5.
2. Nelson JS (2006) Fishes of the World, 4th edition. Hoboken, NJ: Wiley. 624 p.
3. Li J, Xia R, McDowall RM, López JA, Lei G, et al. (2010) Phylogenetic position of the enigmatic *Lepidogalaxias salamandroides* with comment on the orders of lower euteleostean fishes. Mol Phylogenet Evol 57: 932–936.
4. Near TJ, Eytan RI, Dornburg A, Kuhn KL, Moore JA, et al. (2012) Resolution of ray-finned fish phylogeny and timing of diversification. Proc Natl Acad Sci U S A 109: 13698–13703.
5. Zou M, Guo B, Tao W, Arratia G, He S (2012) Integrating multi-origin expression data improves the resolution of deep phylogeny of ray-finned fish (Actinopterygii). Sci Rep 2: 665.
6. Campbell MA, López JA, Sado T, Miya M (2013) Pike and salmon as sister taxa: Detailed intraclade resolution and divergence time estimation of Esociformes + Salmoniformes based on whole mitochondrial genome sequences. Gene 530: 57–65.
7. Davidson WS, Koop BF, Jones SJ, Iturra P, Vidal R, et al. (2010) Sequencing the genome of the Atlantic salmon (*Salmo salar*). Genome Biol 11: 403.
8. Thorgaard GH, Bailey GS, Williams D, Buhler DR, Kaattari SL, et al. (2002) Status and opportunities for genomics research with rainbow trout. Comp Biochem Physiol B Biochem Mol Biol 133: 609–646.
9. Koop BF, von Schalburg KR, Leong J, Walker N, Lieph R, et al. (2008) A salmonid EST genomic study: genes, duplications, phylogeny and microarrays. BMC Genomics 9: 545.
10. Leong JS, Jantzen SG, von Schalburg KR, Cooper GA, Messmer AM, et al. (2010) *Salmo salar* and *Esox lucius* full-length cDNA sequences reveal changes in evolutionary pressures on a post-tetraploidization genome. BMC Genomics 11: 279.
11. Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, et al. (2006) The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol 7: R43.
12. Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y (2004) Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. Proc Natl Acad Sci U S A 101: 1638–1643.
13. Christoffels A, Koh EG, Chia JM, Brenner S, Aparicio S, et al. (2004) Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Mol Biol Evol 21: 1146–1151.
14. Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, et al. (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431: 946–957.
15. Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y (2003) Genome duplication, a trait shared by 22000 species of ray-finned fish. Genome Res 13: 382–390.
16. Allendorf F, Thorgaard G (1984) Tetraploidy and the Evolution of Salmonid Fishes. In: Turner B, editor. Evolutionary Genetics of Fishes: Springer US. pp. 1–53.
17. Shiina T, Dijkstra J, Shimizu S, Watanabe A, Yanagiya K, et al. (2005) Interchromosomal duplication of major histocompatibility complex class I regions in rainbow trout (*Oncorhynchus mykiss*), a species with a presumably recent tetraploid ancestry. Immunogenetics 56: 878–893.
18. Macqueen DJ, Johnston IA (2014) A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proc R Soc B 281.
19. Conant GC, Wolfe KH (2008) Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet 9: 938–950.
20. Dittmar K, Liberles D (2011) Evolution after Gene Duplication: Wiley.
21. Van de Peer Y, Maere S, Meyer A (2009) The evolutionary significance of ancient genome duplications. Nat Rev Genet 10: 725–732.
22. Semon M, Wolfe KH (2008) Preferential subfunctionalization of slow-evolving genes after allopolyploidization in *Xenopus laevis*. Proc Natl Acad Sci U S A 105: 8333–8338.
23. de Boer JG, Yazawa R, Davidson WS, Koop BF (2007) Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. BMC Genomics 8: 422.
24. Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, et al. (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat Commun 5: 3657.
25. Mank JE, Avise JC (2006) Phylogenetic conservation of chromosome numbers in Actinopterygiian fishes. Genetica 127: 321–327.
26. Gregory TR (2013) Animal Genome Size Database. http://www.genomesize.com.
27. Beamish RJ, Merrilees MJ, Crossman EJ (1971) Karyotypes and DNA values for members of the suborder Esocoidei (Osteichthyes: Salmoniformes). Chromosoma 34: 436–447.
28. Ráb P, Crossman EJ (1994) Chromosomal NOR phenotypes in North American pikes and pickerels, genus *Esox*, with notes on the Umbridae (Euteleostei: Esocae). Can J Zool 72: 1951–1956.
29. Crossman EJ, Ráb P (2001) Chromosomal NOR Phenotype and C-Banded Karyotype of Olympic Mudminnow, *Novumbra hubbsi* (Euteleostei: Umbridae). Copeia 2001: 860–865.
30. Craig JF (2008) A short review of pike ecology. Hydrobiologia 601: 5–16.
31. Nilsson PA, Skov C, Farrell J (2008) Current and future directions for pike ecology and management: a summary and synthesis. Hydrobiologia 601: 137–141.
32. Raat AJP (1988) Synopsis of Biological Data on the Northern Pike. FAO United Nations 30: 1–187.
33. Miller LM, Senanan W (2003) A Review of Northern Pike Population Genetics Research and Its Implications for Management. N Am J Fish Manage 23: 297–306.
34. Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A 108: 1513–1518.
35. Hardie DC, Hebert PDN (2003) The nucleotypic effects of cellular DNA content in cartilaginous and ray-finned fishes. Genome 46: 683–706.
36. Vinogradov AE (1998) Genome size and GC-percent in vertebrates as determined by flow cytometry: The triangular relationship. Cytometry 31: 100–109.
37. Hardie DC, Hebert PDN (2004) Genome-size evolution in fishes. Can J Fish Aquat Sci 61: 1636–1646.
38. Poptsova MS, Il'icheva IA, Nechipurenko DY, Panchenko LA, Khodikov MV, et al. (2014) Non-random DNA fragmentation in next-generation sequencing. Sci Rep 4: 4532.
39. Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrom M, et al. (2011) The genome sequence of Atlantic cod reveals a unique immune system. Nature 477: 207–210.
40. Amemiya CT, Alfoldi J, Lee AP, Fan S, Philippe H, et al. (2013) The African coelacanth genome provides insights into tetrapod evolution. Nature 496: 311–316.
41. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, et al. (2013) The zebrafish reference genome sequence and its relationship to the human genome. Nature advance online publication.
42. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, et al. (2012) The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484: 55–61.
43. Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, et al. (2007) The medaka draft genome and insights into vertebrate genome evolution. Nature 447: 714–719.
44. Aparicio S, Chapman J, Stupka E, Putnam N, Chia J-m, et al. (2002) Whole-Genome Shotgun Assembly and Analysis of the Genome of *Fugu rubripes*. Science 297: 1301–1310.
45. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29: 644–652.

46. He L, Wang Q, Jin X, Wang Y, Chen L, et al. (2012) Transcriptome Profiling of Testis during Sexual Maturation Stages in *Eriocheir sinensis* using Illumina Sequencing. PLoS One 7: e33735.
47. Fletcher GL, Hew CL, Davies PL (2001) Antifreeze proteins of teleost fishes. Annu Rev Physiol 63: 359–390.
48. Lynch M, Conery JS (2000) The Evolutionary Fate and Consequences of Duplicate Genes. Science 290: 1151–1155.
49. Katju V, Bergthorsson U (2013) Copy-number changes in evolution: rates, fitness effects and adaptive significance. Frontiers in genetics 4: 273.
50. Lu J, Peatman E, Tang H, Lewis J, Liu Z (2012) Profiling of gene duplication patterns of sequenced teleost genomes: evidence for rapid lineage-specific genome expansion mediated by recent tandem duplications. BMC Genomics 13: 246.
51. Brunet FG, Crollius HR, Paris M, Aury J-M, Gibert P, et al. (2006) Gene Loss and Evolutionary Rates Following Whole-Genome Duplication in Teleost Fishes. Mol Biol Evol 23: 1808–1816.
52. Wang JT, Li JT, Zhang XF, Sun XW (2012) Transcriptome analysis reveals the time of the fourth round of genome duplication in common carp (*Cyprinus carpio*). BMC Genomics 13: 96.
53. Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH (2011) Genome Evolution and Meiotic Maps by Massively Parallel DNA Sequencing: Spotted Gar, an Outgroup for the Teleost Genome Duplication. Genetics 188: 799–808.
54. Gonen S, Lowe NR, Cezard T, Gharbi K, Bishop SC, et al. (2014) Linkage maps of the Atlantic salmon (*Salmo salar*) genome derived from RAD sequencing. BMC Genomics 15: 166.
55. Palaiokostas C, Bekaert M, Davie A, Cowan ME, Oral M, et al. (2013) Mapping the sex determination locus in the Atlantic halibut (*Hippoglossus hippoglossus*) using RAD sequencing. BMC Genomics 14: 566.
56. Miller LM, Kapuscinski AR (1996) Notes: Microsatellite DNA Markers Reveal New Levels of Genetic Variation in Northern Pike. Trans Am Fish Soc 125: 971–977.
57. Gotoh RO, Tamate S, Yokoyama J, Tamate HB, Hanzawa N (2013) Characterization of comparative genome-derived simple sequence repeats for acanthopterygian fishes. Molecular Ecology Resources 13: 461–472.
58. Wang J, Wang C, Qian L, Ma Y, Yang X, et al. (2011) Genetic characterization of 18 novel microsatellite loci in northern pike (*Esox lucius* L.). Genet Mol Biol 34: 169–172.
59. Launey S, Krieg F, Morin J, Laroche J (2003) Five new microsatellite markers for Northern pike (*Esox lucius*). Mol Ecol Notes 3: 366–368.
60. Ouellet-Cauchon G, Normandeau E, Mingelbier M, Bernatchez L (2014) EST-based microsatellites for northern pike (*Esox lucius*) and cross-amplification across all Esox species. Conservation Genet Resour 6: 451–454.
61. Lien S, Gidskehaug L, Moen T, Hayes BJ, Berg PR, et al. (2011) A dense SNP-based linkage map for Atlantic salmon (*Salmo salar*) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. BMC Genomics 12: 615.
62. Rexroad CE, 3rd, Palti Y, Gahr SA, Vallejo RL (2008) A second generation genetic map for rainbow trout (*Oncorhynchus mykiss*). BMC Genet 9: 74.
63. Singer A, Perlman H, Yan Y, Walker C, Corley-Smith G, et al. (2002) Sex-Specific Recombination Rates in Zebrafish (*Danio rerio*). Genetics 160: 649–657.
64. Rondeau EB, Messmer AM, Sanderson DS, Jantzen SG, von Schalburg KR, et al. (2013) Genomics of Sablefish (*Anoplopoma fimbria*): Expressed genes, mitochondrial phylogeny, linkage map and identification of genetic sex markers. BMC Genomics 14: 452.
65. Wolfe K (2000) Robustness–it's not where you think it is. Nat Genet 25: 3–4.
66. Nordine RN, Pommen LW (1985) Peace River Area Charlie lake Sub-Basin Water Quality Assessment and Objectives. http://www.env.gov.bc.ca/wat/wq/objectives/charlie/Charlietech.pdf.
67. Sambrook J, Russell DW (2001) Molecular Cloning: A Laboratory Manual, 3rd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
68. Wegrzyn JL, Lin BY, Zieve JJ, Dougherty WM, Martínez-García PJ, et al. (2013) Insights into the Loblolly Pine Genome: Characterization of BAC and Fosmid Sequences. PLoS One 8: e72439.
69. Flutre T, Duprat E, Feuillet C, Quesneville H (2011) Considering Transposable Element Diversification in *De Novo* Annotation Approaches. PLoS One 6: e16526.
70. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110: 462–467.
71. Jurka J, Klonowski P, Dagman V, Pelton P (1996) CENSOR–a program for identification and elimination of repetitive elements from DNA sequences. Comput Chem 20: 119–121.
72. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26: 2460–2461.
73. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302: 205–217.
74. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.
75. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.
76. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, et al. (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8: 973–982.
77. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25: 1422–1423.
78. Talevich E, Invergo BM, Cock PJ, Chapman BA (2012) Bio.Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. BMC Bioinformatics 13: 209.
79. Smit AFA, Hubley R, Green P (1996–2010) RepeatMasker Open-3.0 http://www.repeatmasker.org.
80. Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, et al. (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res 40: W622–627.
81. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
82. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The Human Genome Browser at UCSC. Genome Res 12: 996–1006.
83. Holt C, Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12: 491.
84. Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19: ii215–ii225.
85. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44–57.
86. Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) *Stacks*: Building and Genotyping Loci *De Novo* From Short-Read Sequences. G3 (Bethesda) 1: 171–182.
87. Messmer AM, Rondeau EB, Jantzen SG, Lubieniecki KP, Davidson WS, et al. (2011) Assessment of population structure in Pacific *Lepeophtheirus salmonis* (Krøyer) using single nucleotide polymorphism and microsatellite genetic markers. Aquaculture 320: 183–192.
88. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27: 573–580.
89. Blacket MJ, Robin C, Good RT, Lee SF, Miller AD (2012) Universal primers for fluorescent labelling of PCR fragments–an efficient and cost-effective approach to genotyping by fluorescence. Mol Ecol Resour 12: 456–463.
90. Roche Applied Science (2006) GS Guide to Amplicon Sequencing. Roche Diagnostics: GmbHUSM-0022.B - December 2006.
91. Voorrips RE (2002) MapChart: Software for the Graphical Presentation of Linkage Maps and QTLs. J Hered 93: 77–78.
92. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, et al. (2014) The UCSC Genome Browser database: 2014 update. Nucleic Acids Res 42: D764–770.
93. Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21: 1859–1875.