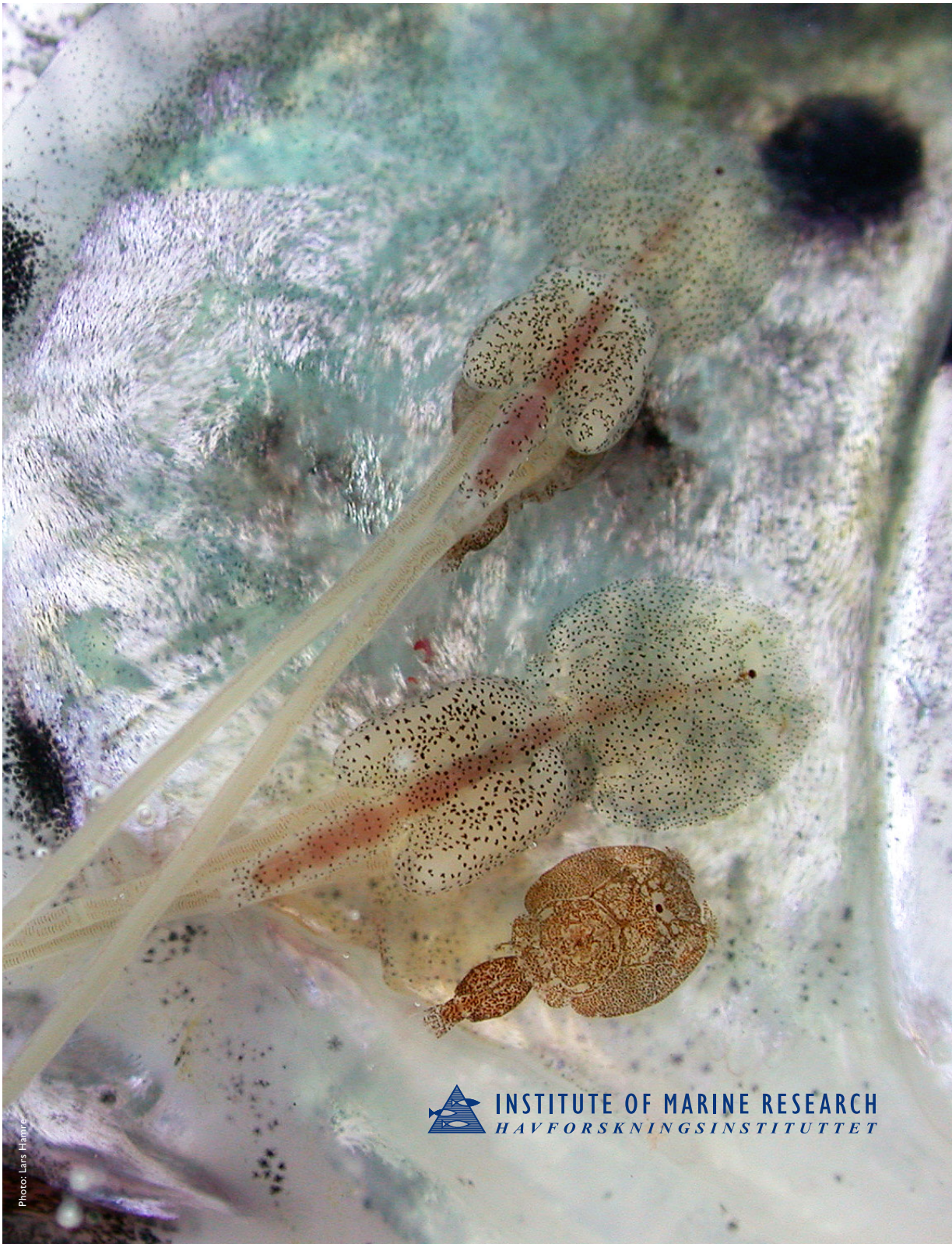
INSTITUTE OF MARINE RESEARCH REPORT
No. 11–2013

Genome sequencing Part 1
Final report, April 2013

Rasmus-Skern Mauritzen, Ketil Malde, Tomasz Furmanek, Frank Nilsen



www.imr.no

IMR-Project: 13291-06
FHF-Project: 900400

Genome sequencing Part 1
Final report, April 2013

Background:

Copepods comprise the largest animal marine biomass, but they have been poorly studied by molecular tools, and no genome has so far been sequenced. Copepods are common as parasites on fishes, and the salmon louse (*Lepeophtheirus salmonis*) is the most important species in terms of economical significance. Due to emerging resistance development, new tools (vaccine and new drugs) for lice control is needed. One of the factors limiting research supporting development of new treatments has been the lack of a complete genome sequence. Consequently a joint effort to sequence the salmon louse genome was embarked upon by Institute of Marine Research (IMR), University of Bergen (represented by the Salmon Louse Research Centre; SLRC), Marine Harvest (MH) and The Norwegian Seafood Research Fund (FHF) in November 2009. *The Salmon Louse Genome Sequencing Project* was divided in two parts: Genome sequencing (Part 1) and bioinformatic processing (Part 2). IMR, MH and FHF agreed to co-finance Part 1 of the project. The project is conducted in cooperation with Max Planck institute, Computational Biology Unit at University of Bergen and University of Victoria.

The salmon louse genome project part 1: *Sequencing*

Results part 1:

The salmon louse genome was sequenced to a final coverage of app. 180X as outlined in Table 1. The sequencing was completed in 2011 and formally completed part 1 of the salmon louse genome project.

Table 1: *Sequencing depth and sequencing details using the different sequencing approaches are shown. The fosmids were sequenced to a clonal genome coverage of app. 3.6X.*

Library	Runs	Platform	Coverage	Read length	Type	insert size
GLW4	4 lanes	Hiseq 2000	97,26	100	paired end	360 bp
GLW13	3 lanes	Hiseq 2000	60,64	100	paired end	500 bp
GLW16	1 lane	Hiseq 2000	7,64	50	mate pair	3-6 kb
454	22,5 plates	454 GS FLX	15,50	450	single read	Na
Fosmid	120000	sanger	0,23	app. 1100	paired end	32-42 kb

Financing of part 1:

Part 1 of the Salmon Louse Sequencing Project was financed by IMR, FHF and MH.

The salmon louse genome project part 2: *Bioinformatic analyses*

Results Part 2:

Initial assemblies were generated using a number of assemblers and various combinations of assemblers (since many assemblers will not accept all data) before scaffolding[1]. The utilization of the resulting genome assemblies and annotations were secured by making the results publicly available in accordance with an access agreement (Appendix A) from March 2012. This availability has resulted in use by 46 scientists from 12 institutions (for details see appendix B). A final assembly was made in January 2013 based on comparisons of assemblers and supporting information on linkage groups supplied by the PrevenT project. Among the tested assemblers (CLC, Newbler and Abyss) none were clearly superior, but one (CLC) produced significantly more errors than the others. The final assembly strategy was selected based on a balance between large scale correctness (fewest possible long distance errors in the assembly) and sequence comprehensiveness and correctness (most possible sequence reads mapping). For details refer to Table 2. The final assembly (generated by Newbler) appears to be of very good quality and has a N50 of 570K and a total size of 695MB.

Table 2: *Assembly statistics. N50 is based on an estimated genome size of 570Mbp. Bad contigs is the number of contigs that are members of >1 linkage groups. % DNA and RNA mapped is the fraction of the reads that mapped to the genome, and the value in parenthesis is the mapping quality. Transcripts are the number of transcripts from the transcriptome assembly that mapped to the genome. LSalAtl2 is the final assembly.*

Assembly	N50	Size	Bad contigs	Basepairs in bad contigs	% DNA mapped	% RNA mapped	Transcripts mapped
CLC	498803	708 820 047	166	139 318 375	88,4 (47)	97,7(35)	28406
Newbler	821670	645 897 683	35	50 996 996	88,5 (46)	96,9(47)	28282
Abyss	563753	723 201 199	32	22 292 126	85,8(42)	90,4(37)	28353
LSalAtl2	599458	684 655 938	35	36 259 186	89,5(46)	97,0(35)	28278

Reliable annotation of genomes requires gene predictions supported by data from sequencing expressed genes (RNA tags). 440 mill RNA tags equally distributed between all stages was consequently sequenced and used to train the gene prediction software. The final assembly contained app. 35K predicted genes (app. 25% of these with annotation), which is comparable to the number for *Daphnia pulex* (app. 31K genes (Colbourne et al., 2011)) and significantly higher than the number found in e.g. *Drosophila melanogaster* (app. 13.5K genes (Adams et al., 2000)).

Financing Part 2:

Part 2 of the Salmon Louse Genome Project was financed by IMR and SLRC.

[1] An assembler organizes raw sequence data into contiguous sequences (contigs). A scaffolder uses additional information from paired sequence reads to organize contigs relative to each other.

References:

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., George, R.A., Lewis, S.E., Richards, S., Ashburner, M., Henderson, S.N., Sutton, G.G., Wortman, J.R., Yandell, M.D., Zhang, Q., Chen, L.X., Brandon, R.C., Rogers, Y.H., Blazej, R.G., Champe, M., Pfeiffer, B.D., Wan, K.H., Doyle, C., Baxter, E.G., Helt, G., Nelson, C.R., Gabor, G.L., Abril, J.F., Agbayani, A., An, H.J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R.M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E.M., Beeson, K.Y., Benos, P.V., Berman, B.P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M.R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K.C., Busam, D.A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J.M., Cawley, S., Dahlke, C., Davenport, L.B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A.D., Dew, I., Dietz, S.M., Dodson, K., Doup, L.E., Downes, M., Dugan-Rocha, S., Dunkov, B.C., Dunn, P., Durbin, K.J., Evangelista, C.C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A.E., Garg, N.S., Gelbart, W.M., Glasser, K., Glodek, A., Gong, F., Gorrell, J.H., Gu, Z., Guan, P., Harris, M., Harris, N.L., Harvey, D., Heiman, T.J., Hernandez, J.R., Houck, J., Hostin, D., Houston, K.A., Howland, T.J., Wei, M.H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G.H., Ke, Z., Kennison, J.A., Ketchum, K.A., Kimmel, B.E., Kodira, C.D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A.A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T.C., McLeod, M.P., McPherson, D., Merkulov, G., Milshina, N.V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S.M., Moy, M., Murphy, B., Murphy, L., Muzny, D.M., Nelson, D.L., Nelson, D.R., Nelson, K.A., Nixon, K., Nusskern, D.R., Pacleb, J.M., Palazzolo, M., Pittman, G.S., Pan, S., Pollard, J., Puri, V., Reese, M.G., Reinert, K., Remington, K., Saunders, R.D., Scheeler, F., Shen, H., Shue, B.C., Siden-Kiamos, I., Simpson, M., Skupski, M.P., Smith, T., Spier, E., Spradling, A.C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A.H., Wang, X., Wang, Z.Y., Wassarman, D.A., Weinstock, G.M., Weissenbach, J., Williams, S.M., WoodageT, Worley, K.C., Wu, D., Yang, S., Yao, Q.A., Ye, J., Yeh, R.F., Zaveri, J.S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X.H., Zhong, F.N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H.O., Gibbs, R.A., Myers, E.W., Rubin, G.M., Venter, J.C., 2000. The genome sequence of Drosophila melanogaster. Science 287, 2185-2195.

Colbourne, J.K., Pfrender, M.E., Gilbert, D., Thomas, W.K., Tucker, A., Oakley, T.H., Tokishita, S., Aerts, A., Arnold, G.J., Basu, M.K., Bauer, D.J., Caceres, C.E., Carmel, L., Casola, C., Choi, J.H., Detter, J.C., Dong, Q., Dusheyko, S., Eads, B.D., Frohlich, T., Geiler-Samerotte, K.A., Gerlach, D., Hatcher, P., Jogdeo, S., Krijgsveld, J., Kriventseva, E.V., Kultz, D., Laforsch, C., Lindquist, E., Lopez, J., Manak, J.R., Muller, J., Pangilinan, J., Patwardhan, R.P., Pitluck, S., Pritham, E.J., Rechtsteiner, A., Rho, M., Rogozin, I.B., Sakarya, O., Salamov, A., Schaack, S., Shapiro, H., Shiga, Y., Skalitzky, C., Smith, Z., Souvorov, A., Sung, W., Tang, Z., Tsuchiya, D., Tu, H., Vos, H., Wang, M., Wolf, Y.I., Yamagata, H., Yamada, T., Ye, Y., Shaw, J.R., Andrews, J., Crease, T.J., Tang, H., Lucas, S.M., Robertson, H.M., Bork, P., Koonin, E.V., Zdobnov, E.M., Grigoriev, I.V., Lynch, M., Boore, J.L., 2011. The ecoresponsive genome of Daphnia pulex. Science 331, 555-561.

Appendix A.

Access agreement

The Salmon Louse Genome Sequencing Project is financed by Institute of Marine Research (IMR), The Sea Louse Research Centre (SLRC), Marine Harvest (MH) and The Norwegian Fisheries and Aquaculture Research Fund (FHF) and is executed by IMR, SLRC, University of Bergen (UoB), Max Planck institute (MPI) and Computational Biology Unit (CBU) in cooperation with University of Victoria. The generated sequence data, annotations etc., hereafter referred to as the results, are the property of the participating parties and are made accessible to project participants through http://sealouse.imr.no. The Salmon Louse Genome Sequencing Project participants will publish the genome in a joint paper as soon as possible and the sequence and annotations will be made freely available at that time. To facilitate research and development of salmon louse treatments the results will be made accessible to third parties prior to publication under the below listed conditions. It is emphasized that all information has not been validated and that granting access does not imply that The Salmon Louse Genome Sequencing Project is liable for any information posted or obliged to provide any support.

Conditions for access:

1. Access to the generated results is granted under the condition that the resource will be used to study genes and/or fragments of the genome only.
2. The results must not be forwarded in any form. The password issued is personal and may not be shared. The holder of access permission is responsible for implementing routines that prevent unauthorized access.
3. Prior to submission of manuscripts or any other form of publication based on information made available through http://sealouse.imr.no the material must be submitted to the project board (participating scientists from IMR, MPI, SLRC, UoB and CBU) for endorsement. This is done by sending the documents to the project leader R. Skern-Mauritzen (rasmus@imr.no).
4. Documents submitted to the academic board for endorsement are confidential.
5. The academic board reserves the right to deny submission or publication of material until publication of the genome. After publication of the genome no restrictions will apply.
6. Submitting documents to the academic board for endorsement does not imply any transfer of rights.
7. Patent applications are not regarded as a publications and notification when filing patent applications is not required. If required for a patent application, relevant information may be forwarded under the condition that use of the information must be related exclusively to the application.
8. Any dispute that may arise as a result of this access agreement is governed by and shall be interpreted in accordance with Norwegian law. Any disputes shall be settled at the court of the IMRs business address.

I hereby confirm that I have read and understood the conditions, and that I accept these.

_____	_____	_____
Signee	*Company/Institution*	*Date/Place*

Appendix B.

Persons with access to the salmon louse genome resources in accordance with the access agreement (Appendix A).

Name	**Institution**
Daniel John Macqueen	*University of St Andrews*
Aleksei Krasnov	*Nofima Marin*
Anna Zofia Komisarczuk	*Universitetet i Bergen*
Ben F Koop	*University of Victoria*
Bjørn Olav Kvamme	*Institute of Marine Research*
Christer Nilsen	*Norwegian Veterinary Institute*
Christiane Eichner	*Universitetet i Bergen*
Christiane Trösse	*Universitetet i Bergen*
Christoffer Marlowe A. Caipang	*Institute of Marine Research*
Craig Morton	*Institute of Marine Research*
Francois Besnier	*Institute of Marine Research*
Frank Nilsen	*Universitetet i Bergen*
Gunnvør Joensen	*Fiskaaling*
Heidi Kongshaug	*Universitetet i Bergen*
Inge Jonassen	*Universitetet i Bergen*
James Bron	*University of Stirling*
Jarle K Hopland	*Institute of Marine Research*
Jon Anders Stavang	*Universitetet i Bergen*
Jong Leong	*University of Victoria*
Jose de la fuente	*Universidad de la Castilla*
Kaur Kiranpreet	*Norwegian veterinary college*
Ketil Malde	*Institute of Marine Research*
Kurt Stueber	*Max Planck Institute*
Liv Sandlund	*Universitetet i Bergen*
Lucien Rufener	*Novartis*
Michael Bekaërt	*University of Stirling*
Michael Dondrup	*Universitetet i Bergen*
Muhammad Tanveer Khan	*Universitetet i Bergen*
Nigel R Finn	*Universitetet i Bergen*
Paul Kersey	*European Bioinformatics Institute*
Petter Frost	*Intervet*
Punit Bhattachan	*Universitetet i Bergen*
Rasmus Skern-Mauritzen	*Institute of Marine Research*
Remi-Andre Olsen	*Universitetet i Bergen*
Richard Reinhardt	*Max Planck Institute*
Rolf B Edvardsen	*Institute of Marine Research*
Rune Male	*Universitetet i Bergen*
Sindre Grotmol	*Universitetet i Bergen*
Stephen Carmichael	*University of Stirling*
Sussie T Dalvin	*Institute of Marine Research*
Svetlana Kalijnaia	*University of St Andrews*
Tomasz Furmanec	*Institute of Marine Research*
Tor Einar Horsberg	*Norwegian veterinary college*
Trond Ove Hjelmervik	*Norwegian veterinary college*
Øyvind Drivnes	*Institute of Marine Research*
Michael Nuhn	*European Bioinformatics Institute*


FHF
FISKERI- OG HAVBRUKSNÆRINGENS FORSKNINGSFOND


marineharvest
excellence in seafood


University
of Victoria


MAX-PLANCK-GESELLSCHAFT


EMBL-EBI


HAVFORSKNINGSINSTITUTTET
INSTITUTE OF MARINE RESEARCH


The Salmon Louse
Genome Project


Sea Lice
Research Centre