



OPEN

The salmon louse genome may be much larger than sequencing suggests

Grace A. Wyngaard¹, Rasmus Skern-Mauritzen², Ketil Malde^{2,3}, Rachel Prendergast¹ & Stefano Peruzzi⁴✉

The genome size of organisms impacts their evolution and biology and is often assumed to be characteristic of a species. Here we present the first published estimates of genome size of the ecologically and economically important ectoparasite, *Lepeophtheirus salmonis* (Copepoda, Caligidae). Four independent *L. salmonis* genome assemblies of the North Atlantic subspecies *Lepeophtheirus salmonis salmonis*, including two chromosome level assemblies, yield assemblies ranging from 665 to 790 Mbps. These genome assemblies are congruent in their findings, and appear very complete with Benchmarking Universal Single-Copy Orthologs analyses finding >92% of expected genes and transcriptome datasets routinely mapping >90% of reads. However, two cytometric techniques, flow cytometry and Feulgen image analysis densitometry, yield measurements of 1.3–1.6 Gb in the haploid genome. Interestingly, earlier cytometric measurements reported genome sizes of 939 and 567 Mbps in *L. salmonis salmonis* samples from Bay of Fundy and Norway, respectively. Available data thus suggest that the genome sizes of salmon lice are variable. Current understanding of eukaryotic genome dynamics suggests that the most likely explanation for such variability involves repetitive DNA, which for *L. salmonis* makes up ≈ 60% of the genome assemblies.

Abbreviations

BUSCO	Benchmarking universal single-copy orthologs
FIAD	Feulgen image analysis densitometry
FCM	Flow cytometry
Fl	Fluorescence
Gbp	Giga base pairs
Mbp	Mega base pairs
PI	Propidium iodide
SNP	Single nucleotide polymorphism

“In the future attention undoubtedly will be centered on the genome, and with greater appreciation of its significance as a highly sensitive organ of the cell, monitoring genomic activities and correcting common errors, sensing the unusual and unexpected events, and responding to them, often by restructuring the genome”—Barbara McClintock’s Nobel Lecture in 1983¹.

A lot has been learned since 1983, and numerous genomes have been sized, sequenced and analyzed. Yet, many questions regarding genomes remain unanswered, the most fundamental potentially being: why do eukaryotic genomes vary so much in size? While complexity appears to correlate with minimum taxon genome size, the actual genome sizes bear no straightforward correlation with eukaryotic organismal complexity, even among closely related taxa, but are increasingly investigated as a trait subject to natural selection and consequently of relevance to studies of ecology and evolution^{2–4}. Selective pressures in copepods have been posed for age at first reproduction in predation intense environments, resulting in smaller genome sizes⁵, as well as selection for larger bodies and genome sizes in cold environments^{6,7}. Causal links between the ‘bulk’ DNA amount, cell division rate, and cell volume, as explained by the nucleotypic hypothesis⁸, may underlie relationships between these cellular parameters and organismal development rates and body size, especially in the copepods which possess eutely⁹.

¹Department of Biology, James Madison University, Harrisonburg, VA, USA. ²Institute of Marine Research, Bergen, Norway. ³Department of Informatics, University of Bergen, Bergen, Norway. ⁴Department of Arctic Marine Biology, UiT-the Arctic University of Norway, Tromsø, Norway. ✉email: stefano.peruzzi@uit.no

While genome size does not appear to govern organismal complexity, some relationships appear to be general: genome size often correlates with the proportion of noncoding, or repetitive, DNA in the genome^{8,10}, cell size^{2,11} and growth rate¹². Furthermore, the evolutionary importance of repetitive elements (mainly transposable elements—TEs) in lateral gene transfer¹³ and generation of new phenotypes¹⁴ is becoming increasingly apparent. This is well illustrated by TEs being responsible for more than 50% of the phenotypes emerging in *Drosophila* laboratory strains¹⁵ and playing a role in adaptive evolution^{16,17}. At the same time, it must be realized that species specific effects may affect genome sizes in ways that appear to be inconsistent with the general trends: for example taxon specific allocation of phosphorus to RNA rather than nonessential non-coding DNA may result in a selection for compact genomes in phosphorus limited environments^{18,19}. As the role(s) of noncoding and repetitive DNA become better understood, the importance of knowing to what extent this component of the genome has been accurately included in the assemblies and annotations becomes increasingly clear. This does not contradict the fact that partial genome assemblies may be both of high quality and immense value.

The salmon louse (*Lepeophtheirus salmonis*, Krøyer 1837) is a marine parasitic copepod of large economic and ecological importance^{20,21}. It belongs to the order Siphonostomatoidea and is found on salmonid fishes in the northern hemisphere. There are two *L. salmonis* subspecies separated by approximately 5 million years of evolution, *L. salmonis salmonis* (Krøyer, 1837) inhabiting the North Atlantic and *L. salmonis onchorhynchii*²¹ inhabiting the Northern Pacific^{22–24}. These parasites alter the physiology, disease susceptibility, growth rates, and behavior of their salmonid hosts^{25–27} and inflict large economic losses²⁸. As the salmonid aquaculture has expanded to the extent that farmed salmonids outnumber wild salmonids by 2–3 orders of magnitude in some regions in the North Atlantic, the salmon lice populations have increased in parallel and currently inflict significant economic and ecological challenges^{28,29}. The combined societal and ecological impacts of *L. salmonis* have spurred intense research and tool development, including modelling to assess ecological risk^{30,31}, development of methodology for surveillance^{25,32}, studies of population genetics^{33–35}, resistance and resilience development against delousing agents^{36,37} and molecular biology^{38–41}. As a result, the salmon louse genome has been sequenced several times using various sequencing platforms, and independent genome assemblies have been made^{41,42}—including two chromosome level assemblies.

As DNA sequencing technologies have advanced and improvements in the identification and annotation of noncoding DNA have gradually followed, there is a growing awareness that genome assembly methods sometimes fail to correctly reconstruct repetitive regions and noncoding DNA [43,44]. Traditional quantitative cytogenetic methods such as flow cytometry and Feulgen microdensitometry are recognized as being reliable with respect to estimating the true amount of nuclear DNA contents and providing estimates of total genome size^{45,46}. Thus, genomic assemblies and cytometric methods possess different strengths with respect to the kinds of information they provide. When these two approaches are applied in combination, they are likely to either validate the independent estimates or provide direction as to seeking explanations for the discrepancies. In light of the importance of the *L. salmonis* genome it seems prudent to use cytometric methods to validate sizes estimated from genome sequencing.

In the present study of the salmon louse, we compare genome size estimates based on two quantitative cytometric methods (flow cytometry and Feulgen image analysis densitometry) performed on multiple samples with unpublished cytometric measurements and estimates based on whole genome sequencing. Where discrepancies are apparent, we propose explanations and a path forward for resolving them. Additionally, as this is the first published estimate of the genome size of a parasitic copepod, the genome size of this siphonostomatoid is discussed in relation to the free-living copepods in the orders Cyclopoida, Calanoida, and Harpacticoida.

Results

Sequencing—and assembly based genome size estimates. The salmon louse genome has been sequenced several times using various sequencing platforms, and six independent genome assemblies have been made—including two chromosome level assemblies (Table 1). The resulting assemblies appear consistent in structure, as revealed by linkage analyses^{34,41,47}, and sizes (Table 1), collectively suggesting the salmon louse genome size to be approximately 600–700 Mbps.

To evaluate the congruence of the assemblies, while not requiring them to conserve synteny, we created libraries of 240 bp synthetic reads from each of the assemblies. These synthetic reads were then mapped to each of the published assemblies using BLAST. The results show that the assemblies are close to interchangeable in terms of their sequence composition (Table 2) and that the differences in the sequence captured by the different sequencing technologies are minor.

It is well known that genome assembly sizes may deviate significantly from the actual genome size¹⁰ and additional sequence-based genome size estimates were therefore produced. First, k-mer analysis was performed using Jellyfish⁵¹. The published GLW4 dataset originating from inbred salmon lice⁴¹ and word sizes (k-mer lengths) of 21–31 yielded estimated genome sizes of 976–1017 Mbps (modal k-mer coverages: 20–17). Repeating the analysis using previously published data from wild salmon lice³⁴ and using the word lengths of 29 and 31 yielded estimated genome sizes of 1086 and 1015 Mbps (modal k-mer coverages: 55 and 53).

Second, sequencing reads were mapped to the *LSalAtl2s* genome using BWA⁵². For each library, modal coverage (*M*) was extracted, and assumed to be representative of diploid coverage. All coverages for the genome were then summed and divided by the modal coverage to estimate the genome size. Under the assumption that repeat sequence occurring *N* times in the genome would have a coverage distribution centered on *N***M*, each location in the repeat would be counted *N* times. Seven inbred salmon lice libraries⁴¹ were used, with modal coverages of 3–23x, and resulting in genome size estimates of 791–1184 Mbps, with low coverage libraries giving the highest size estimates. The analysis was repeated with six libraries from wild lice³⁴, giving coverages of 5–17x, and size estimates of 845–1073, again with low coverage libraries leading to higher size estimates.

Assemblies and accession details	Assembly level	Louse origin	Size (Mbps)	N50 (Kbps)	Sequencing platform (s)	Depth	Reference
UVic_Lsal_1.0 GCA_016086655.1	Chromosome	Pacific	668.1	48,457	Oxford Nanopore, Illumina	45X	Database only***
UStir_LSAA PRJEB43242	Chromosome	Atlantic	632.5 609.3*	5100 43,017**	PacBio	107**	Database only***
LSalAtl2S licebase.org	Scaffold	Atlantic	695.4	478	Illumina HiSeq, 454, Sanger fosmid end	175X	⁴¹
Atlantic female GCA_001005205.1	Scaffold	Atlantic	665.1	16	Illumina HiSeq	45X	Database only***
Atlantic male GCA_001005235.1	Scaffold	Atlantic	665.3	13	Illumina HiSeq	55X	Database only***
ASM18125v2 GCA_000181255.2	Scaffold	Pacific	790.1	10	Illumina, 454	88X	⁴²

Table 1. Key statistics and sequencing platforms for salmon louse assemblies available March 2021. Depths of sequencing coverage are the values associated with the assemblies and are estimated based on assembly-indicated genome sizes (i.e. ≈ 700 Mbps). The ASM18125v2 assembly is quite fragmented and the authors indicate that the assembly overestimates the actual size suggested to be around 600 Mbps (<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA40179>). Only LSaAtl2s is extensively assessed and the predicted gene set appears to be quite complete as it contains 92.4% of the expected genes in a BUSCO analysis and maps $\approx 90\%$ of transcriptome reads^{48,49}. *The UStir_LSAA assembly comprised 158 sequences in total, 15 of which were identified as chromosomes. The value denoted ‘*’ is the assembly size for these 15 chromosomes. **Calculated based on the 632.5 Mbp assembly size. ***Data available in public repositories were included with permission from the assembly authors in adherence to the Fort Lauderdale agreement⁵⁰.

	ASM18125v2	Atlantic female	Atlantic male	UVic_Lsal_1.0	UStir_LSAA	LSaAtl2S
ASM18125v2 GCA_000181255.2		97.74% (0.90)	96.90% (0.90)	96.17% (0.97)	97.12% (0.88)	97.50% (0.88)
Atlantic female GCA_001005205.1	93.77% (0.88)		97.85% (0.98)	93.67% (0.89)	97.22% (0.95)	96.97% (0.93)
Atlantic male GCA_001005235.1	94.01% (0.88)	98.90% (0.98)		94.06% (0.89)	97.36% (0.95)	96.99% (0.93)
UVic_Lsal_1.0 GCA_016086655.1	96.44% (0.97)	98.31% (0.91)	97.52% (0.91)		98.13% (0.89)	98.12% (0.89)
UStir_LSAA PRJEB43242	93.48% (0.87)	97.47% (0.97)	96.59% (0.97)	93.75% (0.88)		97.74% (0.95)
LSaAtl2S licebase.org	94.56% (0.89)	98.58% (0.97)	97.59% (0.97)	94.80% (0.89)	98.57% (0.96)	

Table 2. Consistency of assemblies. The assemblies were converted to 240 bp synthetic reads that were blasted against all other assemblies. The assemblies from which synthetic reads originated are indicated in column headings and the reference assemblies are indicated in the row headings. The results show average query cover % and (the proportion of reads that maps with $> 95\%$ identity).

Genome size estimates based on FCM. The nuclear DNA contents of somatic (2C values) and gametic cells measured by flow cytometry (FCM, Runs 1–5) in gametic, naupliar or adult stages of Atlantic *L. salmonis* are reported in Table 3. Figure 1 shows representative fluorescence (FL histograms of the above cells analyzed together with reference standards). Overall, fluorescence histograms of propidium iodide (PI)-stained somatic cells or gametes indicated good resolution levels with coefficients of variation (CVs) in the range of 1–5% (Table 3).

The estimated 2C DNA contents of somatic cells in naupliar stages of *Ls Tromsø* were similar across replicates (FCM Runs 1 and 2), averaging 3.08 pg DNA per nucleus when using chicken (CEN) as an internal standard (Table 2) and did not vary significantly when both chicken and human (MNCs) standards were used simultaneously (FCM Run 2). Nauplii cannot visually be assigned a sex and thus it cannot be known for certain what sex ratio occurred in these samples although sex is genetically determined with a 50:50 ratio⁴¹.

The average nuclear 1C DNA contents of *Ls* Gulen sperm cells (FCM Run 4) were 1.68 and 1.69 pg DNA per nucleus, and did not differ when estimated using chicken or human standards. The estimated 2C value of oocytes (FCM Run 4) averaged 3.33 pg DNA per nucleus with no significant difference between values based on the two standards. A derived 2C value of sperm cells (twice the 1.69 pg DNA per sperm cell) does not differ significantly from that of the unfertilized oocytes, 3.33 pg DNA per nucleus (Table 2).

The nuclear DNA contents of males and females in both wild caught and laboratory reared *Tromsø* strains were compared in Run 5. The 2C DNA content of somatic cells averaged 3.01 and 3.19 pg DNA/nucleus in female and male specimens of a laboratory strain, respectively. The same trend was observed when analyzing somatic cells of wild specimens; the 2C values averaging 3.08 and 3.19 pg DNA per nucleus in females and males, respectively. Overall, within this FCM run male genome size estimates were consistently larger (ANOVA, $P = 0.0001$) than female genome size estimates whereas, the 2C DNA contents of somatic cells within a sex did not differ significantly between salmon lice of different origin and with no interaction between the two factors. Despite

Run	Origin	Stage	Tissue	N	2C value (pg) vs. CEN Mean ± SEM	2C value (pg) vs. MNCs Mean ± SEM	CV (%)	2C value (Gb)
1	LsTromsø	Nauplii	WB	10	3.08 ± 0.009	na	2–4	3.01
2	LsTromsø	Nauplii	WB	11	3.09 ± 0.005	3.06 ± 0.021	2–3	3.01*
3	LsTromsø	Adult ♀	CT	3	3.10 ± 0.020	3.12 ± 0.008	3–5	3.04*
		Adult ♂	CT	6	3.14 ± 0.139	3.14 ± 0.010	3–5	3.07*
4	LsGulen	Adult ♀	O	5	3.36 ± 0.003	3.31 ± 0.003	2–4	3.26*
		Adult ♂	S	5	1.69 ± 0.006	1.68 ± 0.002	1–3	1.65*
5	LsTromsø	Adult ♀	CT	5	3.01 ^b ± 0.022	na	2–3	2.94
	LsTromsø	Adult ♂	CT	5	3.19 ^a ± 0.016	na	3–4	3.12
	Wild	Adult ♀	CT	11	3.08 ^b ± 0.031	na	2–3	3.01
	Wild	Adult ♂	CT	5	3.19 ^a ± 0.013	na	3–5	3.12

Table 3. Nuclear DNA contents of somatic and gametic cells of *L. salmonis salmonis* as measured using flow cytometry. Runs 1, 2, 3, and 5 measured somatic cells of nauplii or adults of the LsTromsø strain. Run 4 measured gametic cells of the Ls Gulen laboratory strain. Run 5 compared the laboratory reared LsTromsø strain to wild caught adults from naturally infected fish reared in Tromsø. Chicken and/or human white blood cells were used as internal reference standards. CEN Chicken erythrocyte nuclei (2C value = 2.5 pg DNA per nucleus), MNCs human mono-nucleated cells (2C value = 7.0 pg DNA per nucleus), WB whole body, CT cuticular and subcuticular tissues from cephalic region, O oocytes, S sperm, N number of individuals or number of samples in the cases of nauplii analyzed, CV coefficient of variation as a percentage of mean for target nuclei (*L. salmonis* data), Na not available. For naupliar stages (run 1 and 2), each sample consisted of approximately 50 nauplii. *Average value based on the two internal standards. For Run 5, 2C values with superscripts “a” and “b” differ at $P < 0.05$ (two-way ANOVA).

the lack of a statistically significant difference between male and female adults of laboratory reared Ls Tromsø salmon lice recorded in FCM Run 3, the DNA content of these somatic cells did not significantly differ from the laboratory and wild caught adults measured in Run 5, when comparisons were made within a sex.

Feulgen image analysis densitometry (FIAD)—nuclear morphologies of somatic tissues. The squashed somatic tissues possessed a variety of nuclear morphologies, from dense to diffuse (Fig. 2), which yielded a corresponding variety of values of integrated optical density (IOD). Such heterogeneity in nuclear morphology was not observed in either chicken or trout erythrocyte standards, as they were comprised solely of erythrocytes (Fig. 2a,b). We made decisions to select for measurement nuclei with intermediate morphologies that possessed a granular and slightly diffuse appearance (Fig. 2c,d). Nuclei that were very densely staining in appearance, or compact (Fig. 2e), yielded IOD values at the lower end of the range, possibly due to DNA compaction. Nuclei with a very diffuse appearance and sometimes with nuclear membranes possessing uneven edges that might indicate partial degradation (Fig. 2f), tended to possess IOD values at the higher end of the range. The corresponding 2C values were 2.1 pg DNA per nucleus for densely stained nuclei to 3.4 pg DNA per nucleus for diffuse nuclei.

Genome sizes based on FIAD. Based on Feulgen image analysis densitometry of the Ls Gulen laboratory strain and the chicken standard, the average somatic nuclear DNA contents of individual adult males (2.86–2.93 pg DNA per nucleus) were consistently larger than the average nuclear DNA contents of individual females (2.64–2.78 pg DNA per nucleus) (Table 4). Despite the consistency of the trend, the average of three individual females, 2.70 pg DNA per nucleus, did not differ significantly from the average of three individual males, 2.90 pg of DNA per nucleus (Table 4). The two sample *t* tests were based on 3 individuals per sex, rather than 120 nuclei per sex (Table 4; the N values for LsGulen), to avoid pseudoreplication.

A single wild caught adult female collected from a Maine population possessed nuclei in an appendage that were especially well suited for measurement as they were well isolated and lacked any visible background stain. The mean value of these nuclei, 3.07 pg DNA per nucleus, does not differ from the wild female that was caught near Tromsø nor from the laboratory reared Ls Tromsø strain (Tables 3, 4). It should be noted that the Maine specimen was squashed and stained as above, but prior to employing the freeze cracking technique, and thus measurements were restricted to those few nuclei in the spine of a swimming leg that possessed a granular and slightly diffuse appearance, and were also isolated and surrounded by clear background.

Comparison of genome size estimates based on FIAD and FCM. Estimated 2C genome sizes of male and female laboratory reared Ls Gulen adults obtained using FIAD and based on the chicken standard (2.90 and 2.70 pg DNA per nucleus, respectively, Table 4) are within 10% of the estimates obtained using FCM on the Ls Tromsø laboratory strain (Table 3), assuming values of 3.2 and 3.0 pg DNA per nucleus for males and females, respectively, for the Ls Tromsø strain). Each slide containing a population of nuclei from a single adult in the FIAD analyses contained some values that overlapped with estimates obtained using FCM; however, these higher FIAD values did not equal the central tendency of values obtained using FCM. The average nuclear DNA

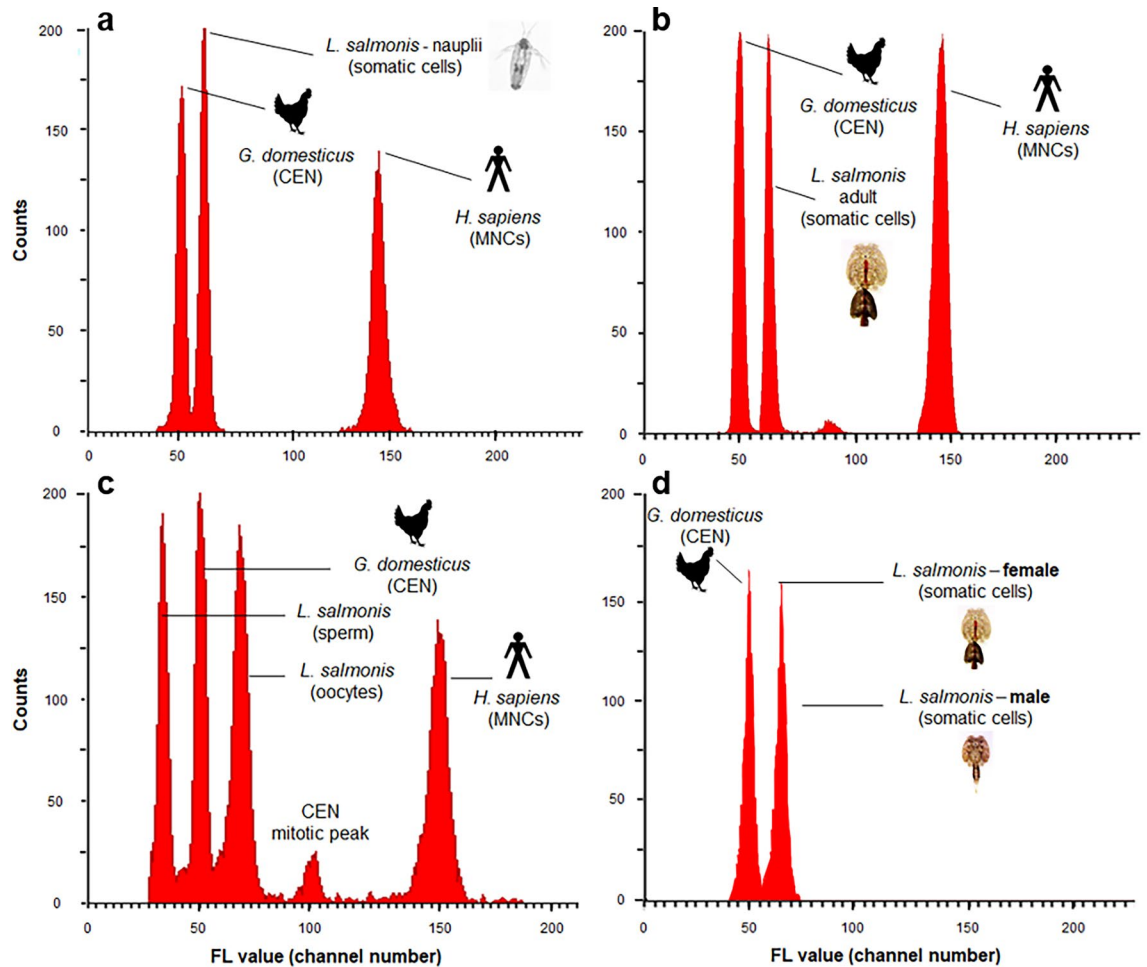


Figure 1. Representative fluorescence histograms of PI-stained somatic and gametic cells of *L. salmonis* *salmonis*. (a) Run 1, somatic cells obtained from whole body squash of nauplii; (b) run 3, somatic cells obtained from cephalic regions of adult females; (c) run 4; oocyte and sperm samples; (d) run 5, somatic cells obtained from cephalic regions of adult males and females. Samples analyzed using chicken and human cells as internal reference standards. Fluorescence (FL) peaks values on the X-axis are reported in arbitrary units expressed as Propidium Iodide (PI) fluorescence channel numbers (FL value). Values on the Y-axis (counts) refer to the number of nuclei counted per channel. *CEN* chicken erythrocytes nuclei, *MNCs* human mono-nucleated cells.

content obtained for the *Ls* Gulen adult females using FIAD, 2.70 pg DNA per nucleus, is within the 15% range of the value based on oocytes 3.26 pg DNA per nucleus (FCM run 4).

Discussion

L. salmonis assemblies are consistently approximately 700 Mbps suggesting that the *L. salmonis* 1C genome may be of approximately the same size, or possibly larger if the assemblies include substantial collapsed repeated regions. This approximate size was readily accepted by the salmon louse research community, as the first cytometric based estimate of a North Atlantic population was 0.58 pg (\approx 567 Mbps)⁵³. Thirteen years later a Bay of Fundy population measured in the same lab was estimated to be 0.96 pg (\approx 939 Mbps)⁵⁴. Yet in the present study the *L. salmonis* subsp. *salmonis* genome size estimates range from 1.3 to 1.6 Gbps when determined by two independent cytometric methods being applied to three laboratory strains and wild salmon lice from two locations. Sequence based extrapolations, in contrast, yield estimates ranging between 0.8 and 1.2 Gbps. In attempting to identify the factors responsible for the different estimates of *L. salmonis* genome size, we emphasize the importance of harnessing complementary approaches to estimate genome size and where discrepancies exist, not to disregard potentially 'missing' portions of DNA which may play an important role in adaptation. Additionally, while it is a common bias to interpret 'old measurements' as wrong when they disagree with new measurements, we remain open to a plethora of explanations to reconcile older and present study findings based on cytometric methods.

Genome assembly sizes are known to be unreliable predictors for genome size. Highly conserved repeats in combination with read errors can be difficult to resolve, and assemblies can have repeated regions collapsed into one sequence or have multiple copies of the same genomic region. More precise estimates can be achieved by examining the sequence reads. We have used two approaches, one using k-mer statistics and another based on mapping statistics against a reference assembly. These methods point to a genome size of 800–1200 based on

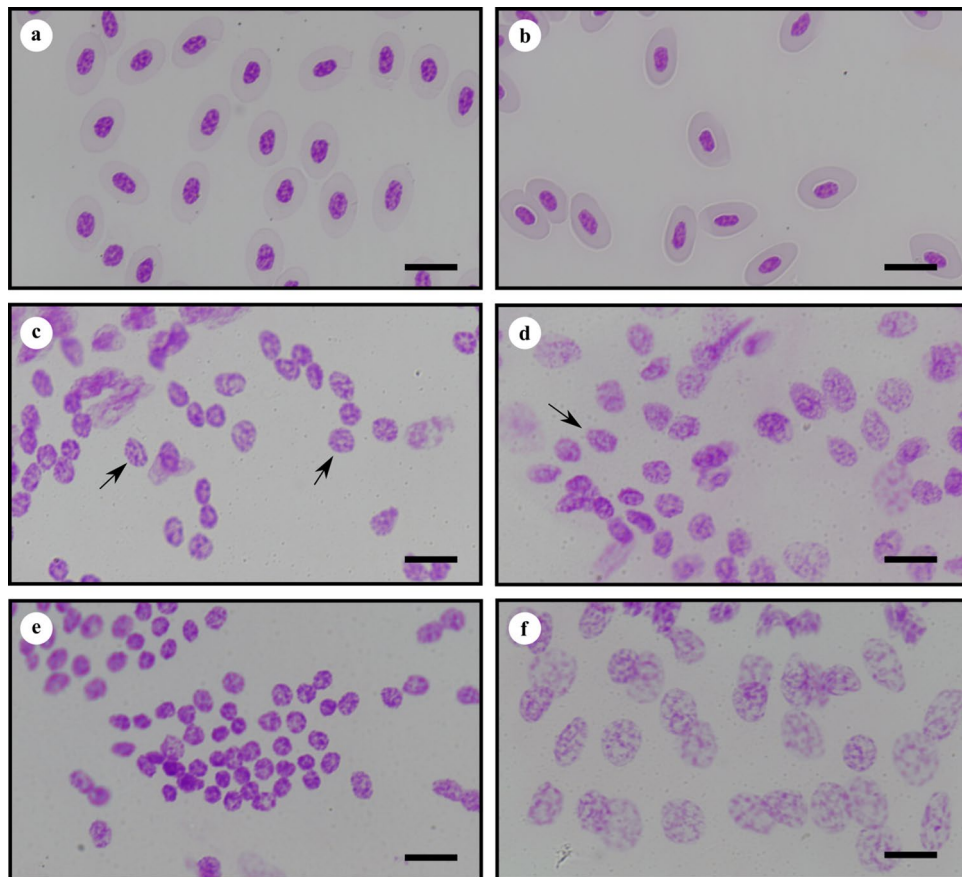


Figure 2. Photomicrographs of squash preparations of somatic nuclei of cephalothorax of *L. salmonis salmonis* and erythrocytes of hen and trout used as standards, stained with Feulgen reaction for DNA. (a) Nuclei of erythrocytes of *G. domesticus*; (b) nuclei of erythrocytes of *O. mykiss*; (c) representative nuclei, indicated by arrows, of male *L. salmonis salmonis* slide 64 whose measurements are given in Table 4; (d) representative nucleus, indicated by arrow, of female *L. salmonis salmonis* slide 63 whose measurements are given in Table 4; (e) nuclei of male *L. salmonis salmonis* slide 64 with high level of DNA compaction; (f) nuclei of male *L. salmonis salmonis* slide 64 with highly diffuse morphology. Scale bars represent 10 μm .

mapping, and 1000–1100 Mbp based on k-mers. One partial explanation for the variability can be unmapped reads (approximately 5% of the reads). If they represent sequence not present in the genome assembly, these genome components will be omitted from the mapping estimate, but included in the k-mer estimate. In addition, most of the sequence data is from female salmon lice, and both approaches will count the average of the haploid sex chromosomes, not their sum.

Flow cytometry is a well-established method for nuclear DNA content analysis and characterization in experimental biology, and is increasingly being used due to its rapidity, precision, and reproducibility. Feulgen image analysis densitometry similarly has experienced a resurgence in its use, partially due to its affordability and applicability when the number of available nuclei to measure is small. Interpretations of genome size estimates based on FCM and FIAD require careful consideration of the advantages and disadvantages of each method for the species and specific tissues under consideration. Explanations of differences among cytometric measurements in general, as well as those in the present study with unpublished estimates^{53,54}, performed on *L. salmonis* could include tissue compaction levels that can cause estimates to vary two-fold, misidentification of haploid and diploid cells, and misidentification of species, the latter of which seems quite unlikely as species identifications were provided by external expert collaborators. Not to be discounted is the possibility of real variation in genome size among populations. Jeffrey's⁵⁴ review highlights the issues of concern, particularly the chemistries and tissues measured most applicable to the present study, and are discussed in Supplemental Methods S1.

Measurements of nuclear DNA contents of naupliar and/or adult stages of two laboratory strains and wild caught *L. salmonis salmonis* based on FCM and two laboratory strains and one wild caught population of adults' stages based on FIAD estimate 2C somatic nuclear contents to be 3 ± 0.3 Gb. Halving the FCM values to obtain the 1C amount and directly measuring sperm DNA content yielded values ranging from 1.47 to 1.65 Gb (Table 3). Gametic nuclei of *L. salmonis salmonis* contained one half the DNA of the somatic nuclei, and indicates a lack of chromatin diminution, a phenomenon in which 1C values cannot be estimated by halving somatic genome sizes⁵⁵.

Male genome size was consistently slightly larger than female genome size, expectedly so due to erosion of the W-chromosome in the heterozygotic female⁵⁶. There was also no evidence of mitoses in the adult somatic

Origin	Stage	Tissue	N	2C value (pg) vs. CEN Mean \pm SEM	2C value (pg) vs. MNCs Mean \pm SEM	CV (%)	2C value (Gb)
Laboratory							
<i>Ls</i> Gulen 61	Adult ♀	CT	13	2.78 \pm 0.012	na	1.6	2.74
<i>Ls</i> Gulen 62	Adult ♀	CT	15	2.64 \pm 0.003	na	3.3	2.60
<i>Ls</i> Gulen 63	Adult ♀	CT	92	2.67 \pm 0.003	na	10.4	2.63
Average				2.70 \pm 0.043	na		2.66
<i>Ls</i> 1a 133	Adult ♀	CT	34	2.64 \pm 0.046	2.60 \pm 0.046	10.2	2.60
<i>Ls</i> 1a 134	Adult ♀	CT	20	2.75 \pm 0.037	2.71 \pm 0.037	6.1	2.67
<i>Ls</i> 1a 137	Adult ♀	CT	16	2.71 \pm 0.025	2.67 \pm 0.025	3.9	2.65
Average				2.70 \pm 0.032	2.66 \pm 0.032		2.66
<i>Ls</i> Gulen 64	Adult ♂	CT	65	2.93 \pm 0.035	na	9.9	2.89
<i>Ls</i> Gulen 65	Adult ♂	CT	36	2.86 \pm 0.037	na	7.9	2.82
<i>Ls</i> Gulen 66	Adult ♂	CT	19	2.90 \pm 0.054	na	8.3	2.85
Average				2.90 \pm 0.020	na		2.85
Wild							
Copscook Bay, Maine	Adult ♀	SP	3	3.07 \pm 0.176	na	9.8	3.00

Table 4. Nuclear DNA contents of somatic cells of individual adult *L. salmonis* as measured using Feulgen image analysis densitometry. Cephalothorax (CT) tissues of laboratory reared adults were obtained from the *Ls* Gulen and *Ls*1a populations. Tissue from the spine of an appendage (SP) was obtained from a wild caught Maine population. Hen (CEN) and male human mononucleated leucocytes (MNCs) were used as internal reference standards to estimate values in picograms (pg). Values based on hen were converted to gigabases (Gb). N refers to number of nuclei measured in each adult. SEM refers to standard error of the mean; CV refers to coefficient of variation of IOD values.

tissues, and therefore most adult somatic cells are suitable for genome size measurement⁹. Furthermore, we found no evidence of significant differences based on cytometric based comparisons of geographical (Norway, Maine), laboratory (*Ls*1a, *Ls* Gulen and *Ls* Tromsø) or wild caught (Norway, Maine) populations.

Genome size estimates of crustaceans based on FIAD are commonly lower than those based on FCM and estimates within 15% of one another are generally considered reliable⁵⁴. Accordingly, the FIAD derived estimates of 2.7 and 2.9 pg DNA per nucleus for adult females and males respectively, are within 10% to 15% lower than the FCM based estimates, depending upon the particular comparison. The most likely explanation for the disparity between FCM and FIAD based estimates is that the laser detection of cells in a suspension used in FCM is less sensitive to background noise, DNA compaction and other conformational changes in the chromatin sometimes encountered when measurements are based on the quantitative intercalation of the Schiff reagent among nucleotides as they sometimes are in squashed tissues in FIAD. These differences between FCM and FIAD based estimates of adults correspond to approximately 0.3 pg in a 2C nucleus, or \approx 150 million base pairs in the 1C genome. We conclude that measurements obtained from FIAD and FCM were internally consistent and that the discrepancy between the results are well within the boundaries expected from earlier studies⁵⁷. Since cytometric measurements are based on direct observations and the derived estimates are both consistent within the methods and discrepancies in accordance with methodological expectations we regard the cytometric results, collectively indicating a *L. salmonis salmonis* genome size of approximately 1.5 Gbps, as the most reliable measurements available.

The above suggests that sequence-based methods underestimate the genome size by approximately 33%. The mapping approach is sensitive to errors in assembly completeness, uniformity of library and sequencing coverage, mapping accuracy and modal mapping estimation. Alas, our analysis did not reveal which of these factors were the more likely to cause the apparent mapping-based underestimation of genome size. As an alternative to the mapping-based estimates we applied the widely used k-mer approach which similarly appeared to miss approximately 33% of the genome size. The most plausible explanation may be that repetitive elements cause the k-mer approach to underestimate genome size as previously observed by Pflug and co-workers⁵⁸. Similarly, genome assembly based on k-mer analysis of the lobster *Homarus americanus* is believed to be missing approximately 28% of the genome⁵⁹. The salmon louse genome is among the crustaceans with highest occurrence of repetitive elements; \approx 60% of the assembly annotated as repeats⁴¹, suggesting that such underestimation of size may not be implausible. In the present study sequence-based genome size estimates consistently provided substantially lower genome size estimates than cytometric measurements. Hence estimates should be regarded with caution until they have been confirmed by direct cytological measurements, as conducted here.

While our cytometric based estimates converge on a genome size of approximately 1.5 Gbps, earlier cytometric measurements disagree: a 1C = 0.58 pg DNA per cell (567 Mbp) estimate by Gregory⁵³ was based on material of Norwegian *L. salmonis salmonis* from a discontinued lab strain supplied by Professor Frank Nilsen of the University of Bergen in the early 2000's (pers. comm. Frank Nilsen) and a later 1C = 0.96 pg DNA per cell (939 Mbp) estimate by Jeffery⁵⁴ which was based on material collected in the Bay of Fundy in the 2010's and supplied by Professor Elizabeth Boulding from the University of Guelph. The most parsimonious explanation of

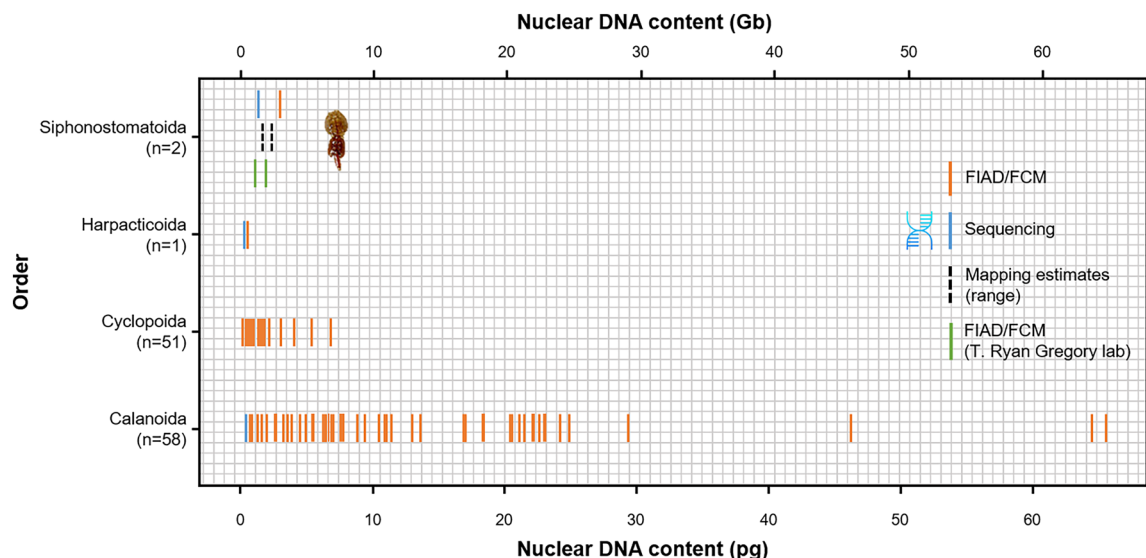


Figure 3. Somatic nuclear DNA contents of copepods. Each species is represented by a vertical bar. Estimates based on cytometric methods, genome assemblies, and mapping are presented in orange, blue, and dotted black, respectively, for the siphonostomatoid *L. salmonis salmonis*. The black dotted lines depict a range of values obtained for *L. salmonis salmonis* using mapping. FIAD and FCM measurements depicted by green vertical lines for *L. salmonis salmonis* are from Gregory⁵³ and Jeffery⁵⁴, respectively. FIAD (orange) and genome assembly (blue) estimates are shown for the harpacticoid *T. californicus*^{63,64}. FIAD (orange) and genome assembly (blue) estimates for the calanoid *E. affinis* are the two left most vertical lines in the distribution of Calanoida^{65,66}. The adult somatic genome size estimates of Cyclopoida and Calanoida (orange) were mostly acquired using FIAD and are from Gregory⁵³. All estimates are presented as 2C values found in the soma because some cyclopoid species and at least one calanoid species possess embryonic chromatin diminution, in which germline genome sizes (conventionally noted as 1C) substantially exceed the sizes of somatic values, thus violating the appropriateness of halving 2C estimates in the soma to obtain 1C estimates of the germline.

the discrepancy is to consider the earlier measurements erroneous. However, the measurements were made by respected authorities in the field, with whom our measurements on other species have agreed, and the salmon lice were supplied by field experts. We therefore believe their measurements are likely to be correct. Our results show that the DNA content of somatic cells is twice that of gametes; therefore, somatic chromatin diminution, found in some copepods, is an unlikely explanation for the observed differences. We therefore consider the discrepancy in results to indicate that large variations in the salmon louse genome size occasionally arise. Such variability would not be unprecedented in copepods and we suggest that this is addressed in *L. salmonis* by cytological measurements of genome sizes of both established strains and wild specimens covering their entire geographical range. Based on FCM geographically based intraspecific variations in genome size of magnitudes 1–9 pg (corresponding to a difference of ≈ 0.5 to 4.5 Gbp in the haploid genome) have been reported in the marine calanoid copepods *Calanus glacialis*, *Calanus hyperboreus*, and *Paraeuchaeta norvegica* populations inhabiting the High Arctic and Southern fjords of Norway⁶. Based on FIAD a difference of 1 pg between German (Schöhsee “house” lake of current Max Planck Institute for Evolutionary Biology in Ploen) and Lake Baikal populations of the freshwater cyclopoid *Mesocyclops leuckarti* was reported^{60,61}. Furthermore, the genome size of the North Sea population of the calanoid *Pseudocalanus elongatus* decreased after being reared in the laboratory for 96 generations⁶². It would not be surprising to encounter additional examples with intraspecific genome size differences in other copepods.

The somatic 2C nuclear DNA contents of *L. salmonis*, as estimated using FCM and FIAD in laboratory and wild populations (2.7–3.2 pg DNA per nucleus and corresponding to genome size of 1.33–1.56 Gbps) are at the lower end of the range of all published values of free-living copepods, which vary more than 300-fold from 0.20 to 64.46 pg DNA per nucleus, corresponding to genome sizes ranging from 195 Mbps to 63 Gbps (Fig. 3). Relative to the range of values in free-living cyclopoids, *L. salmonis salmonis* has an intermediate genome size, although its genome size is larger than the majority of cyclopoid species estimates. Relative to the range of cytometric based values in calanoids, *L. salmonis salmonis* is comparable to the smaller genomes, with the majority of calanoid species possessing larger or far larger genomes.

We speculate that the high abundance of repetitive regions in the *L. salmonis salmonis* genome facilitates the observed variability in genome size by serving as a “size accordion” where the repetitive elements may increase or decrease in copy numbers. This model has earlier been suggested for birds and mammals although the increase in these groups seems to be compensated by a corresponding loss of DNA segments elsewhere, resulting in rather constant genome sizes⁶⁷. We further suggest that a lower limit to the genome size likely exists, in a manner similar to what is indicated for the rotifer *Brachionus asplanchnoidis* in which genome size varies by a factor of 1.9⁶⁸. A consequence of a possible variability in genome size is that the mapping and k-mer based size estimates should not be considered conclusively discredited as they are based on sequence information from different wild samples or strains. For the Ls1a strain that was measured both by FIAD, FCM and included in the mapping and k-mer based analysis, the discrepancy may be assumed to be genuine, although the samples for cytometric

analyses were sampled more than 10 generations after the samples for sequencing were obtained. Hence the *L. salmonis salmonis* LSalAtl2s genome assembly⁴¹ appears to fail to resolve $\approx 50\%$ of the genome that is captured in the cytometric measurements. There are several possible sources that may contribute to this discrepancy: failure to isolate some regions of DNA, failure to capture all DNA in sequencing libraries or sequencing reactions, errors introduced during the bioinformatic analyses and specific challenges caused by TE's and DNA repeats.

The fact that the six independent genome assemblies are congruent in content despite originating from DNA purified from multiple origins of *L. salmonis* in different laboratories and sequenced using various sequencing platforms (Illumina, 454 pyrosequencing, Oxford nanopore, PacBio and Sanger) may suggest that no parts of the genome are systematically missed by the purification and sequencing protocols applied. However, systematic omission of genetic regions across isolation protocols or biased representations of repetitive regions cannot be excluded and would yield biased or incomplete genome representations. A similar distorting effect may be introduced during downstream bioinformatic analysis, for instance, by collapsing repetitive regions⁴⁴. Distinguishing between a bias in sequencing representation and bioinformatic artifacts as the cause of underestimating repetitive regions is challenging. PacBio and Nanopore sequencing platforms produce long reads that are commonly suggested as a tool to address repetitive regions. However, the two most recent *L. salmonis salmonis* assemblies were produced using PacBio (UStir_LSAA, Table 1) and Oxford nanopore combined with Illumina (UVic_Lsal_1.0, Table 1) sequencing. These did not deviate significantly in content or size from earlier assemblies and hence did not resolve the question of the missing DNA. Since there are no indications that specific regions are missed in any of the assemblies (Table 2) it is possible that the majority of the ≈ 800 Mbp of "missing" DNA (cytometric based genome size minus genome assembly size) in the samples measured by FIAD and FCM in the present study is comprised of TEs and DNA repeats that are not accurately captured in the assemblies. The fact that the long read based assemblies do not, at least partly, resolve the challenge may indicate that mini- and microsatellites may be dominating the missing fraction since these are considered more prone to incorrect rendering by long read sequencing than TE's^{44,69}. The existing ≈ 700 Mbp salmon louse assemblies consist of $\approx 60\%$ repetitive regions⁴¹ indicating that ≈ 300 Mbp consists of non-repetitive regions. Hence, the ≈ 1500 Mbp genomes measured in the present study suggestively consist of $\approx 80\%$ (≈ 1200 Mbp) repetitive and/or other kinds of DNA that were otherwise uncaptured regions and $\approx 20\%$ (≈ 300 Mbp) non-repetitive regions.

Challenges in capturing repeated regions are likely exacerbated in mid to large size genomes. At the lower range of genome sizes in copepods are the tidepool harpacticoid copepod *Tigriopus californicus* and the estuarine calanoid copepod *Eurytemora affinis* whose 2C genome sizes estimated by FIAD are 0.5 and 0.6–0.7 pg DNA per nucleus, respectively^{63,65}. Both genome assemblies were significantly smaller by $\approx 20\%$ (≈ 400 Mb for *T. californicus*, ≈ 495 Mb for *E. affinis* for 2C values) which attributed to the inability to sequence all of the repetitive DNA^{64,66} (Fig. 3). While *L. salmonis salmonis* has a genome size at the lower range of the distribution of copepods (Fig. 3), it is still two- to three-fold larger than *T. californicus* and *E. affinis*. A significant proportion of repetitive regions of the salmon louse genome consists of transposable elements, or of unclassified repeated motifs that may in time be annotated as TEs⁴¹. Precisely identifying the portion of the genome that is comprised of TEs and their composition is of interest as TEs are increasingly viewed as drivers of genome plasticity that facilitate the rise of new phenotypes, such as acquiring insecticide resistance in fruitflies^{70,71}. It may be speculated that the high occurrence of repeated regions, including TEs, in the salmon louse genome may have contributed to its documented ability to develop resistance towards new medicinal treatments despite a low diversity of genes typically associated with detoxification and stress response^{37,41,72–74}. If this is the case, the use of medicines in salmon farms that harbor the majority of sea lice in parts of the North Atlantic may have positively selected for high numbers of TEs and hence for a larger genome size. More evidence to support this hypothesis may be provided by evaluating the DNA repeat content in specimens of historical populations that existed prior to the introduction of drugs in aquaculture or present-day populations with little or no known exposure to such drugs. The effect of TEs on drug resistance in *L. salmonis* is an important avenue of study that merits further attention.

Materials and methods

Assembly analyses and sequencing-based genome size estimates. To evaluate the congruence of the six assemblies available in public databases (Table 1), while not requiring them to conserve synteny, we made a script that converted the individual assemblies into non-overlapping 240 bp synthetic reads thus generating 6 synthetic read libraries. These synthetic reads were then mapped to each of the published assemblies using BLAST with the following command line: `blastn -num_threads 16 -evalue 1e-10 -outfmt 6 -num_alignments 10 -penalty -1 -reward 1 -gapopen 3 -gapextend 2`. Since the same fragment may be mapped multiple times we calculated the percentage of synthetic reads that mapped and the fraction of the mapped reads that mapped with $> 95\%$ identity (Table 2).

The genome size was estimated from sequencing and assembly data using three different approaches: modal mapping extrapolation, k-mer analysis and single copy gene mapping extrapolation. The assembly based estimates were derived using the LSalAtl2s assembly⁴¹. The LSalAtl2s assembly was compared to other available assemblies (Table 1) to reveal potential regions that are missing.

Modal mapping extrapolation is based on the assumptions that populations of non-repetitive DNA sequence reads follow a Poisson distribution and this makes up the majority of DNA. By finding the modal coverage and dividing the total number of sequenced bases mapped by this number, we can estimate the genome size. This is done using the Lande-Waterman formula; $G = NL/C$, where G is the genome size, N is the number of reads, L is the average read length, and C is the modal coverage. The modal coverage was determined by plotting the number of sites against nucleotide coverage and identifying the peak value. To facilitate this, sequence reads from the GLW4, GLW13 and GLW16 libraries⁴¹ previously published were mapped against the LSalAtl2S genome assembly using Samtools⁵².

K-mer analysis is based on the assumption that the possible words of a certain size in a genome (k-mers) increase with the size of the genome. A genome consisting primarily of non-repetitive DNA regions will generate an approximately random population of k-mers, and the diversity of k-mers in a population of reads can be used to estimate the genome size. In the present study k-mer analyses were performed using Jellyfish⁵¹ and sequence reads from the GLW4 library derived from a laboratory strain of *L. salmonis salmonis*⁴¹ and libraries derived from *L. salmonis salmonis* collected in the field.

Nuclear DNA content analysis by flow-cytometry (FCM). *Field and laboratory populations.* Specimens of *L. salmonis salmonis* were obtained from several sources: (1) Wild adult males and females were collected from naturally infected farmed Atlantic salmon held at the sea cage facilities of the Aquaculture Research Station in Tromsø (FCM Run 5); (2) An outbred laboratory strain, *Ls Gulen*, was derived from adults collected in *Ls Gulen* (Norway) and reared at the Salmon Louse Research Centre in Bergen (FCM Run 4); (3) The *Ls Tromsø* laboratory strain was established by crossing adults from the *Ls Gulen* strain with a partially outbred strain, *Ls Oslofjord*, originating from specimens collected in Oslofjord (Norway) and reared at the Aquaculture Research Station in Tromsø (FCM runs 1–3).

Collection of samples and tissue preparation. Newly hatched nauplii were obtained from gravid females, crushed in cold citrate buffer⁷⁵ containing 5% dimethyl sulfoxide (DMSO), filtered through a 30 µm nylon mesh and deep-frozen until use. Sperm and eggs were collected from the testes and the genital segment prior to fertilization, respectively, and the resulting samples briefly kept on ice prior to analysis. Somatic (cuticular and subcuticular) tissues obtained from the cephalothorax (Supplemental Materials and Methods Fig. S1) of adult wild or laboratory specimens were crushed, and treated in the same way as the newly hatched nauplii. Specimens were squashed onto slides according to Clower and co-workers⁵⁵ except that a freeze-cracking technique was added.

Flow cytometry analysis. Aliquots of target (sea lice) and internal reference (male human and/or chicken) cells were analyzed using Propidium Iodide (PI) as fluorescent stain following previously reported methods⁷⁶. The mean DNA content of 5000–10,000 cells per sample was measured with a CyFlow®Ploidy Analyser equipped with a green laser.

Nuclear DNA contents of target species were estimated in relation to an assigned 2C value of 7.00 pgDNA/nucleus for human leukocytes and 2.50 pg DNA/nucleus for chicken erythrocytes⁵³ according to the formula:

$$\text{Target species nuclear DNA content (pg)} = (\text{Mean FL value of the sample} \times \text{reference 2CDNA content}) \times (\text{Mean FL value of the reference})^{-1}$$

Feulgen image analysis densitometry. *Field and laboratory populations.* Genome size measurements were obtained from each of three adults (females) of the *Ls1a* laboratory strain described elsewhere⁷⁷, whose ovaries served as the source material of DNA used in the nanopore DNA sequencing and six adults (three males and three females) of *L. salmonis salmonis* from *Ls Gulen* laboratory strain used in the FCM studies. A single adult female was collected from the wild in Copscook Bay, Maine in 2018. Specimens were immediately preserved in undenatured >95% alcohol.

Feulgen staining and scanning microdensitometry. All slides were squashed and stained with Schiff reagent according to previously reported methods^{46,55}, with few modifications. Nuclei were measured using a Zeiss Axio-cope A1 equipped with a 63X oil objective and a Qimaging Bioquant PVI CCD camera. Scanning microdensitometric software (Bioquant Image Analysis; Bioquant Life Sciences 2018 program) was used to determine the IODs of the nuclear DNA contents of individual somatic nuclei. We selected for measurement only nuclei that possessed a granular and slightly diffuse appearance and lacked visible pink background; these nuclei were found mostly at the perimeter or outside the carapace (Fig. 2c,d). Nuclei with relatively small areas and dense staining indicating DNA compaction (Fig. 2e) or very diffuse and large areas (Fig. 2f) are less likely to provide accurate measurements. The Bioquant software used to measure IODs has a conservative estimate of resolution of 0.5 pg DNA per nucleus according to the manufacturer. The mean IOD value of the hen was used to convert the IODs of each *L. salmonis salmonis* specimen to picograms, using the following equation:

$$\text{Pg}_c = (\text{pg}_s/\text{IOD}_s) \times \text{IOD}_c$$

where pg_c is the unknown amount of pg DNA per nucleus of *L. salmonis salmonis*, pg_s , 2.5 pg is the amount of DNA in the standard hen nucleus, IOD_s is the average IOD value of the hen, and IOD_c is the IOD value of *L. salmonis*. Photographs were taken at 100X magnification with a Nikon Eclipse Ti-2 microscope equipped with a PlanApo objective (N.A. 1.45) and QImaging DS RI2 camera.

Reference standards for conversion of integrated optical density (IOD) units to picograms (pg) included mutant white eyed female *Drosophila melanogaster* (0.40 pg DNA per nucleus), erythrocytes of hen *Gallus domesticus* (2.5 pg DNA per nucleus) and trout *Oncorhynchus mykiss* (5.2 pg DNA per nucleus), and leucocytes of male human *Homo sapiens* (7.0 pg DNA per nucleus) whose values were based on previous works^{78,79} and the Animal Genome Size Database⁵³. The calibration curve computed for standards in the staining batch containing the *Ls1a* strain yielded an $R^2 = 0.997$ (Supplementary Methods Fig. S2), indicating quantitative staining over a range of 0.40–7.0 pg DNA per nucleus. Only hen and trout standards (Fig. 2a,b) were used in the staining batch with *Ls Gulen* and Maine specimens.

The mean nuclear DNA contents are reported as 2C values in picograms (pg) and converted to gigabase (Gb) pairs (1 pg DNA = 0.978 Gb) for both FCM and FIAD derived estimates⁸⁰.

Statistical analyses. Differences in nuclear DNA content of somatic tissues between *LsTromsø* nauplii (FCM Run 1 and 2) and *Ls Gulen* germinal (eggs and sperm) or somatic tissues of *LsTromsø* and *Ls Gulen* adult males and females (FCM Run 3–4), as well as those of *Ls Gulen* adults obtained using FIAD, were analyzed by Students *t*-test. Analysis of variance (ANOVA) was used to detect significant differences in nuclear DNA content of somatic tissue of adult wild caught and laboratory *LsTromsø* strain *L. salmons salmons* (FCM Run 5) using fluorescence (FL) PI values as dependent variable and gender and strain (laboratory or wild) as factors. In Exploratory Data Analysis (EDA), Grubbs' test was used to detect presence of outliers and Levene's and Shapiro–Wilk tests were used to test homogeneity of variances among groups. Statistical analyses were performed using IBM SPSS Statistics v.25 software. Differences were accepted as significant when $P < 0.05$. Data are reported as mean \pm standard error (SE). The Shapiro–Wilk's test applied to nuclei within each of 10 specimens measured using FIAD revealed no departures from normality. Differences between male and female genome sizes based on FIAD were tested using two sample, two tailed Student *t*-tests.

Ethical standards. The study has been planned implemented and its results reported in accordance with ARRIVE guidelines (<https://arriveguidelines.org>). Sampling of parasites from infected fish were carried out in facilities approved by the Norwegian Food Safety Authority (Mattilsynet, FOTS). Fish were handled according to the Norwegian regulations for use of fish as laboratory animals (Norwegian Animal Research Authority) and all operations performed by approved personnel at the Tromsø Aquaculture Research Station (FOTS license nr. 110) and at the Institute of Marine Research in Bergen (permit nr. 2009/186329). The fresh chicken blood samples were provided by a professional veterinarian (D.C.R. Da Rocha Marques, Tromsø, Norway) and the human blood samples (isolated MNCs of anonymous donors) by the University Hospital of North Norway (UNN, Tromsø, Norway). The latter experimental work was done in accordance with relevant national guidelines/regulations and approved by Helse Nord RHF (<https://helse-nord.no/> Tromsø, Norway) via a contract for 'Use of blood donor blood for purposes other than patient care' stipulated between the Blood Bank/UNN and S. Peruzzi at UiT (contract nr. SJ1398/V2 of 18 November 2020). The experiments conducted at James Madison University were carried out in accordance with the National Science Foundations' regulations for use of animals in experiments. *Drosophila* tissues and blood of chicken, trout and human were provided by research staff of James Madison University These standard preparations were made on a single day in the early 1990's, stored in the dark at room temperature, and used in all subsequent staining procedures in GAW's laboratory.

Data availability

All relevant data are within the paper and its Supporting Information files. Details on the *LSalAtl2s* scaffold-level assembly are available at Ensembl Metazoa (https://metazoa.ensembl.org/Lepeophtheirus_salmonis).

Received: 7 July 2021; Accepted: 8 April 2022

Published online: 22 April 2022

References

- McClintock, B. The significance of responses of the genome to challenge. *Science* **226**, 792–801. <https://doi.org/10.1126/science.15739260> (1984).
- Gregory, T.R. *The Evolution of the Genome* (ed. Gregory, T.) 740 pp. (Elsevier, Academic Press, 2005).
- Markov, A. V., Anisimov, V. A. & Korotayev, A. V. Relationship between genome size and organismal complexity in the lineage leading from prokaryotes to mammals. *Paleontol. J.* **44**, 363–373 (2010).
- Choi, I.-Y., Kwon, E.-C. & Kim, N.S. The C-and G-value paradox with polyploidy repeatomes, introns, phenomes, and cell economy. *Genes Genomics* **42**, 699–714 (2020).
- Wyngaard, G. A., Rasch, E. M., Manning, N. M., Gasser, K. & Domangue, R. The relationship between genome size, development rate, and body size in copepods. *Hydrobiologia* **532**, 123–137 (2005).
- Leinaas, H. P., Jalal, M., Gabrielsen, T. M. & Hessen, D. O. Inter- and intraspecific variation in body- and genome size in calanoid copepods from temperate and arctic waters. *Ecol. Evol.* **6**, 5585–5595 (2016).
- Hultgren, K. M., Jeffery, N. W., Moran, A. & Gregory, T. R. Latitudinal variation in genome size in crustaceans. *Biol. J. Linn. Soc.* **123**, 348–359 (2018).
- Gregory, T. R. Coincidence, coevolution or causation? DNA content, cell size, and the C-value enigma. *Biol. Rev.* **76**, 65–101 (2001).
- McLaren, I. A. & Marcogliese, D. J. Similar nuclear numbers among copepods. *Can. J. Zool.* **61**, 721–724 (1983).
- Elliott, T.A. & Gregory, T.R. What is in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Phil. Trans. R. Soc. B.* **370**, 20140441. <https://doi.org/10.1098/rstb.2014.0331> (2015).
- Bennett, M.D. & Leitch, I.J. Genome size evolution in plants. In: Gregory, T.R. *The Evolution of the Genome* (ed. Gregory, T.), 89–162 (Elsevier, Academic Press, 2005).
- Hessen, D. O., Daufresne, M. & Leinaas, H. P. Temperature-size relations from the cellular-genomic perspective. *Biol. Rev.* **88**, 476–489 (2013).
- Ivancevic, A.M., Kortschak, R.D., Bertozzi, T. & Adelson, D.L. Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. *Genome Biol.* **19**. <https://doi.org/10.1186/s13059-018-1456-7> (2018).
- Bourque, G. *et al.* Ten things you should know about transposable elements. *Genome Biol.* **19**, 199. <https://doi.org/10.1186/s13059-018-1577-z> (2018).
- Eickbush, T.H. & Furano, A.V. Fruit flies and humans respond differently to retrotransposons. *Curr. Opin. Genet. Dev.* **12**, 669–674. [https://doi.org/10.1016/s0959-437x\(02\)00359-3](https://doi.org/10.1016/s0959-437x(02)00359-3) (2002).
- Rech, G.E. *et al.* Stress response, behavior, and development are shaped by transposable element-induced mutations in *Drosophila*. *PLoS Genet.* **15**, e1007900. <https://doi.org/10.1371/journal.pgen.1007900> (2019).
- Kapun, M. *et al.* Genomic analysis of European *Drosophila melanogaster* populations reveals longitudinal structure, continent-wide selection, and previously unknown DNA viruses. *Mol. Biol. Evol.* **37**, 2661–2678. <https://doi.org/10.1093/conphys/coz072> (2020).

18. Hessen, D. O., Ventura, M. & Else, J. J. Do phosphorus requirements for RNA limit genome size in crustacean zooplankton? *Genome* **51**, 685–691 (2008).
19. Bullejos, F.J., Carrillo, P., Gorokhova, E., Medina-Sanchez, J.M. & Villar-Argaiz, M. Nucleic acid content in crustacean zooplankton: Bridging metabolic and stoichiometric predictions. *PLoS ONE* **9**, e86493. <https://doi.org/10.1371/journal.pone.0086493> (2014).
20. Torrissen, O. *et al.* Salmon lice—impact on wild salmonids and salmon aquaculture. *J. Fish Dis.* **36**, 171–194 (2013).
21. Vollset, K. W. *et al.* Disentangling the role of sea lice on the marine survival of Atlantic salmon. *ICES J. Mar. Sci.* **75**, 50–60 (2018).
22. Skern-Mauritzen, R., Torrissen, O. & Glover, K.A. Pacific and Atlantic *Lepeophtheirus salmonis* (Kroyer, 1838) are allopatric subspecies: *Lepeophtheirus salmonis salmonis* and *L. salmonis oncorhynchi* subspecies novo. *BMC Genet.* **15**, 32. <https://doi.org/10.1186/1471-2156-15-32> (2014).
23. Marincovich, L. & Gladenkov, A. Y. Evidence for an early opening of the Bering Strait. *Nature* **397**, 149–151 (1999).
24. Yazawa, R. *et al.* EST and mitochondrial DNA sequences support a distinct Pacific form of salmon louse *Lepeophtheirus salmonis*. *Mar. Biotechnol.* **10**, 741–749 (2008).
25. Bui, S., Oppedal, F., Stien, L. & Dempster, T. Sea lice infestation level alters salmon swimming depth in sea-cages. *Aquacult. Environ. Interact.* **8**, 429–435 (2016).
26. Fjellidal, P.G., Hansen, T.J., Karlsen, Ø. & Wright D.W. Effects of laboratory salmon louse infection on Arctic char osmoregulation, growth and survival. *Conserv. Physiol.* **7**, coz072. <https://doi.org/10.1093/conphys/coz072> (2019).
27. Barker, S.E. *et al.* Sea lice, *Lepeophtheirus salmonis* (Krøyer 1837), infected Atlantic salmon (*Salmo salar* L.) are more susceptible to infectious salmon anemia virus. *PLoS ONE* **14**, e0213232. <https://doi.org/10.1371/journal.pone.0209178> (2019).
28. Brooker, A. J., Skern-Mauritzen, R. & Bron, J. E. Production, mortality, and infectivity of planktonic larval sea lice, *Lepeophtheirus salmonis* (Kroyer, 1837): Current knowledge and implications for epidemiological modelling. *ICES J. Mar. Sci.* **75**, 1214–1234 (2018).
29. Forseth, T. *et al.* The major threats to Atlantic salmon in Norway. *ICES J. Mar. Sci.* **74**, 1496–1513 (2017).
30. Murray, A.G. & Moriarty, M. A simple modelling tool for assessing interaction with host and local infestation of sea lice from salmonid farms on wild salmonids based on processes operating at multiple scales in space and time. *Ecol. Model.* **443**. <https://doi.org/10.1016/j.ecolmodel.2021.109459> (2020).
31. Sandvik, A.D. *et al.* The development of a sustainability assessment indicator and its response to management changes as derived from salmon lice dispersal modelling. *ICES J. Mar. Sci.*, fsab077. <https://doi.org/10.1093/icesjms/fsab077> (2021).
32. Thomson, C. R. *et al.* Illuminating the planktonic stages of salmon lice: A unique fluorescence signal for rapid identification of a rare copepod in zooplankton assemblages. *J. Fish Dis.* **44**, 863–879 (2021).
33. Glover, K. A. *et al.* Population genetic structure of the parasitic copepod *Lepeophtheirus salmonis* throughout the Atlantic. *Mar. Ecol. Prog. Ser.* **427**, 161–172 (2011).
34. Besnier, F. *et al.* Human-induced evolution caught in action: SNP-array reveals rapid amphi-atlantic spread of pesticide resistance in the salmon ectoparasite *Lepeophtheirus salmonis*. *BMC Genomics* **15**, 937. <https://doi.org/10.1186/1471-2164-15-937> (2014).
35. Fjortoft, H. B. *et al.* Salmon lice sampled from wild Atlantic salmon and sea trout throughout Norway display high frequencies of the genotype associated with pyrethroid resistance. *Aquacult. Environ. Interact.* **11**, 459–468 (2019).
36. Helgesen, K. O., Romstad, H., Aaen, S. M. & Horsberg, T. E. First report of reduced sensitivity towards hydrogen peroxide found in the salmon louse *Lepeophtheirus salmonis* in Norway. *Aquacult. Rep.* **1**, 37–42 (2015).
37. Kaur, K. *et al.* The mechanism (Phe362Tyr mutation) behind resistance in *Lepeophtheirus salmonis* pre-dates organophosphate use in salmon farming. *Sci. Rep.* **7**(1), 12349. <https://doi.org/10.1038/s41598-017-12384-6> (2017).
38. Skern-Mauritzen, R., Frost, P., Hamre, L.A., Kongshaug, H. & Nilsen, F. Molecular characterization and classification of a clip domain containing peptidase from the ectoparasite *Lepeophtheirus salmonis* (Copepoda, Crustacea). *Comp. Biochem. Physiol. B Biochem Mol. Biol.* **146**(2), 289–298 (2007).
39. Eichner, C. *et al.* Characterization of a novel RXR receptor in the salmon louse (*Lepeophtheirus salmonis*, Copepoda) regulating growth and female reproduction. *BMC Genomics* **16**, 81. <https://doi.org/10.1186/s12864-015-1277-y> (2015).
40. Øvergård, A. C. *et al.* Exocrine glands of *Lepeophtheirus salmonis* (Copepoda: Caligidae): Distribution, developmental appearance, and site of secretion. *J. Morphol.* **277**, 1616–1630. <https://doi.org/10.1002/jmor.20611> (2016).
41. Skern-Mauritzen, R. *et al.* The salmon louse genome: Copepod features and parasitic adaptations. *Genomics* **113**, 3666–3680 (2021).
42. Messmer, A. M. *et al.* A 200K SNP chip reveals a novel Pacific salmon louse genotype linked to differential efficacy of emamectin benzoate. *Mar. Genom.* **40**, 45–57 (2018).
43. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing, computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–44 (2012).
44. Tørresen, O. K. *et al.* Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Res.* **47**, 10994–11006 (2019).
45. Rasch, E. M. Feulgen-DNA cytophotometry for estimating C values. *Methods Mol. Biol.* **247**, 163–201 (2004).
46. Hardie, D. C., Gregory, T. R. & Hebert, P. D. N. From pixels to picograms: A beginner's guide to genome quantification by Feulgen image analysis densitometry. *J. Histochem. Cytochem.* **50**, 725–749 (2002).
47. Danzmann, R. G. *et al.* A genetic linkage map for the salmon louse (*Lepeophtheirus salmonis*): Evidence for high male:female and inter-familial recombination rate differences. *Mol. Genet. Genom.* **294**, 343–363 (2019).
48. Eichner, C., Dondrup, M. & Nilsen, F. RNA sequencing reveals distinct gene expression patterns during the development of parasitic larval stages of the salmon louse (*Lepeophtheirus salmonis*). *J. Fish Dis.* **41**, 1005–1029. <https://doi.org/10.1111/jfd.12770> (2018).
49. Heggland, E. I. *et al.* A scavenger receptor B (CD36)-like protein is a potential mediator of intestinal heme absorption in the hematophagous ectoparasite *Lepeophtheirus salmonis*. *Sci. Rep.* **9**, 4218. <https://doi.org/10.1038/s41598-019-40590-x> (2019).
50. Wellcome Trust. Sharing data from large-scale biological research projects: A system of tripartite responsibility. https://wellcome.org/sites/default/files/wtd003207_0.pdf (2003).
51. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**(6), 764–770 (2011).
52. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324> (2009).
53. Gregory, T.R. *Animal Genome Size Database*. <http://www.genomesize.com> (2021).
54. Jeffery, N.W. Genome size diversity and evolution in the Crustacea. *Ph.D Thesis, University of Guelph*; 257 pp. <https://atrium.lib.uoguelph.ca/xmlui/handle/10214/9216> (2015).
55. Clower, M.K., Holub, A.S., Smith, R.T. & Wyngaard, G.A. Embryonic development and a quantitative model of programmed DNA elimination in *Mesocyclops edax* (S.A. Forbes, 1891) (Copepoda: Cyclopoida). *J. Crustac. Biol.* **36**, 661–674 (2016).
56. Bachtrog, D. A dynamic view of sex chromosome evolution. *Curr. Opin. Genet. Dev.* **16**, 578–585 (2006).
57. Jeffery, N. W., Jardine, C. B. & Gregory, T. R. A first exploration of genome size diversity in sponges. *Genome* **56**, 451–456 (2013).
58. Pflug, J.M., Holmes, V.R., Burrus, C., Johnston, J.S. & Maddison, D.R. Measuring genome sizes using read-depth, k-mers, and flow cytometry: Methodological comparisons in beetles (Coleoptera). *G3 (Bethesda)* **10**(9), 3047–3060. <https://doi.org/10.1534/g3.120.401028> (2020).
59. Polinski, J. M. *et al.* The American lobster genome reveals insights on longevity, neural, and immune adaptations. *Sci. Adv.* **7**, eabe290. <https://doi.org/10.1126/sciadv.abe290> (2021).

60. Rasch, E. M. & Wyngaard, G. A. Genome sizes of cyclopoid copepods (Crustacea): Evidence of evolutionary constraint. *Biol. J. Linn. Soc.* **87**, 625–635 (2006).
61. Ivankina, E. A. *et al.* Cytophotometric determination of genome size in two species of cyclops of Lake Baikal (Crustacea: Copepoda, Cyclopoida) in ontogenetic development. *Cell Tissue Biol.* **7**, 192–199 (2013).
62. Escribano, R., McLaren, I. A. & Klein Breteler, W. C. M. Innate and acquired variation of nuclear DNA contents of marine copepods. *Genome* **35**, 602–610 (1992).
63. Wyngaard, G. A. & Rasch, E. M. Patterns of genome size in the copepoda. *Hydrobiologia* **417**, 43–56 (2000).
64. Barreto, F. S. *et al.* Genomic signatures of mitonuclear coevolution across populations of *Tigriopus californicus*. *Nat. Ecol. Evol.* **2**, 1250–1257 (2018).
65. Rasch, E. M., Lee, C. E. & Wyngaard, G. A. DNA-Feulgen cytophotometric determination of genome size for the freshwater invading copepod *Eurytemora affinis*. *Genome* **47**, 559–564 (2004).
66. Eyun, S. I. *et al.* Evolutionary history of chemosensory-related gene families across the Arthropoda. *Mol. Biol. Evol.* **34**, 1838–1862 (2017).
67. Kapusta, A., Suh, A. & Feschotte, A. Dynamics of genome size evolution in birds and mammals. *Proc. Natl. Acad. Sci. USA* **108**, E1460–E1469 (2017).
68. Stelzer, C.-P., Pichler, M., Stadler, P., Hatheuer, A. & Riss, S. Within-population genome size variation is mediated by multiple genomic elements that segregate independently during meiosis. *Genome Biol. Evol.* **11**, 3424–3435. <https://doi.org/10.1093/gbe/evz253> (2019).
69. Liljegen, M. M., de Muinck, E. J. & Trosvik, P. Microsatellite length scoring by Single molecule real time sequencing—Effects of sequence structure and PCR Regime. *PLoS ONE* **11**(7), e0159232. <https://doi.org/10.1371/journal.pone.0159232> (2016).
70. Pimpinelli, S. & Piacentini, L. Environmental change and the evolution of genomes: Transposable elements as translators of phenotypic plasticity into genotypic variability. *Funct. Ecol.* **34**, 428–441. <https://doi.org/10.1111/1365-2435.13497> (2020).
71. Chung, H. *et al.* Cis-regulatory elements in the Accord retrotransposon result in tissue-specific expression of the *Drosophila melanogaster* insecticide resistance gene Cyp6g1. *Genetics* **175**, 1071–1077. <https://doi.org/10.1534/genetics.106.066597> (2007).
72. Coates, A. *et al.* Evolution of salmon lice in response to management strategies: A review. *Rev. Aquacult.* **13**, 1397–1422. <https://doi.org/10.1111/raq.12528> (2021).
73. Besnier, F. *et al.* Identification of quantitative genetic components of fitness variation in farmed, hybrid and native salmon in the wild. *Heredity* **115**, 47–55. <https://doi.org/10.1038/hdy.2015.15> (2015).
74. Aaen, S. M., Helgesen, K. O., Bakke, M. J., Kaur, K. & Horsberg, T. E. Drug resistance in sea lice: A threat to salmonid aquaculture. *Trends Parasitol.* **31**, 72–81 (2015).
75. Vindeløv, L. L., Christensen, I. J. & Nissen, N. I. A detergent-trypsin method for the preparation of nuclei for flow cytometric DNA analysis. *Cytometry* **3**, 323–327 (1983).
76. Tiersch, T. R., Chandler, R. W., Kallman, K. & Wachtel, S. S. Estimation of nuclear DNA content by flow cytometry in fishes of the genus *Xiphophorus*. *Comp. Biochem. Physiol. B* **94**, 465–468 (1989).
77. Hamre, L. A., Glover, K. A. & Nilsen, F. Establishment and characterisation of salmon louse (*Lepeophtheirus salmonis* (Kroyer 1837)) laboratory strains. *Parasitol. Int.* **58**, 451–460 (2009).
78. Mulligan, P. K. & Rasch, E. M. The determination of genome size in male and female germ cells of *Drosophila melanogaster* by DNA-Feulgen cytophotometry. *Histochemistry* **66**, 11–18 (1980).
79. Rasch, E. M. DNA “standards” and the range of accurate DNA estimates by Feulgen absorption microspectrophotometry. In: *Advances in Microscopy, Progress in Clinical and Biological Research* (eds. Cowden, R.R. & Harrison, S.H.), **196**, 137–166 (Alan R. Liss, Inc., 1985).
80. Doležal, J., Bartoš, J., Voglmayr, H. & Greilhuber, J. Nuclear DNA and genome size of trout and human. *Cytometry* **51**, 127–128 (2003).

Acknowledgements

We thank James Madison University (JMU) Physics Department for constructing the liquid nitrogen table, Adrian Streit for methodological advice, Kris Kubow for imaging support and preparation of plate, Harrison Giknovorian for technical assistance with squashing and Feulgen reaction, Emily Schutt for preparation of the graph, Ken Roth for supplying the human blood, Marquis Walker for providing the *Drosophila* and methods for histological preparation, and Michaël Bekaert, Tyler Elliott, Ryan Gregory, Nick Jeffery, and Ben Koop for critical discussions. Work at JMU was supported by NIH 1R15GM104868 and NSF- DBI-1725855 to GAW and others, NSF-DEB 1948267 to GAW and a grant to GAW from the James Madison University Program of Grants for Faculty Assistance. The publication charges for this article have been funded by a grant to SP from the publication fund of UiT, The Arctic University of Norway. The authors are grateful to Anette Hustad, Linn Svendheim and Sverre Rune Hansen at the Aquaculture Research Station in Tromsø and to Sussie Dalvin (IMR, Bergen) for sea lice samples collection. We acknowledge the veterinarian Diogo Costa Ramos Da Rocha Marques for supplying the chicken blood and Goran Kauric at the University Hospital of North Norway (UNN, Tromsø) for processing the human blood samples.

Author contributions

G.A.W., R.S.M. and S.P.: conceptualized and designed the work. G.A.W. and S.P.: conducted cytogenetic (FIAD, Flow Cytometry) analyses and data handling. R.S.M. and K.M.: performed bioinformatic analyses. G.A.W., R.S.M., K.M. and S.P.: drafted and wrote the main manuscript. R.P. developed the freeze-cracking method for FIAD measurements. All authors reviewed the manuscript.

Funding

Open Access funding provided by UiT The Arctic University of Norway. This article was funded by National Institutes of Health (Grant no. 1R15GM104868), James Madison University, Program of Grants for Faculty Assistance, National Science Foundation (Grant no. DBI-1725855).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-10585-2>.

Correspondence and requests for materials should be addressed to S.P.

Reprints and permissions information is available at www.nature.com/reprints.

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



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

© The Author(s) 2022