DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2011, Article ID bar050, doi:10.1093/database/bar050

Original article

SalmonDB: a bioinformatics resource for *Salmo salar* and *Oncorhynchus mykiss*

Alex Di Génova[1], Andrés Aravena[1,2,*], Luis Zapata[1,3], Mauricio González[1,4], Alejandro Maass[1,2] and Patricia Iturra[3]

[1]Laboratory of Bioinformatics and Mathematics of the Genome, Center for Mathematical Modeling (UMI 2807 CNRS) and Center for Genome Regulation (Fondap 15090007), University of Chile, Santiago, Chile, [2]Department of Mathematical Engineering, Faculty of Physical and Mathematical Sciences, University of Chile, Santiago, Chile, [3]ICBM Human Genetics Program, Faculty of Medicine, University of Chile, Santiago, Chile and [4]Laboratorio de Bioinformática y Expresion Génica, INTA, University of Chile, Santiago, Chile

***Corresponding author:** Tel: +56(2) 978 48 70; Fax: +56(2) 688 97 05; Email: andres.aravena@dim.uchile.cl

Submitted 1 July 2011; Revised 21 September 2011; Accepted 16 October 2011

SalmonDB is a new multiorganism database containing EST sequences from *Salmo salar*, *Oncorhynchus mykiss* and the whole genome sequence of *Danio rerio, Gasterosteus aculeatus, Tetraodon nigroviridis, Oryzias latipes* and *Takifugu rubripes*, built with core components from GMOD project, GOPArc system and the BioMart project. The information provided by this resource includes Gene Ontology terms, metabolic pathways, SNP prediction, CDS prediction, orthologs prediction, several precalculated BLAST searches and domains. It also provides a BLAST server for matching user-provided sequences to any of the databases and an advanced query tool (BioMart) that allows easy browsing of EST databases with user-defined criteria. These tools make SalmonDB database a valuable resource for researchers searching for transcripts and genomic information regarding *S. salar* and other salmonid species. The database is expected to grow in the near feature, particularly with the *S. salar* genome sequencing project.

Database URL: http://genomicasalmones.dim.uchile.cl/

Introduction

Atlantic salmon (*Salmo salar*) and Rainbow trout (*Oncorhynchus mykiss*) are some of the fish of importance in aquaculture that have been studied extensively from a biological perspective. Furthermore, scientific interest as a model species has generated research in various aspects such as genetics, physiology, immunology and ecology, among others. This large amount of information has been enriched by the significant increase in genomic research (1).

Several projects and sequencing efforts have produced important genomic resources such as BAC libraries (2), physical (3) and genetic maps (4–6) and ESTs from different stages and tissues from salmon (7, 8). These studies have allowed the search of putative genetic markers such as Single Nucleotide Polymorphism (SNP) (9) and microsatellites (10, 11), the development of DNA microarrays for the global analysis of gene expression (7, 12, 13) and the identification of candidate genes for multiple studies (14, 15). This information has been made publicly available through submissions to NCBI and in publicly accessible websites (16–18). This large amount of salmon EST data represents an opportunity for bioinformatics to explore and make this information available to the scientific community. Together, all these resources complement the sequencing projects for salmon species (8), in particular the upcoming Atlantic salmon genome (1).

A major challenge in genomic studies of Atlantic salmon and Rainbow trout is the complexity of their genomes due to a recent whole genome duplication event (19) and the presence of a large number of repeated elements in their genome (20). The common ancestor of salmonids experienced a genome duplication event 25–120 million years ago (19) yielding to a 6–7% nucleotide similarity between paralogous pairs (21). One strategy that helps to understand the architecture and function of these complex

Database, Vol. 2011, Article ID bar050, doi:10.1093/database/bar050

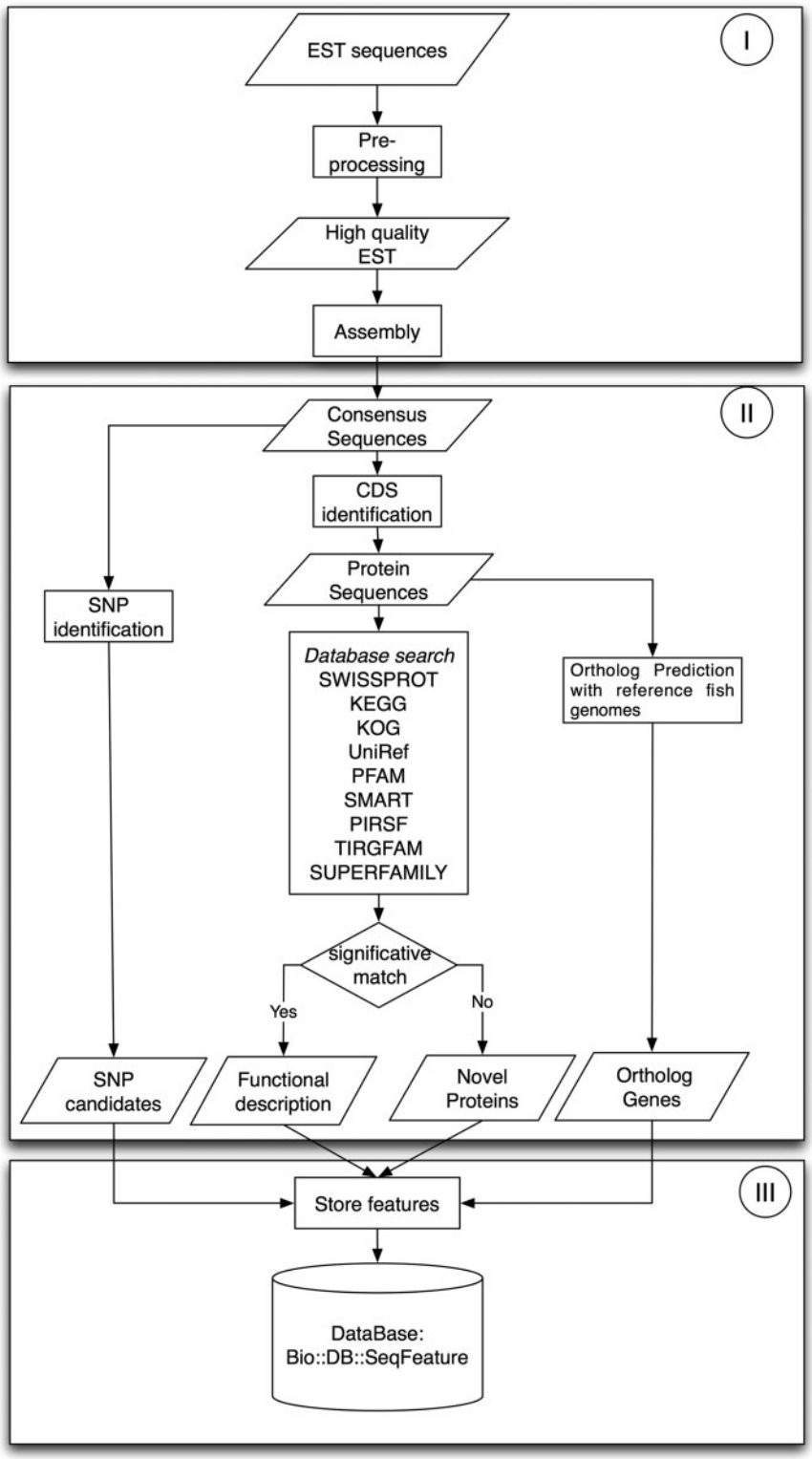

Figure 1. EST workflow: Phase I is preprocessing and assembly; Phase II, sequence annotation and characterization and Phase III, storage of biological information.

genomes is to apply comparative genomics, using available sequenced genomes of related species such as zebrafish (*Danio rerio*), medaka (*Oryzias latipes*) (22) and fugu (*Takifugu rubripes*) (23). In order to facilitate these studies, we have developed SalmonDB, a database which includes unigenes from EST data, functional annotation of putative CoDing Sequences (CDS), orthology relationship to other fish genomes, mapping information to metabolic pathways, tools for primer design and flexible data query, among others.

SalmonDB architecture

SalmonDB was built based on the GBrowse genome viewer—a component of the GMOD project (24)—, the BioMart project (25) and the GOParc ontology and pathways architecture tool, which are part of the GENDB (26) genome annotation system.

GBrowse is a widely used tool for visualization of genome annotations through a web interface. GBrowse was implemented to show unigenes or chromosome features related to blast results, protein domains and SNP predictions. The database backend is implemented in MySQL using the Bio::DB::SeqFeature::Store scheme from the BioPerl project.

In order to perform data mining of our website, an easy-to-use Biomart interface was implemented. BioMart was developed recently and it has now become a fully generic data integration tool. Some model organism databases, such as Ensembl, EBI or InterPro, have implemented it in their websites to simplify the access to their large datasets. We developed a Perl object that takes a Bio::DB::SeqFeture::Store scheme to populate a biomart scheme. Then, we used the BioMart developing tools (Marteditor, MartBuilder and Martrunner) to configure our interface.

Also, to provide visualization of the metabolic pathways and gene ontology, we modified the GOParc code to perform queries to a Bio::DB::SeqFeature::Store database that stores the information related to EC Numbers (pathway data) and GO Numbers (Gene Onotology data).

SalmonDB is based on a modular architechture that allows us to integrate new upcoming tools or informations related to salmonids or other aquaculture species easily.

SalmonDB contents

SalmonDB collects unigene sequences for Atlantic salmon and Rainbow trout. It also stores annotated genome sequences of zebrafish, fugu, stickleback, medaka and pufferfish, which allows inter-species comparison studies.

The current dataset release is based on 495 257 *S. salar* ESTs and 285 359 *O. mykiss* ESTs collected from 93 libraries including different tissues, developmental stages and treatments, downloaded from the public NCBI EST database [GRASP consortium generated most of these data (27)]. The genome sequences and annotations for the other aquaculture species were downloaded from the Ensembl database.

Both EST datasets were processed using an in-house developed workflow, shown in Figure 1, which is divided in three phases: Phase I was dedicated to EST preprocessing and assembly, Phase II was focused on sequence annotation and feature identification and in phase III all results were stored in the SalmonDB database.

Phase I required three programs: Seqclean (28), CAP3 (29) and BLAST (30), which were run to convert raw EST sequences into unigenes. SeqClean was used to remove vector sequences using NCBI's Univec database, Poly-A tails, low-quality segments at the 5′ and 3′ cDNA ends and low complexity regions. All ESTs under 100 bp were discarded. We used CAP3 to assemble high-quality EST sequences sharing over 95% identity over a region >50 bp. Then, we used BLASTN for an all against all contigs and singletons local alignment. We regrouped all sequences sharing over 95% of identity and at least 70% of coverage given from the blast output. Then, we re-assembled all the new groups using CAP3 with the same parameters used in the first assembly step. The assembly information is summarized in Table 1.

In Phase II, unigenes were analyzed to annotate putative protein products and to identify sequence features. First, all unigenes were analyzed with a BLASTX search against the Uniref (31) database to predict putative CDS and to determine the percentage of full-length cDNA contained in it. The CDS was assigned when the unigene had a significative hit ($E < 1E-10$). Unigenes without a significant hit against the Uniref database were further analyzed using ESTscan (32). Putative CDS having at least 30 aminoacids were included in the database, the rest was discarded.

The functional annotation was based on homology detection with known proteins using MPIBLAST (33) searches against Uniref (31), Swissprot (34), KEGG (35) and KOG (36) databases. To improve the functional assignment and classification, we used MPIHMMER against PFAM (37), SMART (38), TIGRFAM (39), SUPERFAMILY (40) and PIRSF (41) databases to search for motifs in all unigenes. Motifs were assigned when a domain was detected with an $E < 10^{-5}$.

Putative SNPs within sequences in the assembly were detected using the AMOS toolkit (42) and in-house developed scripts optimized for SNP discovery in complex genomes. Those sites within CDS regions, with more than four covering reads, that differ at least 20% from the consensus sequence and that were not inside repetitive sequences were

Database, Vol. 2011, Article ID bar050, doi:10.1093/database/bar050

Table 1. Final assembly details

Assembly statistics	*Salmo salar*	*Oncorhynchus mykiss*
Number of total reads	495 257	285 359
Total Unigenes in first assembly	150 720	125 077
Total Unigenes in reassembly (BLAST-CAP3)	103 221	97 667
Total Unigenes after CDS prediction	59 336	62 233
Number of reads in final assembly	387 294	213 218
Number of singletons	31 915	38 884
Average read length	619	666
Unigene length (average ± SD)	872 ± 434	880 ± 322
Average unigene depth	7	3
Maximum unigene depth	2005	1444

Total number of reads and unigenes assembled using the described pipeline.

Table 2. General SalmonDB statistics

Database	*Salmo salar*	*Oncorhynchus mykiss*
Unigenes	59 336	62 233
Total SNP	35 879	42 238
UNIREF	50 067	52 351
KEGG	30 085	31 908
SWISSPROT	41 472	44 803
KOG	33 000	35 436
PFAM	20 625	22 306
TIGRFAM	3191	3715
SMART	10 493	11 088
PIRSF	1658	1978
SUPERFAMILY	24 394	25 447

Total number of unigenes matching a database hit. On average each *S. salar* unigene has 4.2 attributes, while *O. mykiss* unigenes have 4.4.

marked as putative SNPs. Also, non-synonymous alleles and protein positions were predicted for each SNP.

Finally, we used Orthomcl (43) clustering to predict orthologs between the reference fish genomes and the salmonid species. A total of 273 395 proteins were clustered with Orthomcl using an *E*-value cutoff of 10^{-10} and a moderate inflation value of 2.5. The analysis produced a total of 28 365 clusters, where only the ortholog and paralog clusters containing salmonid species were stored in the database.

In Phase III, all sequence features were stored in MySQL as Bio::SeqFeatureI objects, cross-referenced to the corresponding external databases. Cross-references include EC, KO and KOG numbers for the Blast hits, and SMART, PIRSF, SUPERFAMILY, TIGRFAM, PFAM, INTERPRO and GO numbers for HMMER domain hits. The idea was to create a controlled vocabulary useful in other applications. A summary of SalmonDB contents is shown in Table 2.

Using SalmonDB

The database can be accessed through a web interface as seen in Figure 2. The main views are the Unigene, Genome, GO and KEGG browsers, the Blast server and the BioMart interface. It also has a help navigation page that explains step by step how to use the different tools in the website.

The Unigene Browser (Figure 2a) contains different sequence features, including CDS prediction, unigene coverage, BLAST and HMMER hits, GC content and putative SNPs, each presented as GBrowse tracks. The Genome Browser includes the complete *D. rerio, O. latipes, T. rubripes, G. aculeatus* and *T. nigroviridis* genomes and shows the genomic localization of genes, the exon/intron organization and their corresponding transcripts. Every genome contains external links to the Ensembl database for more detail. SalmonDB provides access to KEGG pathway information, through a KEGG Browser (Figure 2d), where you can select a specific EC number or browse through any pathway to find all participating unigenes. This is useful for mapping the relationships within a whole system of annotated enzymes. It is specially valuable for those who are interested in biological pathways. Moreover, SalmonDB could be queried for Gene Ontologies using the GO Browser (Figure 2c).

A web form allows the use of BLAST to find matches to an user-supplied sequence in the SalmonDB unigene databases (*S. salar*, *O. mykiss*) or the SalmonDB reference genome databases (one can search against the genome, the mRNA dataset or peptide dataset from any of the aquaculture species stored in SalmonDB). The BLAST output is dynamically linked to the Unigene and Genome Browsers (Figure 2).

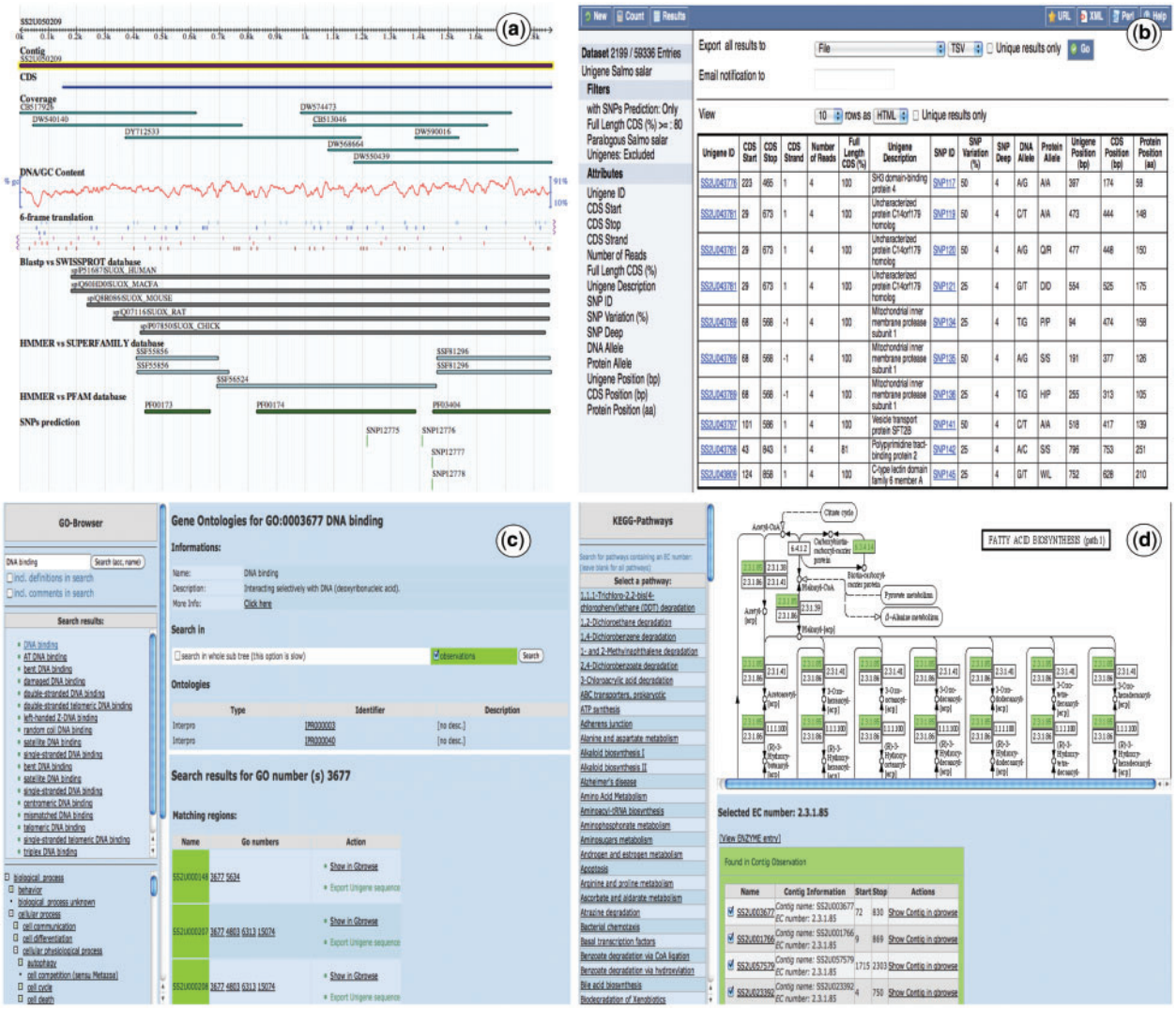

Figure 2. Snapshots of the SalmonDB web interface. (**a**) Unigene browser: the Unigene SS2U057650 is shown with several tracks (features), the blast alignment can be shown for each hit. (**b**) Biomart: the MartView interface is shown using the *S. salar* dataset and several filters selected on the left navigation panel. It also shows the ouptut table with multiple attributes shown on the left. (**c**) Go Browser: result of the search for GO term GO:003872 in the *S. salar* Unigene database. (**d**) KEGG Browser: the pathway associated to alanine and aspartate metabolism is shown using the *S. salar* Unigene database.

BioMart (Figure 2b) is an outstanding feature for SalmonDB. It provides a step by step interface that allows searching the entire database with predefined criteria. It has the advantage that one can select any data filter combination and access only the information needed by clicking on those attributes. It is fast and depends on the information stored in the local database. Complex questions can be solved through a simple query. As an example, suppose that a researcher wants to find all unigenes participating in the nutrient reservoir activity metabolism (they all share a specific GO number, GO:0045735) and that contain a putative SNP within their sequence. First, one would click on the GO filter and specify GO:0045735 number. The next step is to click on the 'SNP predicted only' filter to search for just those unigenes that have a SNP present. The search will return an output table with all unigene hits and the information that was selected in the attributes form. This information can be useful to identify potential SNP markers associated to dietary responses related to nutrient storage in salmons. The website has a step by step help navigation page for using BioMart in more detail. Recently, SalmonDB biomart has been included as part of the central biomart portal (44).

Additionally, database searches can be performed with a keyword term, accession number or any ID from the cross reference of the databases mentioned before by entering the term in the quick box search or in the Gbrowse search box.

Database, Vol. 2011, Article ID bar050, doi:10.1093/database/bar050

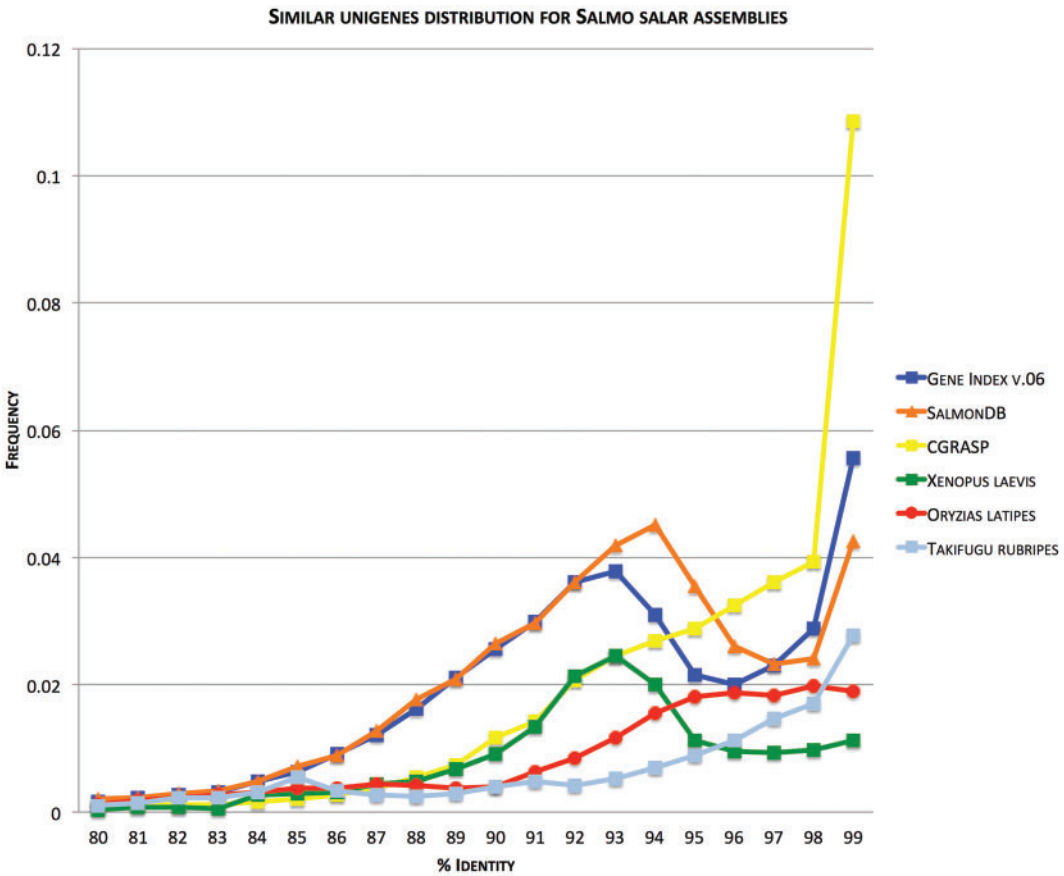

Figure 3. Frequency of aligned Unigenes plotted against percent identity. Figure (modified from [45]) shows frequency of top-pairwise alignment ($E < 1e\text{-}10$; query and subject coverage = 0.9) between Unigenes generated through our assembly pipeline plotted against identity score (SalmonDB, orange). It also shows the relationships among the contig consensus sequences of gene index EST assembly (Gene Index, blue) and cGRASP EST assembly (CGRASP, yellow) for Atlantic salmon. The same analysis is included for Fugu (*Takifugu rubripes,* light blue) and Medaka (*Oryzias latipes*, dark red) mRNAs obtained from Ensembl and the African Clawed Frog (*Xenopus laevis*, green) Unigenes obtained from NCBI. Since there is no standard metodolgy to compare EST assemblies (e.g. Genome assembly has N50 value), a good approximation is to observe the expected pattern for a duplicated genome using this strategy. We include the African clawed frog because it has a well-documented recent genome duplication. The expected pattern is shown in the figure with a peak around 93–94%. The same is expected for Salmon which suffered from a whole genome duplication ~100 million years ago. SalmonDB and gene index assembly show these accumulation of paralogs around 93–94% identity.

The other available databases rely on EST assembly and gene annotation data (17, 18), or the physical map based on BAC fingerprinting with BAC end sequence data (16). A comparison of the assemblies based on percent similarity among the final number of unigenes is shown on Figure 3. This plot shows the expected peak for a recent genome duplication event (45). Also, the complementary capabilities of each database and assembly statistics such as percent full-length cDNAs are shown in Tables 3 and 4, respectively.

SalmonDB is intended to fully exploit genetic information regarding salmon and provides several tools and pre-calculated analyzed data that can be easily browsed through the BioMart interface. It is also possible to perform fast comparative genomic research with other salmon databases and fish reference genomes. Among several tools, it is possible to design primers within salmon sequences and search for these primer sequences in the other genomes. This could enable an effective comparison of intron/exon boundaries among salmon and other fishes. Among other important features, SalmonDB provides with several putative SNPs that are accessible for all scientific community in order to validate and use them for genotyping experiments. All these combined information can help the researcher to conduct experiments and, therefore, improve results.

Future development of SalmonDB

In the near future, we will incorporate genomic information provided from the Atlantic salmon genome

Table 3. Global comparison of available salmon databases

	SalmonDB	GRASP	ASALBASE	Gene index
Data				
Data source	All public ESTs	Public ESTs, BAC ends	BAC clones, BAC ends and EST cluster	NCBI ESTs
Base pair quality	No	Yes	No	No
EST assembly	CAP3, clustering	Phrap	No	Clustering, CAP3
Physical map	No	No	Yes	No
Genetic map	No	No	Yes	No
Expression data	No	Yes	No	No
Tools				
Blast homology search	Yes	Yes	No	Yes
Quick search box	Yes	No	Yes	No
Primer design	Yes	No	No	No
RepeatMasking	No	Yes	No	No
GO annotation browser	Yes	No	No	Yes
KEGG annotation browser	Yes	No	No	Yes
Advanced search with Biomart	Yes	No	No	No
Analysis				
Ortholog prediction	Yes	Yes	Yes	No
Paralog prediction	Yes	No	No	No
SNP prediction	Yes	No	No	Yes
CDS prediction	Yes	Yes	No	Yes
Other markers	No	No	Yes	No
Full-length cDNA prediction	Yes	Yes	Yes	No
Alternative splicing forms prediction	No	No	No	Yes
Others				
Web interface	Gbrowse	Gbrowse	Gbrowse, custom	custom
Other organism data	5 fish species	Other salmonids and salmon lice	4 fish species and Human	Other TIGR organisms

cGRASP information was extracted directly from the http://web.uvic.ca/grasp/ website that includes features from external links. Gene index information was obtained from the website http://compbio.dfci.harvard.edu/tgi/.

Table 4. Assembly statistics comparison of available salmon databases

	SalmonDB	Gene index	cGRASP
Unigenes	59 336	99 285	81 236
Total length (Mb)	51	84	71
Min length	100	100	75
Max length	4563	5828	4780
Average length	872	854	881
Median length	771	755	758
Full-length cDNA	5939[a]	7124	7625
% Full-length protein	10.01	7.18	9.39

Table shows statistics for the three Atlantic salmon assemblies. Total number of unigenes constructed using each database pipeline, total sequence length from all unigenes and their statistics. Also, we show the number of full-length cDNAs calculated using blastx against nr database (counted as full-length when the unigene cover 99% or more of the protein).
[a]Number of full-length cDNAs from SalmonDB biomart is 7465. This number was calculated using translated sequences (blastp) instead of blastx against nr.

sequencing project (1) and publicly available transcriptomic data from Illumina/Solexa or Roche/454 sequences (46).

We expect to incorporate additional tools in order to allow scientists to explore the genetic and physical maps of *S. salar*. Also, we are integrating our database with the existing resources for salmonids using cross references to Gene Index TCs and cGRASP unigenes. Therefore, a link between similar unigenes (98% identity and 95% coverage for both sequences) will be provided in order to navigate through the different databases.

Several ongoing projects on salmon require an easy to access database with several tools available. Next-generation sequencing technologies will bust up the amount of information related to sequences. Thus, our experience in constructing databases (44, 47), NGS pipeline development and SNP discovery for salmon sequences will alow us to build a new version of the database every year with the goal of providing up to date information to end users.

Finally, Chile is part of the International Collaboration to sequence the Atlantic Salmon Genome (ICSASG) (1). Thus, the access to data will allow us to exploit different pipelines, tools and methodologies regarding salmon genome sequences. In the future, our goal is to become an important reference database for the salmonid species.

Acknowledgements

We would like to thank Dr William Davidson for his comments on the first manuscript.

Funding

The development, creation and hosting of SalmonDB was supported by CORFO-INNOVA (grant 07CN13PBT-41); Fondecyt (1110427) and Fondap (No 15090007); Basal Grant CMM Projects. Funding for open access charge: Fondap (No 15090007).

Conflict of interest. None declared.

References

1. Davidson,W., Koop,B., Jones,S. *et al*. (2010) Sequencing the genome of the Atlantic salmon (Salmo salar). *Genome Biol.*, **11**, 403.
2. Thorsen,J., Zhu,B., Frengen,E. *et al*. (2005) A highly redundant bac library of Atlantic salmon (Salmo salar): an important tool for salmon projects BMC. *Genomics*, **6**, 50.
3. Ng,S.H., Artieri,C.G., Bosdet,I.E. *et al*. (2005) A physical map of the genome of Atlantic salmon, Salmo salar. *Genomics*, **86**, 396–404.
4. Moen,T., Hayes,B., Baranski,M. *et al*. (2008) A linkage map of the Atlantic salmon (Salmo salar) based on EST-derived SNP markers. *BMC Genomics*, **9**, 223.
5. Danzmann,R.G., Davidson,E.A., Ferguson,M.M. *et al*. (2008) Distribution of ancestral proto-actinopterygian chromosome arms within the genomes of 4r-derivative salmonid fishes (rainbow trout and Atlantic salmon). *BMC Genomics*, **9**, 557.
6. Palti,Y., Genet,C., Luo,M. *et al*. (2011) A first generation integrated map of the rainbow trout genome. *BMC Genomics*, **12**, 180.
7. Rise,M.L., vonSchalburg,K.R., Brown,G.D. *et al*. (2004) Development and application of a salmonid est database and cDNA microarray: data mining and interspecific hybridization characteristics. *Genome Res.*, **14**, 478–490.
8. Koop,B., vonSchalburg,K., Leong,J. *et al*. (2008) A salmonid EST genomic study: genes, duplications, phylogeny and microarrays. *BMC Genomics*, **9**, 545.
9. Hayes,B., Laerdahl,J., Lien,S. *et al*. (2007) An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (Salmo salar) expressed sequences. *Aquaculture*, **265**, 82–90.
10. Ng,S., Chang,A., Brown,G. *et al*. (2005) Type I microsatellite markers from Atlantic salmon (Salmo salar) expressed sequence tags. *Mol. Ecol. Notes*, **5**, 762–766.
11. Vasemägi,A., Nilsson,J. and Primmer,C.R. (2005) Expressed sequence tag-linked microsatellites as a source of gene-associated polymorphisms for detecting signatures of divergent selection in Atlantic salmon (Salmo salar l.). *Mol. Biol. Evol.*, **22**, 1067–1076.
12. Taggart,J., Bron,J., Martin,S. *et al*. (2008) A description of the origins, design and performance of the traits–sgp Atlantic salmon (Salmo salar l. cDNA microarray. *J. Fish Biol.*, **72**, 2071–2094.
13. Rise,M.L., vonSchalburg,K.R., Cooper,G.A. *et al*. (2007) Salmonid DNA microarrays and other tools for functional genomics research. In: Zhangjiang,L. (ed). *Aquaculture Genome Technologies*. Blackwell Publishing, Oxford, UK, pp. 369–412.
14. Leaver,M., Villeneuve,L., Obach,A. *et al*. (2008) Functional genomics reveals increases in cholesterol biosynthetic genes and highly unsaturated fatty acid biosynthesis after dietary substitution of fish oil with vegetable oils in Atlantic salmon (Salmo salar). *BMC Genomics*, **9**, 299.
15. Panserat,S. and Kaushik,S. (2010) Regulation of gene expression by nutritional factors in fish. *Aquaculture Res.*, **41**, 751–762.
16. The Atlantic salmon genomics database (ASALBASE). http://www.asalbase.org/ (10 June 2011, date last accessed).
17. The University of Victoria Grasp site (GRASP). http://web.uvic.ca/grasp/ (10 June 2011, date last accessed).
18. The TIGR gene index database (TIGR). http://compbio.dfci.harvard.edu/tgi/ (10 June 2011, date last accessed).
19. Allendorf,F. and Thorgaard,G. (1984) Tetraploidy and the evolution of salmonid fishes. In: Turner,B.J. (ed). *Evolutionary Genetics of Fishes*. Plenum Press, New York, pp. 1–53.
20. deBoer,J.G., Yazawa,R., Davidson,W.S. *et al*. (2007) Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. *BMC Genomics*, **8**, 422.
21. Leong,J.S., Jantzen,S.G., vonSchalburg,K.R. *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.
22. Kasahara,M., Naruse,K., Sasaki,S. *et al*. (2007) The medaka draft genome and insights into vertebrate genome evolution. *Nature*, **447**, 714–719.
23. Aparicio,S., Chapman,J., Stupka,E. *et al*. (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science*, **297**, 1301–1310.

24. Stein,L.D., Mungall,C., Shu,S. *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
25. Smedley,D., Haider,S., Ballester,B. *et al*. (2009) Biomart–biological queries made easy. *BMC Genomics*, **10**, 22.
26. Meyer,F., Goesmann,A., McHardy,A.C. *et al*. (2003) Gendb–an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.*, **31**, 2187–2195.
27. Consortium for genomic research on all salmonids program (cGRASP). http//: www.cgrasp.org/ (10 June 2011, date last accessed).
28. Seqclean: a script for automated trimming and validation of ESTs or other DNA sequences by screening for various contaminants, low quality and low-complexity sequences. http://compbio.dfci.harvard.edu/tgi/software/.
29. Huang,X. and Madan,A. (1999) Cap3: A DNA sequence assembly program. *Genome Res.*, **9**, 868–877.
30. Altschul,S.F., Madden,T.L., Schäffer,A.A. *et al*. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
31. Suzek,B.E., Huang,H., McGarvey,P. *et al*. (2007) Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, **23**, 1282–1288.
32. Iseli,C., Jongeneel,C.V. and Bucher,P. (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 138–148.
33. Archuleta,J., Tilevich,E. and chunFeng,W. (2007) IEEE Int. Conf Softwar. Maint., 10.
34. Bairoch,A., Boeckmann,B., Ferro,S. *et al*. (2004) Swiss-prot: juggling between evolution and stability. *Brief. Bioinformatics*, **5**, 39–55.
35. Kanehisa,M. (2002) The kegg database. *Novartis Found. Symp.*, **247**, 91–101; discussion 101–103, 119–128, 244–252.
36. Tatusov,R.L., Fedorova,N.D., Jackson,J.D. *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
37. Finn,R.D., Tate,J., Mistry,J. *et al*. (2008) The pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
38. Letunic,I., Doerks,T. and Bork,P. (2009) Smart 6: recent updates and new developments. *Nucleic Acids Res.*, **37**, D229–D232.
39. Haft,D.H., Selengut,J.D. and White,O. (2003) The tigrfams database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
40. Wilson,D., Madera,M., Vogel,C. *et al*. (2007) The superfamily database in 2007: families and functions. *Nucleic Acids Res.*, **35**, D308–D313.
41. Nikolskaya,A.N., Arighi,C.N., Huang,H. *et al*. (2006) Pirsf family classification system for protein functional and evolutionary analysis. *Evol. Bioinform. Online*, **2**, 197–209.
42. Phillippy,A.M., Schatz,M.C. and Pop,M. (2008) Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol.*, **9**, R55.
43. Feng,C., Aaron,J., Mackey,Christian J, Stoeckert,Jr. and David,S. Roos (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.
44. Guberman,J.M., Arnaiz,O., Baran,J. *et al*. (2011) BioMart Central Portal: An Open Database Network for the Biological Community, In press.
45. Koop,B., von Schalburg,K., Leong,J. *et al*. (2008) A salmonid EST genomic study: genes, duplications, phylogeny and microarrays. *BMC Genomics*, **9**, 545.
46. Salem,M., Rexroad,C., Wang,J. *et al*. (2010) Characterization of the rainbow trout transcriptome using sanger and 454-pyrosequencing approaches. *BMC Genomics*, **11**, 564.
47. The Potato Genome Sequencing Consortium (PGSC). (2011) Genome sequence and analysis of the tuber crop potato. *Nature*, **475**, 189–195.